

A look under the hood of the Interactive Deep Learning Enterprise (No-IDLE)

DANIEL SONNTAG, MICHAEL BARZ, and THIAGO GOUVÊA, German Research Center for Artificial Intelligence (DFKI), Germany

This DFKI technical report presents the anatomy of the No-IDLE prototype system (funded by the German Federal Ministry of Education and Research) that provides not only basic and fundamental research in interactive machine learning, but also reveals deeper insights into users' behaviours, needs, and goals. Machine learning and deep learning should become accessible to millions of end users. No-IDLE's goals and scientific challenges centre around the desire to increase the reach of interactive deep learning solutions for non-experts in machine learning. One of the key innovations described in this technical report is a methodology for interactive machine learning combined with multimodal interaction which will become central when we start interacting with semi-intelligent machines in the upcoming area of neural networks and large language models.

1 INTRODUCTION

In recent years, machines have surpassed humans in the performance of specific and narrow tasks such as some aspects of image recognition or decision making along clinical pathways in the medical domain (weak AI). Although it is very unlikely that machines will exhibit broadly-applicable intelligence comparable to or exceeding that of humans in the next 30 years (strong AI), it is to be expected that machines will reach and exceed human performance on more and more applied tasks. To develop the positive aspects of AI, manage its risks and challenges, and ensure that everyone has the opportunity to help in building an AI-enhanced society and to participate in its benefits, in this project, human intelligence and machine learning (ML) take the centre stage: *Interactive Machine Learning (IML) is the design and implementation of algorithms and intelligent user interface frameworks that facilitate ML with the help of human interaction.*

Our focus is to improve the interaction between humans and machines, by leveraging state-of-the-art human-computer interaction (HCI) approaches, as well as solutions that involve state-of-the-art ML techniques. In this project, we focus on *Interactive Deep Learning (IDL): deep learning (DL) approaches for IML.* We want computers to learn from humans by interacting with them in natural language for example and by observing them. Our goal in No-IDLE ¹ is to improve the interaction between humans and machines to update DL models, by leveraging both state-of-the-art human-computer-interaction and DL approaches. Basic and fundamental research in this corridor project should also reveal deeper insights into users' behaviours, needs, and goals. Machine learning and DL should become accessible to millions of end users, and be functionally more advanced than current recommender systems in online shops that provide suggestions for items that are most pertinent to a particular user. Explicit (ontological) knowledge representation and reasoning capabilities are however not part of this focused project, but a follow-up project would highly benefit from them. In addition, we emphasise the role of multimodal interaction and mixed-initiative interaction. While focusing on IDL in this corridor project, we pose the development of a methodology for IDL as a challenge problem. A methodology for IDL will become central when we start interacting more with semi-intelligent machines. As a layer used to represent the interactions, opinions and feedback, it is critical that IML is well understood and defined. Also, there has been recent and relatively rapid success of AI and ML solutions that arise from neural network architectures. But neural networks lack the interpretability and transparency needed to understand the underlying decision process and learned

¹<https://www.dfki.de/en/web/research/projects-and-publications/project/no-idle>

representations. Making sense of why a particular model misclassifies test data instances or behaves poorly at times is a challenging task for model developers and is an important problem to address [Hohman et al. 2018]. A related argumentation is that despite their huge successes, largely in problems which can be cast as classification problems, the effectiveness of neural networks is still limited by their un-debuggability, and their inability to “explain” their decisions in a human understandable and reconstructable way [Goebel et al. 2018].

In No-IDLE, we explore the relationship between DL, HCI, and explainable AI (XAI). For example, by approaching the problem from the HCI perspective, recent work has shown the benefits of visualising complex data in virtual reality (VR), e.g., in data visualisation [Donalek et al. 2014], and big data analytics [Moran et al. 2015]. In one HCI subtask in No-IDLE for example, we extend an interactive image clustering method in VR [Prange and Sonntag 2021], where the user can explore and then fine-tune the underlying DL model through intuitive hand gestures. While HCI constitutes a key approach, we will attack the IML problem from multiple angles. Informed by emerging directions in both research and commercialisation of IML systems [Oviatt et al. 2019; Zacharias et al. 2018], we will deploy our expertise in multimodal-multisensor interfaces (MMI) and natural language processing (NLP), while also tapping on the broader interdisciplinary community, to deliver on the mission to improve interaction between humans and machines. Past application projects of DFKI’s IML group include deep active learning such as described in [Shui et al. 2020], explanatory interactive image captioning [Biswas et al. 2020], IDL systems for melanoma detection [Sonntag et al. 2020] and wildlife monitoring [Gouvêa et al. 2023], toolkits for building multimodal systems and applications [Barz et al. 2021a; Oviatt et al. 2019], and interactions with ML systems as domain-specific explanations [Hartmann et al. 2021]. In No-IDLE, we bring these approaches, technologies and our experience together to apply them to a special use case, namely interactive photo book creation, to test and evaluate the basic and fundamental research in this corridor project. The proposed project builds upon this broad experience and research results of the IML group in the areas of human-computer interaction (HCI), machine learning (ML), multimodal human-computer interaction (MMI), and natural language processing (NLP).

In a nutshell, in No-IDLE we explore IDL from four different perspectives (HCI, ML, NLP, MMI). No-IDLE is a basic research project to advance our understanding of IML. We expect practical contributions to be made while bringing the four working groups of IML closer together to work on a specific application around IML for photo book creation and the exploitation of the findings in ongoing DFKI consortial and industrial projects.

2 USE CASE: INTERACTIVE PHOTO BOOK CREATION

The research questions raised in No-IDLE will be investigated in the context of a specific use case: the interactive creation of a photo book. Consider the following scenario:

Family Smith (a family of four) takes many photos from all kinds of events and occasions and regularly likes to create personal photo books and calendars for themselves and as gifts for family members and friends. Selecting the right photos, arranging them and writing captions is fun but very time consuming, and while they appreciate it as a means of their personal expression and creativity, they would like to speed up the process, especially with respect to the more tedious parts like selecting among similar photos or finding a basic arrangement. At the same time, they would like to maintain control and a personal connection to the results. Each family member has their own personal taste: some are more inclined to funny situations and photos of people, other prefer scenic views and interesting lighting and their personal style of arrangement, some like to put the photos simply side by side, others like to make use of interesting frames, clip art and creative arrangements. In addition, the goal and target audience influence their choices. For instance, they like to create diary type photo books of their travels for their own archive but like to tell image stories of the same

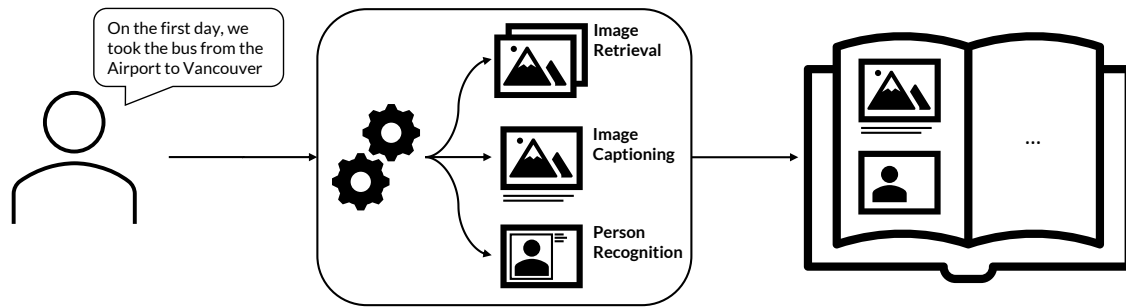


Fig. 1. We plan to combine several modules based on deep learning models to create photo book pages from natural language input. These modules include, for instance, image retrieval, image captioning, and person recognition.

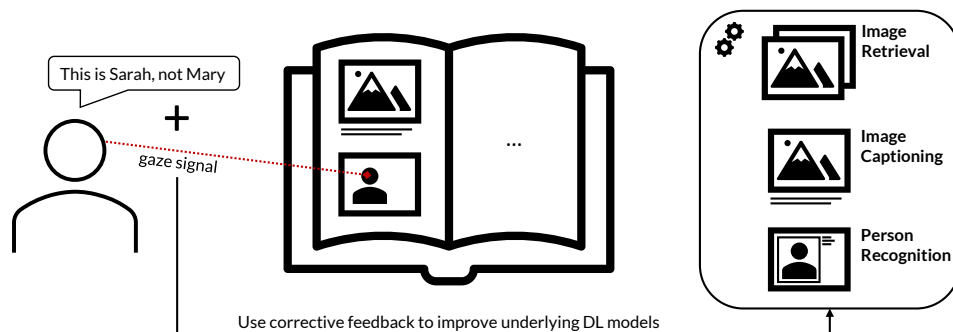


Fig. 2. The user can provide multimodal feedback to the photo book tool to alter the created content. For instance, we plan to jointly interpret the user's gaze signal and spoken utterances to improve person recognition. An example is shown in figure 3.

trip for showing them or gifting them to others. When they create books or calendars for special holidays or birthday gifts, they typically select photos that somehow match the occasion but that also contain the gifted person if possible.

Thankfully, they find out about the AI software that integrates techniques developed within NO-IDLE. Using these, a photo book can be created by providing a set of images and by sequentially describing the occasion in natural language, be it a holiday trip or a wedding party. They can also describe the style and purpose of the photo book to guide the creation process. To make an example, imagine that they plan to create a photo book about their last family trip to Canada. They start off by telling the system: "This will be a photo book for aunt Mary about our last trip to Canada. We would like to add some dramatic touch to it". In return, the photo book creation tool suggests a suitable caption and basic style for the photo book. If not suitable, they can edit the caption or adapt the style, e.g., by selecting another frame type for captions or another font family. They would continue by describing how they perceived their vacation to the photo book tool just like they would describe it to another human: "On the first day, we took the bus from the airport to Vancouver" (see figure 1). As a response, the system creates a single page with suitable photos, i.e., from getting on the bus at the airport, a photo of the skyline of Vancouver from inside the bus and one with aunt Mary who was waiting for them at the bus stop. Since this is the first time family Smith is using this tool, the automatic caption generation module is uncertain whether its output is suitable and, hence, actively asks for feedback.

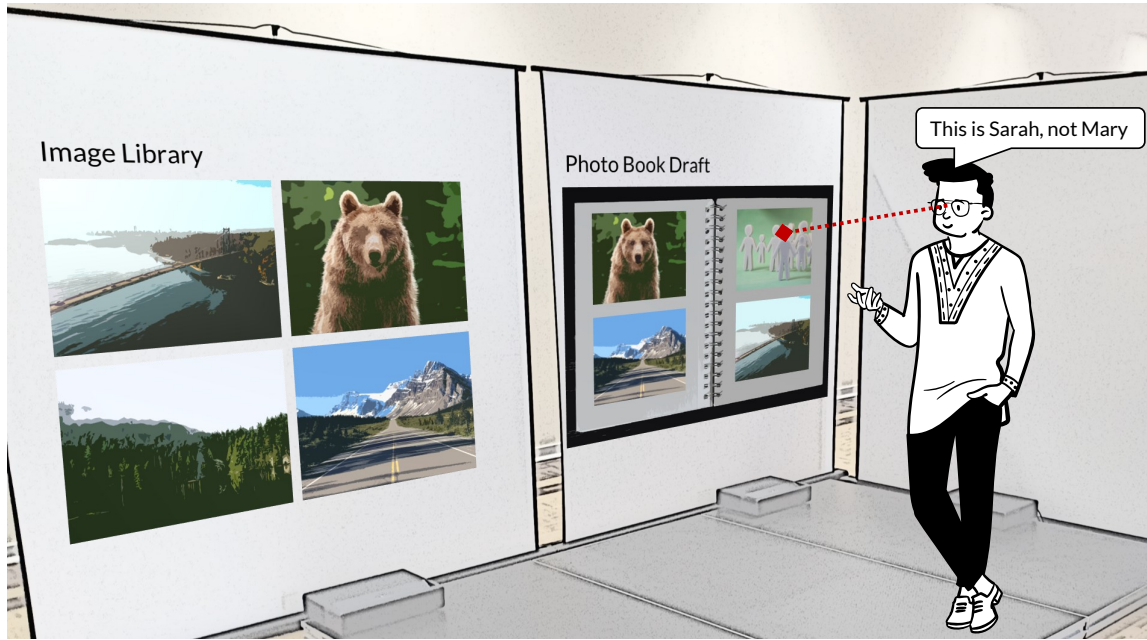


Fig. 3. Example of a multimodal user input to our photo book application (based on an existing demo setup). The user provides corrective feedback in natural language by saying “This is Sarah, not Mary”. The system uses his gaze to resolve the face that was referred to and uses the new information to update the underlying deep learning models as depicted in figure 2.

Being happy with this partial result, the family continues to describe the events saying “The incident with the bears was extremely funny and the woods were so impressive”. The newly generated pages of the photo include pictures of the bear and the woods from their hiking trip, but none with aunt Mary, so they complain about this. “Please add a picture with Mary here”. As the system does not know yet how Mary looks, it shows extracted faces from the provided photos and asks to select a picture of Mary. Mrs. Smith looks at a picture and says “that’s my sister Mary”. The system uses the gaze signal to identify the face that was referred to and learns to recognise Mary. Eventually, family Smith reports how their vacation ended: “it was also something how aunt Mary had to take us to the airport on short notice because our car broke down and we almost thought we wouldn’t make it and how they welcomed us back at the airport after we landed.” One of the images shows Sarah in front of aunt Mary’s car, but the caption states “This is aunt Mary after carrying us to the airport”. Mr. Smith corrects the system by saying “this is Sarah, not Mary” (see figures 2, 3, and 4). The system automatically corrects the caption and corrects the label for the detected face. From now on, the system will be better at differentiating between Sarah and her sister Mary. Alternatively, Mr. Smith could edit the caption to “This is Sarah in front of her car after carrying us to the airport last minute.” and the feedback contained in this post-edit would be used to update the image captioning model.

While the initial draft of the photo book is already quite good, the Smiths want to add some personal touch and they also spot some errors browsing through the suggestions. The system supports an immersive mode using VR or just a normal desktop/tablet-based presentation. While they could use either mode and intuitive hand or touch gestures in combination with gaze-tracking/spoken dialogue to rearrange and edit each caption and photo by pointing or touching a photo and selecting from better alternatives presented by the system, by rating a photo as not suitable, or by providing

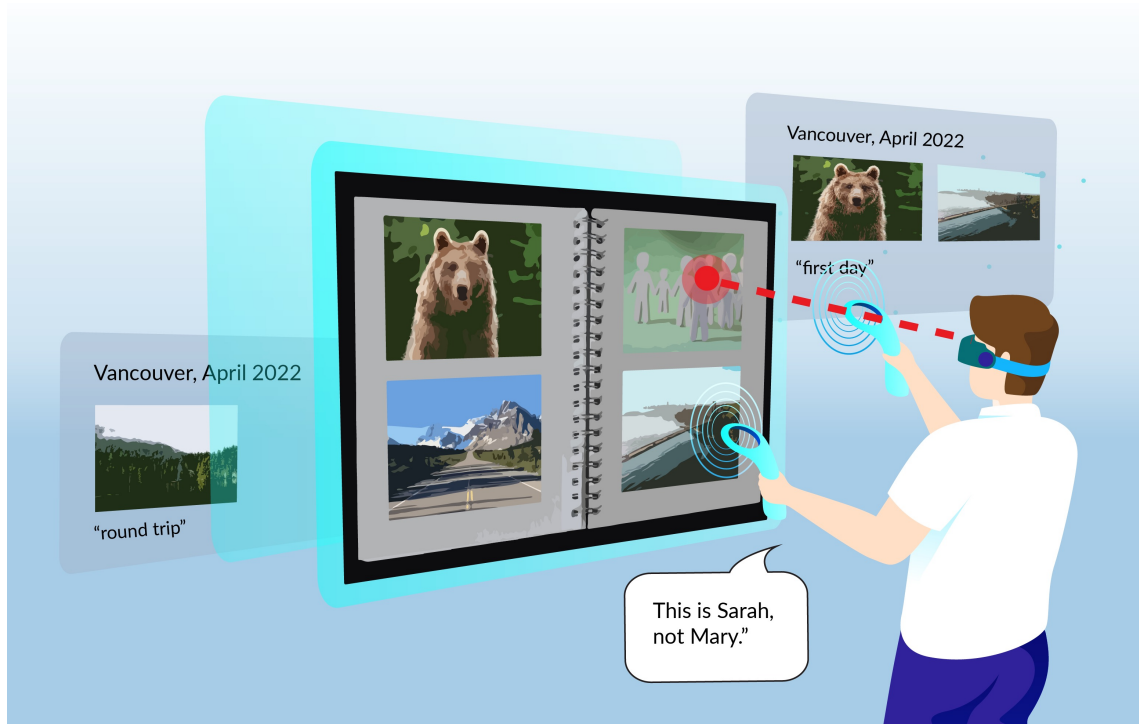


Fig. 4. Visualisation of the virtual reality scenario. Images and the photo book are presented in an immersive virtual environment. Through multimodal interaction (pointing, eye-/gaze-tracking, natural speech) the user engages with the system and provides corrective feedback by saying "This is Sarah, not Mary". The system uses implicit and explicit pointing or gaze to resolve the face that was referred to and uses the new information to update the underlying deep learning models as depicted in figure 2. In addition to the multimodal setup depicted in figure 3, VR tracking provides detailed spatial tracking information that will be included in the data analysis.

feedback to a caption, the system also provides some higher-level tools: for story-based books the overall time and dramatic flow of the story and the included events are visualised along a time line (which works especially well in VR thanks to almost unlimited virtual space). To avoid clutter, each event is represented by some iconic photos and a summarising caption, generated by the system. The Smiths can now put more or less emphasis on certain events, add or remove whole events, or "zoom" in and identify key characters and photos. For diary-type or location-centered books, the photos are clustered accordingly and visualised over a floating map and again the Smiths can now edit and provide feedback using rich multi-modal input. The system will continue to learn from the user input and actively ask for help in uncertain cases. The rich input/output modalities (especially in the VR case) will benefit user and system on several levels. They will make active learning by the system more effective because multimodality can be used to disambiguate and to compensate for noise in single modalities. They will also improve the user experience because they allow for a more intuitive and effective interaction and visualisation and as they provide more data about the user to the system, the system can learn more effectively (using not only explicit but implicit inputs) about the user preference and can adapt the information load.

Over the past decade, researchers have studied similar scenarios [Sandhaus et al. 2008] and proposed partial solutions for certain sub-task. For instance, different methods ranging from semantic modelling [Sandhaus and Boll 2011] and meta data analysis [Boll et al. 2006, 2007] to deep learning solutions [Withöft et al. 2022] have been investigated for retrieving and filtering photos according to general criteria or personal preferences [Maszuhn et al. 2021]. Some of these works have also looked at data from social media activity to learn about user preferences or events [Rabbath et al. 2011a,b]. Other works have looked at the presentation layer, for instance, at how to create aesthetic layouts [Sandhaus et al. 2011] or how to design novel augmented reality interaction techniques to allow users to easily annotate their photos [Henze and Boll 2011]. However, integrated solutions for a complete system are still missing, which highlights both the relevance but also the challenge of the presented scenario. While the goal of this project is not to develop a market-ready photo book application software, we are certain that we will be able to implement the use case as an AI testbed to extend the current state-of-the-art in interactive deep learning. We propose a unique and integrated approach that draws on our expertise from machine learning, NLP, multimodal interaction and HCI research.

3 GOALS AND SCIENTIFIC CHALLENGES OF NO-IDLE

With the convergence of artificial intelligence and machine Learning, IDL is where the HCI community meets the DL community [Amershi et al. 2014; Dudley and Kristensson 2018; Holzinger 2016; Sonntag 2010; Teso and Hinz 2020; Zacharias et al. 2018]. No-IDLE’s goals and scientific challenges centre around the desire to increase the reach of DL solutions (and ML solutions in general): DL for non-experts in ML and improving DL models when not enough data is available (e.g., due to highly individualised tasks like photo book creation) or data quality is not sufficient. In addition, to fully automate tasks in practical applications such as our use case of interactive photo book creation can be extremely difficult and even undesirable. As a consequence, our goals are to find a computational and design methodology to gracefully combine automated services with direct user input or manipulation. We investigate our scientific goals in the context of our photo book application. However, the technologies developed shall be beneficial for other domains as well such as healthcare or smart manufacturing. They can be summarised as follows:

- (1) Define and declare the role of humans in IDL (HCI): (1) realising the importance of studying users; (2) reducing the need for supervision by ML practitioners; (3) explore interactivity in a tight coupling between the system and the user; (4) handle human ambiguity and confusion and instil trust and confidence through feedback and explanations; (5) explore gamification and serious games in the context of IDL and IML in general.
- (2) Provide a way for users to (1) understand why the system had made a particular prediction, and (2) adjust the (DL) learner’s reasoning if its prediction was wrong. To this end, the system should provide an explanation for its predictions, and incorporate corrective feedback given by the user. How can this be done in practical terms? For providing useful explanations of model predictions, we will investigate the feasibility of solving tasks with interpretable (DL) models rather than black box models [Rudin and Radin 2019].
- (3) Active and passive user input needs to be interpreted carefully to establish an efficient and effective interaction between humans and an AI system. The challenge includes to interpret signals from multiple input modalities (e.g., gaze and spoken instructions). It may be required to interpret the input signals according to a user or context model (e.g., reflecting a user’s preferences or the interaction context). In No-IDLE, we develop multimodal interaction techniques for incremental photo book creation with the goal to improve model training through rich multimodal user feedback and to improve the user experience through robust and intuitive interfaces. At the same time, we should avoid the limitations of human cognitive abilities.

- (4) Implement mixed initiative interaction, an opportunity to explore interfaces that can leverage knowledge and capabilities of domain experts more efficiently and effectively. The ML system and the domain expert should engage in a two-way dialogue to facilitate more accurate learning from less data compared to the classical approach of passively observing labelled data. In the context of our photo book use case, we aim at using, e.g., active learning and principles from human-in-the-loop expert systems. The greater goal is to perform application tasks more satisfactorily: human-machine teams shall surpass the efficiency/effectiveness of humans or machines in this task alone [van Zoelen et al. 2023].

3.1 Natural Language Processing (NLP)

Our approach for supporting photo book creation relies on several components based on deep learning models for image and multimedia data, in particular face and body shape recognition, text-to-image retrieval, image captioning, visual storytelling, and Visual Question Answering (VQA) models. The different components are triggered based on a user's commands (e.g., "On the first day, we took the bus from the airport to Vancouver" triggers the text-to-image retrieval component and the image captioning component). We plan to model this by either explicitly mapping triggering keywords to components, or by applying more sophisticated semantic parsers. The optimal way of processing user input will be determined in the course of the project based on insights from user studies. In this part of the No-IDLE project, we investigate three core research problems associated with the application and interaction with DL models in the context of our use case: (1) how to adapt state-of-the-art multimedia DL models to process user-specific texts and images, which, in contrast to the generic data the models are usually applied to, requires to account for specific information related to the user and the events they want to present in their photo book; (2) how to improve the DL components based on user feedback collected in the refinement phase based on the IML paradigm; (3) how to use model explanations to achieve optimal interaction between user and model and best support the photo book creation process.

Users have a personal relationship with objects and concepts displayed in the images of their photo book, and providing support in the photo book creation process requires modelling image content from a user's perspective. For example, we need to take into account that a user will refer to named entities in an image by proper name rather than a common noun (*Mary* instead of *a woman*). In No-IDLE, we investigate how to adapt multimedia and multimodal DL models to account for such user-specific information. For cross-modal (text-to-image) retrieval, we plan to implement state-of-the-art DL retrieval models [Alikhani et al. 2022; Jia et al. 2021; Zhang et al. 2020], which retrieve items based on embedding similarities in a shared representation space, in combination with rule-based filters that take into account output from a person recognition model as well as available image metadata, such as time stamps and geolocation. For example, given a user query *Show me the pictures of Peter and Mary playing football when we visited Vancouver*, the component retrieves images given the query *Two people playing football* and returns the subset of images for which the person recognition model indicates Peter and Mary being present, and the geolocation indicates an image taken in Vancouver. In contrast to image captions that can be found in general purpose datasets such as MS COCO [Lin et al. 2014] or Flickr30k [Plummer et al. 2015], the captions generated by our captioning component should be (1) entity-aware (e.g., instead of generic descriptions of objects or concepts, the captions contain proper names for named entities), (2) stylised, and (3) controllable (see table 1 for examples). Existing models for entity-aware captioning usually first generate a template caption with place-holders for named entities, which is then filled with information retrieved from associated text or knowledge bases [Biten et al. 2019; Lu et al. 2018]. Ramnath et al. [Ramnath et al. 2014] propose an approach for personalised template-filling with information such as geolocation, time stamp, detected landmarks, recognised faces, which we plan to extend to incorporate finer-grained location information specified by the user. To

generate stylised captions, we will explore caption generation reflecting sentiment [Mathews et al. 2016], specific styles [Gan et al. 2017; Guo et al. 2019], and taking into account a user’s active vocabulary [Chunseong Park et al. 2017]. In the refinement phase, when additional captions are generated for newly retrieved images, the user should be able to exert fine-grained control over the concepts to be included in the caption, e.g., by actively modifying an abstract scene graph representation based on which the caption is generated [Chen et al. 2020]. In contrast to generating captions for images in isolation, the visual storytelling component generates a sequence of captions that form a coherent story for a retrieved sequence of images [Huang et al. 2016; Jung et al. 2020; Wang et al. 2020]. Similar to the captioning component, the visual story component needs to be entity-aware and controllable. To this end, we will investigate to what extent approaches for adapting the captioning model can be transferred to the visual storytelling task. Finally, in the refinement phase, a VQA component can directly answer the user’s questions about image content, such as *What was the name of the mountain in the background?*, or *Did Peter join us for the trip to Lake Baikal?*. Here, we will focus on implementing models for answering questions that cannot be answered from information in the image alone, but require additional knowledge about named entities and specific events, that could for example be provided by a knowledge graph [Shah et al. 2019].

In order to improve the above described components based on feedback collected in the photo book refinement phase, we implement an IML framework that allows us to iteratively update the models based on new information via incremental and focused updates [Amershi et al. 2014]. Training and improving the models in an IML framework is crucial to our use case, as we cannot assume large amounts of labelled personalised data to be available at once, and therefore need to learn from user-specific data incrementally. In No-IDLE, we explore how IML can be applied to improve the multimodal DL components for photo book creation, considering three scenarios: (1) debugging trained models, e.g., identifying and correcting spurious patterns learned by the model [Lertvittayakumjorn and Toni 2021]. Here, we assume an explanation-based interactive loop to be particularly helpful; (2) adapting pre-trained models to user-specific data with small amounts of annotations [Yao et al. 2021] (3) personalising models [Kulesza et al. 2015], e.g., for generating captions following stylistic preferences of users. We focus on improving models based on explanatory feedback provided by the user, i.e., instead of providing only label-level feedback (e.g., a correct answer to a VQA model), the user additionally provides information that states *why* the provided answer is the correct one. Interacting on the basis of explanations has the potential to benefit both the user and the model: on the user side, providing richer feedback beyond the label level is in line with their preferred way of interaction [Amershi et al. 2014; Ghai et al. 2021]. From the modelling perspective, learning from explanatory feedback instead of label-level feedback can improve data efficiency [Hancock et al. 2018; Ye et al. 2020] and generalisation [Yao et al. 2021]. We focus on the two most commonly considered types of human explanations, which are *highlight explanations*, i.e., subsets of input elements deemed relevant for assigning a specific label; and *free-text explanations*, i.e., natural language statements providing information about why specific label should be assigned [Wiegrefe and Marasovic 2021]. Several ways for improving models (except for [Selvaraju et al. 2019] these were developed for models that process either text or image data) based on such human explanations have been proposed [Hartmann et al. 2021; Hase and Bansal 2021]: using natural language explanations as additional inputs [Co-Reyes et al. 2019; Rajani et al. 2019; Rupprecht et al. 2018], using explanation generation as auxiliary task [Camburu et al. 2018; Hase et al. 2020; Narang et al. 2020; Wiegrefe et al. 2021], directly constraining intermediate representations [Rieger et al. 2020; Ross et al. 2017; Selvaraju et al. 2019; Shao et al. 2021], or exploiting explanations to generate additional training instances [Awasthi et al. 2020; Hancock et al. 2018; Yao et al. 2021; Ye et al. 2020]. We will investigate how to combine and extend these methods to update multimodal DL models based on multimodal feedback. Most of these approaches have only been tested in offline setups, where the model can be

trained on the entire explanatory feedback at once. As a first step, we will investigate which methods are applicable in an interactive setup where models are updated incrementally. As all DL components process the same user-specific data, we assume that it might be useful to share user-specific information among the components by exploiting user feedback to update multiple components at once. To this end, we will experiment with a multi-task architecture with hard parameter sharing, which trains n models for n tasks with a subset of parameters being shared among them [Caruana 1993; Collobert et al. 2011], e.g., sharing the multi-modal encoder while maintaining task-specific classifier layers (or decoders for language generation tasks). By updating the encoder based on feedback collected for one task, the information will be available to models for the other tasks as well. For evaluating our methods for interactive deep learning, we will follow previous work in re-splitting existing task-specific datasets (e.g., Microsoft COCO [Lin et al. 2014] and Flickr30k [Plummer et al. 2015] for image captioning and text-to-image retrieval, VQA_{v2} [Goyal et al. 2017] and KB-VQA [Wang et al. 2017] for visual question answering, VIST [Huang et al. 2016] for visual story telling) into new data splits that allow to evaluate specific model behaviour, e.g., if a model relies less on language bias [Agrawal et al. 2018], or if a model has better continual learning abilities [Del Chiaro et al. 2020; Greco et al. 2019].

The central component of an IML system is a tight interactive loop between user and ML model, in which the model presents its current state of knowledge to the user, and the user provides feedback to the model accordingly [Amershi et al. 2011; Dudley and Kristensson 2018; Wang et al. 2021]. The former part of the loop could be supported by showing an explanation for why the model made a specific prediction or took a specific action. The ability to provide explanations for predictions, i.e., information about the reasons for why a specific prediction was made, is considered essential for large-scale adoption of AI systems by end-users [Barredo Arrieta et al. 2020; Gunning 2017]. In No-IDLE, we will investigate how to use model explanations to achieve optimal interaction between user and model. For DL black-box models, this requires choosing an adequate mechanism to construct explicit representations of explanations that can be provided to the user [Kim et al. 2021]. While for image processing models, saliency methods can provide useful visualisations of important input regions, such methods are less intuitive for text inputs. Here, the compositional nature of language calls for more expressive attribution methods that can model interactions between input tokens [Bastings and Filippova 2020]. We focus on the generation of suitable explanations for generative or predictive multi-modal tasks, e.g., by generating natural language explanations while at the same time marking image regions that were relevant for a prediction [Park et al. 2018a]. For presenting the explanation to the target end-user, we investigate the personalisation of explanations [Ghai et al. 2021; Mohseni et al. 2021; Ras et al. 2018; Sokol and Flach 2020; Tomsett et al. 2018] to elicit high quality feedback and increase user satisfaction. How to evaluate model explanations is an active research topic [DeYoung et al. 2020; Doshi-Velez and Kim 2017; Jacovi and Goldberg 2020; Pruthi et al. 2022] and we will focus on using previously proposed metrics for comparing model-generated explanations with human-generated explanations on publicly available multi-modal datasets, in particular VQA-X and e-ViL [Kayser et al. 2021; Park et al. 2018b].

The main deliverables of this part of the project are:

- (1) Implementation of multi-modal DL components for photo book creation support that are entity-aware and controllable. For image captioning and visual story telling, the components should be able to generate text in a specific style.
- (2) Implementation of an IML framework which allows to update the DL components based on explanatory user feedback collected in the photo book refinement phase. In addition to learning from explanatory feedback, the model should retain its knowledge while learning new things, which calls for the application of continual learning methods [Biesialska et al. 2020; d’Autume et al. 2019; Li et al. 2020] within the feedback loop to prevent

catastrophic forgetting [Kirkpatrick et al. 2017]. Incompleteness and uncertainty of human explanations [Tan 2021] should be accounted for when implementing a feedback mechanism into the model as a software package. To this end, we will build on insights from the core ML part of the project that investigates the use of Bayesian modelling for feedback integration as described in section 3.3.

- (3) Implementation of XAI methods for multimodal models which provide explanations for black box DL model decisions and take into account user-specific information, e.g., background knowledge and the motivation for consuming the explanation.

3.2 Multimodal-Multisensor Interaction (MMI)

In No-IDLE, we aim at developing interactive training mechanisms that enable continuous improvements of DL models. A central aspect of this interactive loop is human feedback. We investigate the effect of integrating multimodal user input on the effectiveness, efficiency, and usability of interactive model training. We target models of our photo book application which include, for instance, models for recognizing specific persons and objects (see section 3.3) and natural language generation models (see section 3.1).

One goal is to implement a gaze-driven dialogue that can support the initial creation and iterative refinement of a photo book. The multimodal feedback from the user shall enable the underlying DL models to learn new concepts, to differentiate between instances of a concept, and to improve the detection/recognition of know classes. We plan to implement simple state-based dialogues to realise interactive model training with human gaze as additional input modality (e.g., based on the open source dialogue platform Rasa²). The goal is not to develop beyond state-of-the-art multimodal dialogue systems, but to investigate the effect of integrating gaze (or pointing gestures) in simple speech-based instructions on the usability and effectiveness of interactive machine learning systems. For instance, a face recognition model could wrongly detect Sarah as Mary as described in section 2. When the user detects that the person identification system failed, he could provide a corrective feedback in natural language: "This is Sarah, not Mary". The system should analyse the user's gaze to identify to which face he referred in his utterance. Figure 3 illustrates this interaction based on an existing demo setup with three wall-sized screens. This corrective feedback shall be used to improve the underlying deep learning models (see figure 2). While, in No-IDLE, we put a focus on gaze-based input, pointing gestures will be considered for this kind of reference resolution as well, especially in the context of AR/VR interaction settings or when interacting with a wall-sized screen. Also, multimodal interaction can benefit from system-initiated interaction. This is particularly interesting in combination with active learning techniques that shall be developed by the ML group (see section 3.3). We want to explore the effectiveness (does the system actually learn to recognise new persons and objects), efficiency (what time is required for the model until it can recognise a new class), and usability (is the system usable for lay users) of different approaches in collaboration with the HCI group (see section 3.4). Another goal is to produce captions that are more focused on what the user wants to describe. We plan to integrate aggregated [Cornia et al. 2018; Sugano and Bulling 2016] or sequential [Meng et al. 2021; Pont-Tuset et al. 2020; Takmaz et al. 2020] human attention traces estimated from the multimodal input signal (gaze and pointing) into the generation process. We hypothesise that incorporating multimodal interaction signals can improve the robustness of and the user experience during the interaction with an interactive machine learning system. Eventually, this should improve the quality of human feedback and, hence, the efficiency of model updates during training. Also, we expect

²<https://rasa.com/open-source/>

that multimodal interaction can lead to a better understanding of how a model works, to a better understanding of the model’s strengths and weaknesses, and eventually to more trust in the model’s decisions.

Human gaze is well known for carrying non-verbal cues that can be used intelligent user interfaces: the eye movement behaviour depends on the task in which a user is currently engaged [DeAngelus and Pelz 2009], which provides an implicit insight into their intentions and allows an external observer or intelligent user interface to make predictions about the ongoing activity [Flanagan and Johansson 2003; Gredebäck and Falck-Ytter 2015; Rothkopf et al. 2016; Rotman et al. 2006]. For instance, knowing which objects in a scene are fixated is a valuable context information for spoken feedback in personalised photo book creation. In particular, when deictic references must be resolved [Matuszek 2018; Mehlmann et al. 2014]. Also, there is a strong link between gaze behaviour and spoken language: speakers fixate elements “less than a second before naming them” [Griffin and Bock 2000] and the coordination of hand-movements depends on human vision, e.g., when “directing the hand or object in the hand to a new location” [Land et al. 1999]. Human gaze can also be used to analyse or model the behaviour of a user (user modelling), e.g., to learn about a user’s ongoing activity [Bulling et al. 2013; Steil and Bulling 2015], their preferences [Barz et al. 2022; Lallé et al. 2021], intentions [Barz et al. 2020b; Huang and Mutlu 2016], or state [Bulling and Zander 2014; Huang et al. 2019]. Observing eye movement behaviour during interaction with an interactive machine learning system could reveal situations in which the user disagrees with the model output. If these situations coincide with the model being uncertain about the output, this may be a good point in time to trigger a feedback request to the user (system-initiative).

In No-IDLE, we focus on human gaze and pointing gestures as additional interaction modalities. We investigate the impact of using multimodal interaction signals on recognising objects or persons as context-information and to personalise the natural language generation process in the context of the photo book creation and refinement process. The challenge is that relevant persons and objects, their appearance, or similar properties can significantly vary between users and the occasion for creating such a book [Barz and Sonntag 2021]. However, pre-trained models cannot account for such dynamic circumstances and adaptive models or agents are required that incrementally and continuously learn from human collaborators or interlocutors. The main deliverable is a software extension of an existing DFKI system, the multisensor-pipeline (MSP)³. The resulting modules shall be integrated and evaluated in the photo book creation process based on the experimental procedure as depicted in section 3.5:

- (1) Implementation of a module that enables to learn about unseen classes (objects) when the context shifts (class-incremental learning) and to improve the recognition of known classes via multimodal user interaction based on, e.g., transfer learning [Käding et al. 2017] and active learning (see section 3.3). Similarly, we plan the implementation of a module to differentiate between multiple instances of the same class through multimodal human-machine interaction. We focus on the differentiation between multiple persons according to our photo book use case (e.g., to filter for images showing a particular person). This part will benefit from novel active learning approaches as described in section 3.3. We will also investigate in how far this module can be used to track meta information like ownership (see the COPDA project⁴). Real-time tracking of multiple instances could be achieved by a combination of (multi-)object tracking [Li et al. 2019] and models that estimate object properties such as colour, size, and shape [Thomason et al. 2016]. Such models are of particular interest when grounded in natural language, which would facilitate expressive explanations for classification results (related to section 3.1 and the XAINES project⁵).

³<https://github.com/DFKI-Interactive-Machine-Learning/multisensor-pipeline>

⁴<https://www.dfki.de/en/web/research/projects-and-publications/projects-overview/project/copda>

⁵<https://www.dfki.de/en/web/research/projects-and-publications/projects-overview/project/xaines>

- (2) Implementation of a module of a new image clustering and object tracking method that can help "quick start" multimodal interactive model training upon domain shifts, because a single (user-provided) label can be propagated to multiple samples, e.g., to an image cluster or to samples from object tracking (semi-supervised learning). The idea is to cluster fixated image contents in the photo book application and, once a label is provided via speech, to propagate this label to the whole cluster. Similarly, **few shot learning (FSL)** from the ML task (see section 3.3) should help to overcome this cold-start problem. Few-shot image classification [Wertheimer et al. 2021] or few-shot object detection [Fan et al. 2019] enables image classification or object detection, respectively, using around five example images.
- (3) Implementation of multimodal interaction techniques based on eye tracking for the photo book application. This includes approaches to provide feedback on model outputs multimodally (e.g., correcting labels for misclassified persons, triggering post-editing of generated captions, and guiding the caption generation process), but also general multimodal interaction with photo book representations in desktop or VR settings (for instance rearranging images, selecting better photos, or similar selection and manipulation actions).

3.3 Machine Learning (ML)

An important factor which contributes to the recent success of **DL** (apart from superior computing power and training algorithms) is the availability of labelled data. In fact, neural networks are known to be data-hungry (e.g., popular benchmark datasets range from tens of thousands of labelled samples as in CIFAR-10 to millions as in ImageNet dataset). However, data labelling is a costly, human labour intensive activity. In certain domains such as healthcare and biomedicine where considerable expertise may be required, data labelling becomes a limiting step in the realisation of the value of **ML**. This is also the case for the creation of personalised photo books. For instance, when the system should learn to differentiate between faces and body shapes of a set of persons in order to select images containing them or not while the persons may differ per user and photo book. Thus, it is imperative to build **ML** algorithms which are capable of learning from significantly fewer labelled samples to save human time.

A set of methods known as active learning [Monarch 2021; Settles 2010] tackle this problem by allowing the system to identify a subset of maximally informative samples from a given pool of unlabelled data to be queried for additional labelling/feedback. In the context of **IML** in this proposal, active learning plays a key role in how a learning system requests, receives, and learns from user input. In combination with the **HCI** tasks (section 3.4), this forms a joint task for mixed-initiative interaction: **ML** system and human domain expert engage in a two-way dialogue, facilitating learning from less data compared to the classical approach of passive consumption of labelled data. One direction to explore are new input techniques that allow users to provide more informative feedback [Ratner et al. 2016], compared to traditional low dimensional labels.

Popular methods in active learning might be uncertainty-based [Joshi et al. 2009; Konyushkova et al. 2019; Tong and Koller 2001], density- or diversity-based approaches [Gissin and Shalev-Shwartz 2019; Sourati et al. 2018], ensemble methods [Beluch et al. 2018; Freund et al. 1997; McCallumzy and Nigamy 1998], and expected error reduction [Roy and McCallum 2001]. A common problem of pure uncertainty-based methods is that the selection strategy depends on the performance of an existing model. This could be problematic in the early phase of training since outcomes are likely to be unreliable, leading the algorithm to query poor examples and thus lead to inefficiencies. Similarly, in pure density-based approaches data labelling could be redundant if the present model produces already high confident predictions. Recently, methods have been proposed which try to mitigate this problem by combining and balancing uncertainty and diversity of the new samples w.r.t. the data distribution [Ash et al. 2020; Huang et al. 2010; Ozdemir

et al. 2018; Smailagic et al. 2018; Yang et al. 2017]. Bayesian approaches have also been proposed [Gal et al. 2017; Kapoor et al. 2007; Kirsch et al. 2019], but they do not scale well to deep networks with large datasets. Other recent works include Fisher information [Ash et al. 2021; Sourati et al. 2018] and learning to select from data [Konyushkova et al. 2017].

We tailor active learning technologies to be applied in No-IDLE in the context of our photo book scenario. The goal is to train a model that is able to differentiate between individual persons contained in a set of photos with little labelling effort by the user. The basis for this feature are computer vision models that enable a robust detection and location of faces and body shapes. For any set of images, these models can provide a pool of unlabelled face and body images. This is helpful to filter for images showing humans versus, e.g., landscape photos. However, for personalised photo books, we want the system to be able to differentiate between individual persons to filter for photos with specific persons. For instance, a user request could be “please add an image of Mary in front of our rental car”. The persons involved may vary as they are highly dependent on the user and the occasion for creating the photo book. We will (1) implement and evaluate new sampling techniques/active learning approaches that enable model training with small amounts of labelled data and (2) investigate when system-initiative feedback requests should be shown and how they should be designed in order to maintain a good user experience. A good opportunity to trigger a feedback request could be right after a user takes the initiative to provide a new name (i.e., a label) for a person/face or corrects a label. For instance, if a user tells the system “this is Mary”, the system could query for the most informative unlabelled instances that may also show Mary like “Ah, this is Mary. I guess, I’ve seen her on other pictures too. Is this Mary again [system shows another face image]?”.

In this proposal, we aim to address the following ML problems:

- (1) On the experimental side, we first investigate the performance of existing uncertainty functions for various neural network architectures on image classification/segmentation tasks (see, e.g., figure 5).
- (2) On the experimental side, this point is related to studying if we should only use the DL black box models in the IML process when we perhaps do not need to. The point brought forward in [Rudin and Radin 2019] is that one might consider that (in IML) maybe interpretable deep-learning models can be constructed, or transparent models be used in conjunction with DL models according to the user feedback. In machine learning, these black box models are created directly from data by an algorithm, meaning that humans, even those who design them, cannot understand how variables are being combined to make predictions. Also see surrogate models for this purpose, co-creating a transparent model from the predictions. A global surrogate model is an interpretable model that is trained to approximate the predictions of a black box model. We can draw conclusions about the DL black box model by interpreting the surrogate model [Burkart and Huber 2021].
- (3) On the practical side including Few-Shot-Learning: the motivation for this ML task comes from MMI (section 3.2), where we want the system to learn new objects during an interactive training session with the user, given that the user has provided feedback/labels for a few examples. In the literature, this problem could be tackled using techniques from FSL [Tian et al. 2020]. The main challenge is how to learn a good latent embeddings of the inputs and the labels, and to align them together in such a way that certain attributes from both inputs and labels can be transferred to unseen objects.

The research outcomes and main deliverables will include the design of new uncertainty functions, which will be used in IML-related tasks such as NLP (section 3.1) and MMI (section 3.2). Additionally, together with HCI (section 3.4) we will promote an active role for the human-in-the-loop: besides providing labels, we want to explore different

ways of providing/correcting explanations, aligning important features learned by the machine with human intuition, interpreting learned models, and finding a common ground with general HCI tasks, including a more generic approach for generating explanations and insights into the effectiveness of few-shot learning.

3.4 Human-Computer Interaction (HCI)

We explore the role of humans in IDL. From our own previous work [Herrlich et al. 2017] and from the literature [Oviatt 2006; Picard 2000; Ryan and Deci 2000], the relevance of motivation, emotion and factors like cognitive load on how interfaces and systems are used and, consequently, how these factors should be taken into account during interface design is quite clear. IDL presents both a potential solution and an additional challenge in this regard [Amershi et al. 2014]. Furthermore, we want to transfer insights from our previous works in the medical domain and virtual reality. We have studied expert users such as medical doctors⁶ and explored VR for IDL, e.g., for image classification in VR [Prange and Sonntag 2021], and as a general prototyping and evaluation environment for human-centered interaction design [Klonig and Herrlich 2020; Omar Jubran et al. 2021; Queck et al. 2022; Reinschluessel et al. 2017; Vera Eymann et al. 2021].

Referring to the example “photo book” application scenario described above, we plan to explore the combination of VR and IDL as a multi-modal, immersive interaction environment. This environment supports rich data input signals, for example, gaze and eye tracking, tracking of spatial movements and features such as pointing using a controller or freehand gestures and recording 3D trajectories over time as well as audio and speech input. It also integrates multi-modal output signals in the form of 3D graphics, spatial audio and simple forms of tactile feedback. Last but not least, it provides unlimited virtual space. As we sketched in the application scenario, we want to investigate how to leverage the potential of VR for IDL but the VR environment also provides an ideal test bed for generating and comparing data and models to be used in the real world because it is much easier to control and deploy. While existing works in this area have investigated specific components and tasks of the example usage scenario, e.g., the selection of aesthetically pleasing photos [Withöft et al. 2022], taking a specific look at the human factors with respect to the rich input and output modalities within virtual reality is a novel idea and has not been explored in the context of IDL to the best of our knowledge.

From an HCI perspective, the goals can be summarised as exploring new ways for learning systems to interact with their users, namely: (1) how user-driven learning cycles can involve more rapid, focused, and incremental model updates; (2) how to reduce the need for supervision by ML practitioners; (3) As a result of these rapid interaction cycles common in IML, even users with little or no machine-learning expertise should be able to steer machine-learning behaviours through low-cost trial and error or focused experimentation with inputs and outputs. How can this be supported from the HCI perspective? (4) Transparency can help provide better labels (contextual features, ML predictions, etc.) towards explainable IML. The experimental setup should include explainable IML, where the user feedback is derived after the system explains its results, to avoid “right answers for the wrong reasons”, see, e.g., [Anders et al. 2022]. (5) Understanding how people actually interact—and want to interact—with machine-learning systems is critical to designing systems that people can use effectively [Simard et al. 2017].

More specifically, we plan to study basic properties like mental and physical load, attention split problems, confusion, and emotional affect. These provide the foundation to investigate more complex effects regarding user intention and strategies, trust, and confidence in using the system. Furthermore, we expect a large impact of explainability techniques

⁶https://medicalcps.dfki.de/wp-content/uploads/2017/08/KDI_V2_Pro_v04_2.mp4

on these factors. We plan to experiment with different graphical and textual or spoken explanations. By studying these factors from the user's perspective we intend to optimise the effectiveness of active learning techniques.

We plan to run comparative studies within VR, for example, exploring different interaction designs, information presentation and DL techniques. The idea is to measure human factors as listed above, e.g., cognitive load, but also other factors of the user experience, such as emotional affect and motivational measures such as user engagement and study their impact on active learning efficiency and effectiveness.

Considering the potential effect of user motivation, experimenting with forms of gamification [Deterding et al. 2011a,b] and serious games within the framework of the example scenario seems relevant. One approach in that regard will be to turn the respective task, e.g., finding photos with certain contents, describing a picture, sorting or clustering pictures, inserting a missing or best fitting picture into visual photo book story, into challenges by introducing a time limit (soft or hard), rewards (short, mid, long term) and potentially forms of social relatedness (synchronous or asynchronous forms of multi-player). Gamification could also be used to provide a measurement of the quality of the DL model by using it to acquire user ratings of the overall output.

As a side note, to facilitate user participation in our experiments, we plan to set up an open lab space in the centre of the city of Oldenburg (in the CORE Oldenburg) to increase participation and recruit volunteers with diverse demographic backgrounds.

The main deliverables in this area are:

- (1) Implementation of different interaction modalities within virtual reality, e.g., free hand gestures vs. controller based selection or manipulation vs. NLP and possible combinations.
- (2) Studies about the influence of conscious and unconscious gestures, e.g., certain movements or posture that relate to confusion or decision insecurity; gaze or eye tracking (here there is a very strong link to multimodality).
- (3) Implementation of different feedback forms and modalities to encode information about the DL results and decision process, from "simple" visual features (colour, location, etc.) to audio or tactile channels.
- (4) Concepts and studies of the effect of more playful approaches (serious games and gamification) with respect to user motivation and user feedback quality and quantity for IDL.

3.5 Evaluation Plan

In this subsection we provide details about our general evaluation process and study plan. Of course, due to the novelty of the research, the plan will have to be adjusted throughout the project as it depends on the progress and results of the technical parts and work packages. We want to emphasise that the guiding overall focus of all evaluation activity is to investigate and improve the IDL process as discussed in the specific subsections, e.g., how can the observed user behaviour and user experience be utilised as a means for improving efficiency and effectiveness of IDL. This also is reflected in the way that VR is used within this project, i.e., as powerful tool for studying user behaviour and collecting data using photo book creation as an example application as opposed to investigating the use of VR for photo book creation, which is explicitly not a focus point of this project.

Firstly, we plan to conduct a number of smaller studies that look at very specific aspects and that lay the foundation for a larger study towards the end of the project. At the beginning, we will focus on fundamentals and isolated elements and shift to investigating more complex combinations of system features and tasks over time. This will also be reflected in the methods we apply. At the beginning we will employ methods of a more exploratory and formative type, for

instance, case studies using methods such as interviews, cognitive walk-troughs, think-aloud, observation and forms of moderated discussion. Of course, this does not exclude also collecting quantitative data already in this phase if possible.

The main study approach of a more summative character will be using an experimental setup comparing two conditions (control + intervention) or (if applicable) a factorial design with up to three or four conditions using appropriate tools and collecting quantitative measures like completion times, labelling accuracy in addition to (preferably validated) questionnaires for subjective feedback especially for measuring user experience and usability, e.g., SUS [Brooke 1986], PANAS-X [Watson and Clark 1994] and other SDT-based [Ryan and Deci 2000] tools related to motivation and also physical and mental load (e.g., NASA-TLX [Hart 2006; Hart and Staveland 1988]).

The final decision for the experimental design with respect to independent or dependent groups (within-subjects vs. between subjects design) hinges on factors like the expected learning effect vs. fatigue effects and is subject to the specific experimental design for each study based on testing and pre-studies to quantify these confounding effects.

In addition, the VR setup in particular but also the eye-tracking scenario provide unique opportunities to collect objective data, most importantly, eye-tracking and movement data, e.g., trajectories of the controllers. We will also look into additional psycho-physiological measures, such as heart rate that are relatively easy to measure with off-the-shelf wearables.

We will base the number of participants on comparable studies and standards in HCI, typically in the range of 20-80 participants per experiment. The general experimental procedure includes the following steps:

- (1) Introduction and welcome of participants and collecting their informed consent.
- (2) A training or accommodation phase, which is especially important in the VR case.
- (3) A calibration phase or procedure, which can also include collecting base levels of certain measures.
- (4) The main part, i.e., participants perform specified tasks under different conditions, e.g., different forms of visual feedback, input gestures or active learning prompts. Some data are collected continuously through logging other data (e.g., subjective feedback) are collected after each condition (in accordance to the respective measure or questionnaire).
- (5) Collection of post-experimental and independent data (e.g., demographics).
- (6) De-briefing and “Goodbye”.

Throughout the procedure participants will be able to take breaks as needed (especially in the VR scenario) and we will adhere to scientific standards including getting approval of the DFKI ethics committee. The statistical analysis of individual measures will be carried out using linear models such as ANOVA for comparing means or non-parametric tests like Friedman [Cairns 2019]. In addition, forms of time series analysis and clustering will be looked into for analysing and correlating spatial measures such as body, hand, or controller movements. We will also consider post-hoc experiments based on recorded user inputs to test additional IML approaches. This can be done by simulating the interaction signals of our study participants if the model outputs have no immediate impact on the interaction flow.

4 EXISTING HARDWARE AND SOFTWARE FRAMEWORKS AT DFKI IML

By harnessing the power of foundation models [Ali et al. 2019], i.e., any ML model which is trained on a large-scale dataset and can be adapted to a wide range of downstream tasks, the research community is optimistic about their social applicability [Bommasani et al. 2021], especially in the healthcare discipline with integrated human interaction. Especially, patient care via disease treatment usually requires expert knowledge that is limited and expensive. Foundation models trained on the abundance of data across many modalities (e.g., images, text, molecules) present clear opportunities

to transfer knowledge learned from related domains to a specific domain and further improve efficiency in the adaptation step by reducing the cost of expert time. As a result, a fast prototype application can be employed without collecting significant amounts of data and training large models from scratch. In the opposite direction, end-users who will directly use or be influenced by these applications can provide feedback to power these foundation models toward creating tailored models for the desired goal of IDL, based on DFKI IML's existing software frameworks: [Nguyen et al. 2020; Nunnari and Sonntag 2021; Sonntag et al. 2020; Zacharias et al. 2018].

The planned multimodal multisensor interfaces in No-IDLE will be based on the multisensor-pipeline (MSP)⁷, our lightweight, flexible, and extensible framework for prototyping MMI based on real-time sensor input [Barz et al. 2021a]. The MSP ecosystem will benefit from the developments in No-IDLE, because novel modules will be released as open source to the research community. No-IDLE will take advantage from recent and upcoming developments in the BMBF Project GeAR⁸ (ends in September 2022): we are developing methods that reduce the human effort in the process of annotating mobile eye tracking data as described in [Barz and Sonntag 2021]. In GeAR, we target semi-automatic annotation for analytical applications (post-hoc) rather than real-time interactive model training, which is integrated into the application itself.

5 EXISTING APPLICATION DOMAINS AND DEMO SCENARIOS AT DFKI IML

We build the MMI and HCI components of this project upon four past application domains and demo scenarios, which we detail in the respective figure captions:

- Interactive Doctor Feedback (use case from BMBF Ophthalmology-AI⁹) project (see figure 5)
- Interactive Image Classification in VR (see figure 6)
- Explanatory IML (use case from XAINES project, see figure 7): In XAINES, we develop models that provide explanations for predictions in an explanation-feedback loop, which can serve to improve the model based on human feedback, and to personalize explanations. These models will serve as a starting point for developing interactive DL models for the No-IDLE photo book use case.
- The multimodal interaction systems in No-IDLE will be build based on our experience and outcomes from recent research projects (SciBot, GeAR). This includes methods for real-time interpretation of multimodal sensor streams such as mobile eye tracking data [Barz et al. 2022, 2021b; Barz and Sonntag 2021; Barz et al. 2020b; Bhatti et al. 2021; Kapp et al. 2021] (for an example, see figure 8), but also pen-based input signals [Barz et al. 2020a]. In addition, we will use and further develop our framework for building multimodal, real-time interactive interfaces, the *multisensor-pipeline* [Barz et al. 2021a].

6 CONCLUSION

We presented the anatomy of the No-IDLE prototype system (funded by the German Federal Ministry of Education and Research) and described basic and fundamental research in interactive machine learning while addressing users' behaviours, needs, and goals. We described goals and scientific challenges that centre around the desire to increase the reach of interactive deep learning solutions for non-experts in machine learning, followed by a methodology for interactive machine learning combined with multimodal interaction which will become central when we start interacting with semi-intelligent machines in the upcoming area of neural networks and large language models. Future work

⁷<https://github.com/DFKI-Interactive-Machine-Learning/multisensor-pipeline>

⁸<https://www.dfki.de/en/web/research/projects-and-publications/projects-overview/project/gear>

⁹<https://www.dfki.de/en/web/research/projects-and-publications/projects-overview/project/ophthalmo-ai>

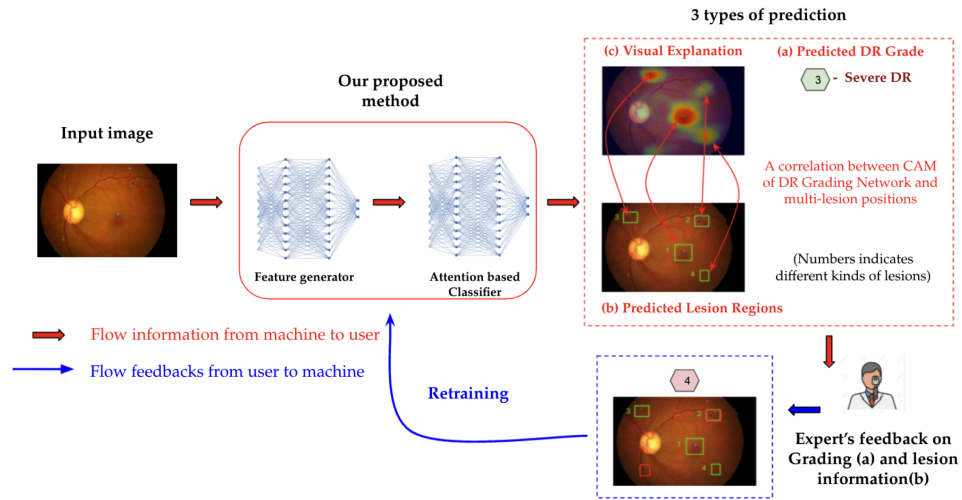


Fig. 5. High level overview of our proposed method in the IDL workflow of the Ophthalmo-AI project (BMBF). Given a retinal image, our DL models will generate 3 types of predictions (DR grade, lesion region, visual explanation) simultaneously. Ophthalmologists can observe the predictions and provide feedback for model fine-tuning.

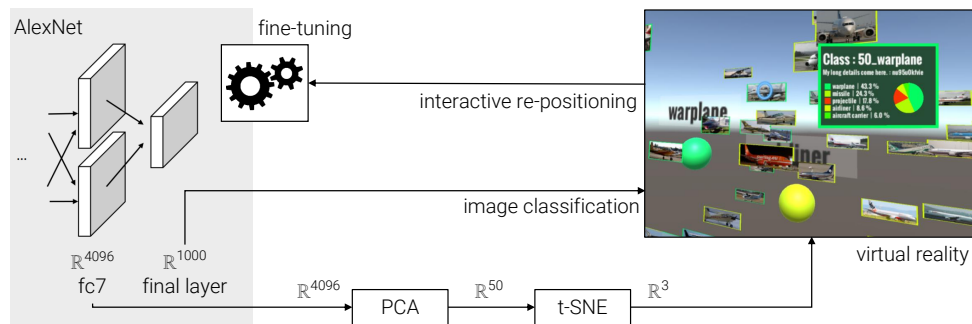


Fig. 6. Architecture of our approach in [Prange and Sonntag 2021] based on PCA and t-SNE dimensionality reduction. Based on a pre-trained AlexNet we calculate 3D coordinates for each image. In VR, information related to a particular image is displayed if the user looks at it.

includes "No-IDLE meets ChatGPT". The overall objective of this follow-up project will be to leverage the opportunities arising from large language models and technologies for the No-IDLE project. No-IDLE aims to enhance the interaction between humans and machines for the purpose of updating deep learning models, integrating cutting-edge human-computer interaction techniques and advanced deep learning approaches. Considering the recent advances in LLMs and their multimodal capabilities, the overall objective of "No-IDLE meets ChatGPT" should be well motivated.

ACKNOWLEDGMENTS

This work is funded by the German Federal Ministry of Education and Research under grant number 01IW23002.


Task	Natural Language Inference (NLI)	Commonsense Reasoning	Visual Question Answering (VQA)
Data	<p><i>P:</i> A 2-3 year old blond child is kneeling on a couch.</p> <p><i>H:</i> The child has brown hair.</p> <p>→ Contradiction</p>	<p>Question: What would not be true about a basketball if it had a hole in it but it did not lose its general shape?</p> <p>A: punctured, B: full of air, C: round</p> <p>→ Answer B</p>	<p>Image: </p> <p>Question: What is the person doing?</p> <p>→ Skiing</p>
Expl.	<p>The child would not have brown hair if he/she was blond.</p>	<p>Air cannot stay in any object that has a hole in it.</p>	<p>... because they are on skis and in a skiing outfit.</p>

Fig. 7. Examples of existing datasets with human explanations for natural language inference [Camburu et al. 2018], commonsense reasoning [Rajani et al. 2019], and visual question answering [Park et al. 2018a]. Explanations are either free-form (bottom line) or subsets of the input data (highlights in blue). These datasets can be used for both learning to generate natural language explanations as well as simulating explanatory feedback fed to the model in the sense of explanatory IML, see [Teso and Kersting 2019], where in each human-in-the-loop step, the learner explains its prediction to the user, and the user can provide explanatory feedback back to the model in order to improve it. Whereas explanatory IML mainly focuses on correcting *right for the wrong reason* behaviour, we will also explore how to use explanatory feedback to adapt models to user-specific input data.



Fig. 8. Our prototype based on Microsoft’s HoloLens 2 classifies and augments fixated objects in real-time [Barz et al. 2021b]. It displays classification labels and the duration of recent attention events to the user as a hologram. The demo video can be viewed here: <https://www.youtube.com/watch?v=bdNCIVz9yIE>. In No-IDLE, we plan to enable interactive model adaptation based on foundation models: For instance, the user could create a specific instance of "reflex camera" and name it "Nikon camera" via speech (as shown in the image). This is related the COPDA project which aims to establish and maintain object relations like ownership. Other examples include that users may correct wrong classifications or teach new classes to the ML system in a mixed-initiative dialogue.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don’t just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4971–4980.

Shahbaz Ali, Hailong Sun, and Yongwang Zhao. 2019. Model Learning: A Survey on Foundation, Tools and Applications. In *arXiv:1901.01910*.

Malihe Alikhani, Fangda Han, Hareesh Ravi, Mubbasir Kapadia, Vladimir Pavlovic, and Matthew Stone. 2022. Cross-Modal Coherence for Text-to-Image Retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 10 (Jun. 2022), 10427–10435. <https://doi.org/10.1609/aaai.v36i10.21285>

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.



Generic:	Two boys are playing frisbee on the beach.
Personalised:	Peter and Tom are playing frisbee at the Pyla campsite.
Stylised:	A heated game of frisbee on the Pyla court.
Controllable:	David has enough and returns to the cabin.

Table 1. Example image captions for an image taken from MS COCO, showing the difference between generic image captions and entity-aware, stylised, and controllable image captions require for photo book creation support.

- Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. 2011. Effective end-user interaction with machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 25.
- Christopher J. Anders, Leander Weber, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. 2022. Finding and removing Clever Hans: Using explanation methods to debug and improve deep models. *Inf. Fusion* 77 (2022), 261–295. <https://doi.org/10.1016/j.inffus.2021.07.015>
- Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. 2021. Gone Fishing: Neural Active Learning with Fisher Embeddings. In *NeurIPS*.
- Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. In *ICLR*.
- Abhijeet Awasthi, Sabyasachi Ghosh, Rasna Goyal, and Sunita Sarawagi. 2020. Learning from Rules Generalizing Labeled Exemplars. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=SkeuexBtDr>
- Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Michael Barz, Kristin Altmeyer, Sarah Malone, Luisa Lauer, and Daniel Sonntag. 2020a. Digital Pen Features Predict Task Difficulty and User Performance of Cognitive Tests. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2020, Genoa, Italy, July 12-18, 2020*, Tsvi Kuflik, Ilaria Torre, Robin Burke, and Cristina Gena (Eds.). ACM, 23–32. <https://doi.org/10.1145/3340631.3394839>
- Michael Barz, Omair Shahzad Bhatti, Bengt Lüers, Alexander Prange, and Daniel Sonntag. 2021a. Multisensor-Pipeline: A Lightweight, Flexible, and Extensible Framework for Building Multimodal-Multisensor Interfaces. In *Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 13–18. <https://doi.org/10.1145/3461615.3485432> event-place: Montreal, QC, Canada.
- Michael Barz, Omair Shahzad Bhatti, and Daniel Sonntag. 2022. Implicit Estimation of Paragraph Relevance From Eye Movements. *Frontiers in Computer Science* 3 (2022), 136. <https://doi.org/10.3389/fcomp.2021.808507>
- Michael Barz, Sebastian Kapp, Jochen Kuhn, and Daniel Sonntag. 2021b. Automatic Recognition and Augmentation of Attended Objects in Real-time using Eye Tracking and a Head-mounted Display. In *ACM Symposium on Eye Tracking Research and Applications (ETRA '21 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1–4. <https://doi.org/10.1145/3450341.3458766>
- Michael Barz and Daniel Sonntag. 2021. Automatic Visual Attention Detection for Mobile Eye Tracking Using Pre-Trained Computer Vision Models and Human Gaze. *Sensors* 21, 12 (Jan. 2021), 4143. <https://doi.org/10.3390/s21124143> Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.
- Michael Barz, Sven Stauden, and Daniel Sonntag. 2020b. Visual Search Target Inference in Natural Interaction Settings with Machine Learning. In *ACM Symposium on Eye Tracking Research and Applications (ETRA '20 Full Papers)*, Andreas Bulling, Anke Huckauf, Eakta Jain, Ralph Radach, and Daniel Weiskopf (Eds.). Association for Computing Machinery, 1–8. <https://doi.org/10.1145/3379155.3391314>
- Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Online, 149–155. <https://doi.org/10.18653/v1/2020.blackboxnlp-1.14>
- William H. Beluch, Tim Genewein, Andreas Nurnberger, and Jan M. Kohler. 2018. The power of ensembles for active learning in image classification. In *CVPR*.
- Omair Shahzad Bhatti, Michael Barz, and Daniel Sonntag. 2021. EyeLogin - Calibration-free Authentication Method for Public Displays Using Eye Gaze. In *ACM Symposium on Eye Tracking Research and Applications (ETRA '21 Short Papers)*. Association for Computing Machinery, New York, NY, USA, 1–7. <https://doi.org/10.1145/3448018.3458001>
- Magdalena Biesialska, Katarzyna Biesialska, and Marta R Costa-jussà. 2020. Continual Lifelong Learning in Natural Language Processing: A Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*. 6523–6541.
- Rajarshi Biswas, Michael Barz, and Daniel Sonntag. 2020. Towards Explanatory Interactive Image Captioning Using Top-Down and Bottom-Up Features, Beam Search and Re-ranking. *KI-Künstliche Intelligenz* 34, 4 (2020), 571–584.

- Ali Furkan Biten, Lluís Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12466–12475.
- Susanne Boll, Philipp Sandhaus, Ansgar Scherp, and Sabine Thieme. 2006. MetaXa—Context- and Content-Driven Metadata Enhancement for Personal Photo Books. In *Advances in Multimedia Modeling (Lecture Notes in Computer Science)*, Tat-Jen Cham, Jianfei Cai, Chitra Dorai, Deepu Rajan, Tat-Seng Chua, and Liang-Tien Chia (Eds.). Springer, Berlin, Heidelberg, 332–343. https://doi.org/10.1007/978-3-540-69423-6_33
- Susanne Boll, Philipp Sandhaus, Ansgar Scherp, and Utz Westermann. 2007. Semantics, content, and structure of many for the creation of personal photo albums. In *Proceedings of the 15th ACM international conference on Multimedia (MM '07)*. Association for Computing Machinery, New York, NY, USA, 641–650. <https://doi.org/10.1145/1291233.1291385>
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshthe Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the Opportunities and Risks of Foundation Models. *CoRR* abs/2108.07258 (2021). arXiv:2108.07258 <https://arxiv.org/abs/2108.07258>
- John Brooke. 1986. System usability scale (SUS): a quick-and-dirty method of system evaluation user information. *Reading, UK: Digital equipment co ltd* 43 (1986), 1–7.
- Andreas Bulling, Christian Weichel, and Hans Gellersen. 2013. EyeContext: Recognition of High-level Contextual Cues from Human Visual Behaviour. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 305–308. <https://doi.org/10.1145/2470654.2470697>
- Andreas Bulling and Thorsten O. Zander. 2014. Cognition-Aware Computing. *IEEE Pervasive Computing* 13, 3 (July 2014), 80–83. <https://doi.org/10.1109/MPRV.2014.42>
- Nadia Burkart and Marco F. Huber. 2021. A Survey on the Explainability of Supervised Machine Learning. *J. Artif. Intell. Res.* 70 (2021), 245–317. <https://doi.org/10.1613/jair.1.12228>
- Paul Cairns. 2019. *Doing better statistics in human-computer interaction*. Cambridge University Press.
- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. *Advances in Neural Information Processing Systems* 31 (2018), 9539–9549.
- R Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*. Citeseer, 41–48.
- Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. 2020. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9962–9971.
- Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 895–903.
- John D Co-Reyes, Abhishek Gupta, Suvansh Sanjeev, Nick Altieri, John DeNero, Pieter Abbeel, and Sergey Levine. 2019. Meta-Learning Language-Guided Policy Learning. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=HkgSEnA5KQ>
- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research* 12, ARTICLE (2011), 2493–2537.
- Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2018. Paying more attention to saliency: Image captioning with saliency and context attention. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 2 (2018), 1–21.
- Cyprien de Masson d’Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 13132–13141.
- Marianne DeAngelus and Jeff B. Pelz. 2009. Top-down control of eye movements: Yarus revisited. *Visual Cognition* 17, 6-7 (Aug. 2009), 790–811. <https://doi.org/10.1080/13506280902793843> Publisher: Routledge.
- Riccardo Del Chiaro, Bartłomiej Twardowski, Andrew Bagdanov, and Joost Van de Weijer. 2020. Ratt: Recurrent attention to transient tasks for continual image captioning. *Advances in Neural Information Processing Systems* 33 (2020), 16736–16748.
- Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011a. From game design elements to gamefulness: defining ‘gamification’. In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*. 9–15.
- Sebastian Deterding, Miguel Sicart, Lennart Nacke, Kenton O’Hara, and Dan Dixon. 2011b. Gamification. using game-design elements in non-gaming contexts. In *CHI’11 extended abstracts on human factors in computing systems*. 2425–2428.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4443–4458.
- C. Donalek, S. G. Djorgovski, A. Cioc, A. Wang, J. Zhang, E. Lawler, S. Yeh, A. Mahabal, M. Graham, A. Drake, S. Davidoff, J. S. Norris, and G. Longo. 2014. Immersive and collaborative data visualization using virtual reality platforms. In *2014 IEEE International Conference on Big Data (Big Data)*. 609–614.
- Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
- John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (June 2018). <https://doi.org/10.1145/3185517> Place: New York, NY, USA Publisher: Association for Computing Machinery.

- Qi Fan, Wei Zhuo, and Yu-Wing Tai. 2019. Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector. *CoRR* abs/1908.01998 (2019). <http://arxiv.org/abs/1908.01998> arXiv: 1908.01998.
- J. Randall Flanagan and Roland S. Johansson. 2003. Action plans used in action observation. *Nature* 424, 6950 (Aug. 2003), 769–771. <https://doi.org/10.1038/nature01861>
- Yoav Freund, H Sebastian Seung, Eli Shamir, , and Naftali Tishby. 1997. Selective sampling using the query by committee algorithm. *Machine Learning* 28 (1997), 2–3.
- Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *ICML*.
- Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3137–3146.
- Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.
- Daniel Gissin and Shai Shalev-Shwartz. 2019. Discriminative Active Learning. *CoRR* abs/1907.06347 (2019). arXiv:1907.06347 <http://arxiv.org/abs/1907.06347>
- Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. 2018. Explainable AI: The New 42?. In *Machine Learning and Knowledge Extraction*, Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl (Eds.). Springer International Publishing, Cham, 295–303.
- Thiago S. Gouvêa, Hannes Kath, Ilira Troshani, Bengt Lüers, Patricia P. Serafini, Ivan B. Campos, André S. Afonso, Sergio M. F. M. Leandro, Lourens Swanepoel, Nicholas Theron, Anthony M. Swemmer, and Daniel Sonntag. 2023. Interactive Machine Learning Solutions for Acoustic Monitoring of Animal Wildlife in Biosphere Reserves. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*. ijcai.org, 6405–6413. <https://doi.org/10.24963/IJCAI.2023/711>
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6904–6913.
- Claudio Greco, Barbara Plank, Raquel Fernández, and Raffaella Bernardi. 2019. Psycholinguistics Meets Continual Learning: Measuring Catastrophic Forgetting in Visual Question Answering. In *Proceedings of the Association for Computational Linguistics*. 3601–3605.
- Gustaf Gredebäck and Terje Falck-Ytter. 2015. Eye Movements During Action Observation. *Perspectives on Psychological Science* 10, 5 (2015), 591–598. <https://doi.org/10.1177/1745691615589103> _eprint: <https://doi.org/10.1177/1745691615589103>.
- Zenzi M. Griffin and Kathryn Bock. 2000. What the Eyes Say About Speaking. *Psychological Science* 11, 4 (2000), 274–279. <https://doi.org/10.1111/1467-9280.00255> _eprint: <https://doi.org/10.1111/1467-9280.00255>.
- David Gunning. 2017. Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web* 2, 2 (2017), 1.
- Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. 2019. Mscap: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4204–4213.
- Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training Classifiers with Natural Language Explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1884–1895.
- Sandra G. Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908. Issue: 9.
- Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.
- Mareike Hartmann, Ivana Kruijff-Korbayová, and Daniel Sonntag. 2021. Interaction with Explanations in the XAINES Project.
- Peter Hase and Mohit Bansal. 2021. When can models learn from explanations? a formal framework for understanding the roles of explanation data. *arXiv preprint arXiv:2102.02201* (2021).
- Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 4351–4367.
- Niels Henze and Susanne Boll. 2011. Who’s That Girl? Handheld Augmented Reality for Printed Photo Books. In *Human-Computer Interaction – INTERACT 2011 (Lecture Notes in Computer Science)*, Pedro Campos, Nicholas Graham, Joaquim Jorge, Nuno Nunes, Philippe Palanque, and Marco Winckler (Eds.). Springer, Berlin, Heidelberg, 134–151. https://doi.org/10.1007/978-3-642-23765-2_10
- Marc Herrlich, Parnian Tavakol, David Black, Dirk Wenig, Christian Rieder, Rainer Malaka, and Ron Kikinis. 2017. Instrument-mounted displays for reducing cognitive load during surgical navigation. *International Journal of Computer Assisted Radiology and Surgery* 12, 9 (Sept. 2017), 1599–1605. <https://doi.org/10.1007/s11548-017-1540-6>
- Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2018. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics* (2018).
- Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131.
- Chien-Ming Huang and Bilge Mutlu. 2016. Anticipatory Robot Control for Efficient Human-Robot Collaboration. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI '16)*. IEEE Press, 83–90. event-place: Christchurch, New Zealand.
- Michael Xuelin Huang, Jijia Li, Grace Ngai, Hong Va Leong, and Andreas Bulling. 2019. Moment-to-Moment Detection of Internal Thought from Eye Vergence Behaviour. (Jan. 2019). <http://arxiv.org/abs/1901.06572> arXiv: 1901.06572.
- Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. 2010. Active Learning by Querying Informative and Representative Examples. In *NIPS*.

- Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1233–1239.
- Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4198–4205.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *International Conference on Machine Learning*. PMLR, 4904–4916.
- Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. 2009. Multi-class active learning for image classification. In *CVPR*.
- Yunjae Jung, Dahun Kim, Sanghyun Woo, Kyungsu Kim, Sungjin Kim, and In So Kweon. 2020. Hide-and-Tell: Learning to Bridge Photo Streams for Visual Storytelling. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 07 (Apr. 2020), 11213–11220. <https://ojs.aaai.org/index.php/AAAI/article/view/6780>
- Ashish Kapoor, Kristen Grauman, Raquel Urtasun, , and Trevor Darrell. 2007. Active learning with gaussian processes for object categorization.. In *ICCV*.
- Sebastian Kapp, Michael Barz, Sergey Mukhametov, Daniel Sonntag, and Jochen Kuhn. 2021. ARETT: Augmented Reality Eye Tracking Toolkit for Head Mounted Displays. *Sensors* 21, 6 (March 2021), 2234. <https://doi.org/10.3390/s21062234> Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.
- Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1244–1254.
- Mi-Young Kim, Shahin Atakishiyev, Housam Khalifa Bashier Babiker, Nawshad Farruque, Randy Goebel, Osmar R Zaiane, Mohammad-Hossein Motallebi, Juliano Rabelo, Talat Syed, Hengshuai Yao, et al. 2021. A multi-component framework for the analysis and design of explainable artificial intelligence. *Machine Learning and Knowledge Extraction* 3, 4 (2021), 900–921.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 114, 13 (2017), 3521–3526.
- Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. 2019. BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning. In *NIPS*.
- Johannes Klonig and Marc Herrlich. 2020. Integrating 3D and 2D Views of Medical Image Data in Virtual Reality for Efficient Navigation. In *2020 IEEE International Conference on Healthcare Informatics (ICHI)*. 1–7. <https://doi.org/10.1109/ICHI48887.2020.9374344> ISSN: 2575-2634.
- Ksenia Konyushkova, Sznitman Raphael, and Pascal Fua. 2017. Learning Active Learning from Data. In *NeurIPS*.
- Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. 2019. Geometry in Active Learning for Binary and Multi-class Image Segmentation. *Computer Vision and Image Understanding* (2019), 1077–3142.
- Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.
- Christoph Käding, Erik Rodner, Alexander Freytag, and Joachim Denzler. 2017. Fine-Tuning Deep Neural Networks in Continuous Learning Scenarios. In *Computer Vision – ACCV 2016 Workshops*, Chu-Song Chen, Jiwen Lu, and Kai-Kuang Ma (Eds.). Springer International Publishing, Cham, 588–605.
- Sébastien Lallé, Dereck Toker, and Cristina Conati. 2021. Gaze-Driven Adaptive Interventions for Magazine-Style Narrative Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 27, 6 (2021), 2941–2952. <https://doi.org/10.1109/TVCG.2019.2958540>
- Michael Land, Neil Mennie, and Jennifer Rusted. 1999. The Roles of Vision and Eye Movements in the Control of Activities of Daily Living. *Perception* 28, 11 (1999), 1311–1328. <https://doi.org/10.1068/p2935> _eprint: <https://doi.org/10.1068/p2935>.
- Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-based human debugging of nlp models: A survey. *Transactions of the Association for Computational Linguistics* 9 (2021), 1508–1528.
- Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. 2019. SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4277–4286. <https://doi.org/10.1109/CVPR.2019.00441>
- Yuanpeng Li, Liang Zhao, Kenneth Church, and Mohamed Elhoseiny. 2020. Compositional Language Continual Learning. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=rklnDgHtDS>
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
- Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, and Shih-Fu Chang. 2018. Entity-aware Image Caption Generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 4013–4023. <https://doi.org/10.18653/v1/D18-1435>
- Matthias Maszuhn, Larbi Abdenebaoui, and Susanne Boll. 2021. A User-Centered Approach for Recognizing Convenience Images in Personal Photo Collections. In *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*. 1–4. <https://doi.org/10.1109/CBMI50038.2021.9461908> ISSN: 1949-3991.
- Alexander Mathews, Lexing Xie, and Xuming He. 2016. SentiCap: Generating Image Descriptions with Sentiments. *Proceedings of the AAAI Conference on Artificial Intelligence* 30, 1 (Mar. 2016). <https://doi.org/10.1609/aaai.v30i1.10475>

- Cynthia Matuszek. 2018. Grounded Language Learning: Where Robotics and NLP Meet. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 5687–5691. <https://doi.org/10.24963/ijcai.2018/810>
- Andrew Kachites McCallumzy and Kamal Nigamy. 1998. Employing em and pool-based active learning for text classification. In *ICML*.
- Gregor Mehlmann, Markus Häring, Kathrin Janowski, Tobias Baur, Patrick Gebhard, and Elisabeth André. 2014. Exploring a Model of Gaze for Grounding in Multimodal HRI. In *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI '14)*. Association for Computing Machinery, New York, NY, USA, 247–254. <https://doi.org/10.1145/2663204.2663275> event-place: Istanbul, Turkey.
- Zihang Meng, Licheng Yu, Ning Zhang, Tamara L Berg, Babak Damavandi, Vikas Singh, and Amy Bearman. 2021. Connecting what to say with where to look by modeling human attention traces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12679–12688.
- Sina Mohseni, Nilofar Zarei, and Eric D Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11, 3-4 (2021), 1–45.
- Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.
- A. Moran, V. Gadepally, M. Hubbell, and J. Kepner. 2015. Improving Big Data visual analytics with interactive virtual reality. In *2015 IEEE High Performance Extreme Computing Conference (HPEC)*. 1–6.
- Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. WT5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546* (2020).
- Duy Minh Ho Nguyen, Abraham Obinwanne Ezema, Fabrizio Nunnari, and Daniel Sonntag. 2020. A Visually Explainable Learning System for Skin Lesion Detection Using Multiscale Input with Attention U-Net. In *KI 2020: Advances in Artificial Intelligence - 43rd German Conference on AI, Bamberg, Germany, September 21-25, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12325)*, Ute Schmid, Franziska Klügl, and Diedrich Wolter (Eds.). Springer, 313–319. https://doi.org/10.1007/978-3-030-58285-2_28
- Fabrizio Nunnari and Daniel Sonntag. 2021. A Software Toolbox for Deploying Deep Learning Decision Support Systems with XAI Capabilities. In *EICS '21: ACM SIGCHI Symposium on Engineering Interactive Computing Systems, Virtual Event, The Netherlands, 8-11 June 2021*, Panos Markopoulos, Jun Hu, and Philippe A. Palanque (Eds.). ACM, 44–49. <https://doi.org/10.1145/3459926.3464753>
- Omar Jubran, Vera Eymann, Nicole Burkard, Jan Spilski, Marc Herrlich, Daniela Czernochowski, and Thomas Lachmann. 2021. Expanding on Behavioral Data Collection in an Adapted N-Back Task for Virtual Reality. In *Contributions to the 63rd Tagung Experimentell arbeitender Psychologen*. Pabst Science Publishers, Lengerich, Germany, Ulm, Germany.
- Sharon Oviatt. 2006. Human-centered design meets cognitive load theory: designing interfaces that help people think. In *Proceedings of the 14th ACM international conference on Multimedia (MM '06)*. Association for Computing Machinery, New York, NY, USA, 871–880. <https://doi.org/10.1145/1180639.1180831>
- Sharon Oviatt, Björn Schuller, Philip Cohen, Daniel Sonntag, Gerasimos Potamianos, and Antonio Krüger. 2019. *The Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions*. Morgan & Claypool.
- Firat Ozdemir, Zixuan Peng, Christine Tanner, Philipp Fuerstahl, and Orcun Goksel. 2018. Active Learning for Segmentation by Optimizing Content Information for Maximal Entropy. In *MICCAI Workshop: Deep Learning in Medical Image Analysis (DLMIA)*.
- Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018a. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8779–8788.
- Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018b. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 8779–8788.
- Rosalind W. Picard. 2000. *Affective computing*. MIT press.
- Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*. 2641–2649.
- Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *European conference on computer vision*. Springer, 647–664.
- Alexander Prange and Daniel Sonntag. 2021. A Demonstrator for Interactive Image Clustering and Fine-Tuning Neural Networks in Virtual Reality. In *KI 2021: Advances in Artificial Intelligence - 44th German Conference on AI, Virtual Event, September 27 - October 1, 2021, Proceedings (Lecture Notes in Computer Science, Vol. 12873)*, Stefan Edelkamp, Ralf Möller, and Elmar Rueckert (Eds.). Springer, 194–203. https://doi.org/10.1007/978-3-030-87626-5_14
- Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C. Lipton, Graham Neubig, and William W. Cohen. 2022. Evaluating Explanations: How Much Do Explanations from the Teacher Aid Students? *Transactions of the Association for Computational Linguistics* 10 (04 2022), 359–375. https://doi.org/10.1162/tacl_a_00465 arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00465/2006971/tacl_a_00465.pdf
- Dirk Queck, Iannis Albert, Nicole Burkard, Philipp Zimmer, Georg Volkmar, Bastian Dänekas, Rainer Malaka, and Marc Herrlich. 2022. SpiderClip: Towards an Open Source System for Wearable Device Simulation in Virtual Reality. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, 1–7. <https://doi.org/10.1145/3491101.3519758>
- Mohamad Rabbath, Philipp Sandhaus, and Susanne Boll. 2011a. Automatic creation of photo books from stories in social media. *ACM Transactions on Multimedia Computing, Communications, and Applications* 7S, 1 (Nov. 2011), 27:1–27:18. <https://doi.org/10.1145/2037676.2037684>
- Mohamad Rabbath, Philipp Sandhaus, and Susanne Boll. 2011b. Multimedia retrieval in social networks for photo book creation. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval (ICMR '11)*. Association for Computing Machinery, New York, NY, USA, 1–2. <https://doi.org/10.1145/1991996.1992068>

- Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4932–4942.
- Krishnan Ramnath, Simon Baker, Lucy Vanderwende, Motaz El-Saban, Sudipta N Sinha, Anitha Kannan, Noran Hassan, Michel Galley, Yi Yang, Deva Ramanan, et al. 2014. Autocaption: Automatic caption generation for personal photos. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 1050–1057.
- Gabriëlle Ras, Marcel van Gerven, and Pim Haselager. 2018. Explanation methods in deep learning: Users, values, concerns and challenges. In *Explainable and interpretable models in computer vision and machine learning*. Springer, 19–36.
- Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems* 29 (2016), 3567–3575.
- Anke Verena Reinschluessel, Joern Teuber, Marc Herrlich, Jeffrey Bissel, Melanie van Eikeren, Johannes Ganser, Felicia Koeller, Fenja Kollasch, Thomas Mildner, Luca Raimondo, Lars Reisig, Marc Ruedel, Danny Thieme, Tobias Vahl, Gabriel Zachmann, and Rainer Malaka. 2017. Virtual Reality for User-Centered Design and Evaluation of Touch-free Interaction Techniques for Navigating Medical Images in the Operating Room. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, New York, NY, USA, 2001–2009. <https://doi.org/10.1145/3027063.3053173> event-place: Denver, Colorado, USA.
- Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. 2020. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International Conference on Machine Learning*. PMLR, 8116–8126.
- Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. In *IJCAI*.
- Constantin A. Rothkopf, Dana H. Ballard, and Mary M. Hayhoe. 2016. Task and context determine where you look. *Journal of Vision* 7, 14 (July 2016), 16–16. <https://doi.org/10.1167/7.14.16>
- Gerben Rotman, Nikolaus F. Troje, Roland S. Johansson, and J. Randall Flanagan. 2006. Eye Movements When Observing Predictable and Unpredictable Actions. *Journal of Neurophysiology* 96, 3 (2006), 1358–1369. <https://doi.org/10.1152/jn.00227.2006> eprint: <https://doi.org/10.1152/jn.00227.2006>.
- N. Roy and A. McCallum. 2001. Toward optimal active learning through sampling estimation of error reduction. In *ICML*.
- Cynthia Rudin and Joanna Radin. 2019. Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. *Harvard Data Science Review* 1, 2 (2019). <https://doi.org/10.1162/99608f92.5a8a3a3d>
- Christian Rupprecht, Iro Laina, Nassir Navab, Gregory D Hager, and Federico Tombari. 2018. Guide me: Interacting with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8551–8561.
- Richard M. Ryan and Edward L. Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist* 55, 1 (2000), 68.
- Philipp Sandhaus and Susanne Boll. 2011. Semantic analysis and retrieval in personal and social photo collections. *Multimedia Tools and Applications* 51, 1 (Jan. 2011), 5–33. <https://doi.org/10.1007/s11042-010-0673-1>
- Philipp Sandhaus, Mohammad Rabbath, and Susanne Boll. 2011. Employing Aesthetic Principles for Automatic Photo Book Layout. In *Advances in Multimedia Modeling (Lecture Notes in Computer Science)*, Kuo-Tien Lee, Wen-Hsiang Tsai, Hong-Yuan Mark Liao, Tsuhan Chen, Jun-Wei Hsieh, and Chien-Cheng Tseng (Eds.). Springer, Berlin, Heidelberg, 84–95. https://doi.org/10.1007/978-3-642-17832-0_9
- Philipp Sandhaus, Sabine Thieme, and Susanne Boll. 2008. Processes of photo book production. *Multimedia Systems* 14, 6 (Dec. 2008), 351–357. <https://doi.org/10.1007/s00530-008-0136-y>
- Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. 2019. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2591–2600.
- Burr Settles. 2010. Active learning literature survey. *University of Wisconsin, Madison* 52, 55-66 (2010), 11.
- Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. KVQA: Knowledge-Aware Visual Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 8876–8884. <https://doi.org/10.1609/aaai.v33i01.33018876>
- Xiaoting Shao, Arseny Skryagin, Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. 2021. Right for Better Reasons: Training Differentiable Models by Constraining their Influence Function. In *Proceedings of Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*.
- Changjian Shui, Fan Zhou, Christian Gagné, and Boyu Wang. 2020. Deep active learning: Unified and principled method for query and training. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1308–1318.
- Patrice Y. Simard, Saleema Amershi, David Maxwell Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo A. Ramos, Jina Suh, Johan Verwey, Mo Wang, and John Wernsing. 2017. Machine Teaching: A New Paradigm for Building Machine Learning Systems. *CoRR abs/1707.06742* (2017). arXiv:1707.06742 <http://arxiv.org/abs/1707.06742>
- Asim Smailagic, Hae Young Noh, Pedro Costa, Devesh Walawalkar, Kartik Khandelwal, Mostafa Mirshekari, Jonathon Fagert, Adrián Galdrán, and Susu Xu. 2018. MedAL: Deep Active Learning Sampling Method for Medical Image Analysis. In *ICMLA*.
- Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 56–67.
- Daniel Sonntag. 2010. *Ontologies and adaptivity in dialogue for question answering*. Vol. 4. IOS Press.
- Daniel Sonntag, Fabrizio Nunnari, and Hans-Jürgen Profitlich. 2020. The Skincare project, an interactive deep learning system for differential diagnosis of malignant skin lesions. Technical Report. *arXiv preprint arXiv:2005.09448* (2020).

- Jamshid Sourati, Ali Gholipour, Jennifer G. Dy, Sila Kurugol, and Simon K. Warfield. 2018. Active Deep Learning with Fisher Information for Patch-wise Semantic Segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (2018), 83–91.
- Julian Steil and Andreas Bulling. 2015. Discovery of Everyday Human Activities from Long-Term Visual Behaviour Using Topic Models. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 75–85. <https://doi.org/10.1145/2750858.2807520> event-place: Osaka, Japan.
- Yusuke Sugano and Andreas Bulling. 2016. Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv:1608.05203* (2016).
- Ece Takmaz, Sandro Pezzelle, Lisa Beinborn, and Raquel Fernández. 2020. Generating Image Descriptions via Sequential Cross-Modal Alignment Guided by Human Gaze. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4664–4677.
- Chenhao Tan. 2021. On the Diversity and Limits of Human Explanations. *arXiv preprint arXiv:2106.11988* (2021).
- Stefano Teso and Oliver Hinz. 2020. Challenges in Interactive Machine Learning: Toward Combining Learning, Teaching, and Understanding. , 127–130 pages.
- Stefano Teso and Kristian Kersting. 2019. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 239–245.
- Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J. Mooney. 2016. Learning Multi-Modal Grounded Linguistic Semantics by Playing “I Spy”. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, 3477–3483. event-place: New York, New York, USA.
- Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. 2020. Rethinking Few-Shot Image Classification: a Good Embedding Is All You Need?. In *ECCV*.
- Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552* (2018).
- Simon Tong and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. *JMLR* 2 (2001), 45–66.
- Emma M. van Zoelen, Tina Mioch, Mani Tajaddini, Christian Fleiner, Stefani Tsaneva, Pietro Camin, Thiago S. Gouvêa, Kim Baraka, Maaiké H. T. de Boer, and Mark A. Neerincx. 2023. Developing Team Design Patterns for Hybrid Intelligence Systems. In *HHAI 2023: Augmenting Human Intellect - Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence, June 26-30, 2023, Munich, Germany (Frontiers in Artificial Intelligence and Applications, Vol. 368)*, Paul Lukowicz, Sven Mayer, Janin Koch, John Shawe-Taylor, and Ilaria Tiddi (Eds.). IOS Press, 3–16. <https://doi.org/10.3233/FAIA230071>
- Vera Eymann, Omar Jubran, Nicole Burkard, Jan Spilski, Marc Herrlich, Thomas Lachmann, and Daniela Czernochowski. 2021. Quantifying Cognitive Load by Combining Eye Tracking and EEG in a Virtual Reality Environment. In *Contributions to the 63rd Tagung Experimentell arbeitender Psychologen*. Pabst Science Publishers, Lengerich, Germany, Ulm, Germany.
- Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Henge. 2017. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 1290–1296.
- Ruize Wang, Zhongyu Wei, Piji Li, Qi Zhang, and Xuanjing Huang. 2020. Storytelling from an image stream using scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9185–9192.
- Zijie J Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting Humans in the Natural Language Processing Loop: A Survey. In *Proceedings of the First Workshop on Bridging Human-Computer Interaction and Natural Language Processing*. 47–52.
- David Watson and Lee Anna Clark. 1994. The PANAS-X: Manual for the positive and negative affect schedule-expanded form. (1994). Publisher: University of Iowa.
- Davis Werthimer, Luming Tang, and Bharath Hariharan. 2021. Few-Shot Classification With Feature Map Reconstruction Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8012–8021.
- Sarah Wiegrefe and Ana Marasovic. 2021. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. <https://openreview.net/forum?id=ogNcxJn32BZ>
- Sarah Wiegrefe, Ana Marasović, and Noah A. Smith. 2021. Measuring Association Between Labels and Free-Text Rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 10266–10284. <https://doi.org/10.18653/v1/2021.emnlp-main.804>
- Ani Withöft, Larbi Abdenebaoui, and Susanne Boll. 2022. ILMICA - Interactive Learning Model of Image Collage Assessment: A Transfer Learning Approach for Aesthetic Principles. In *MultiMedia Modeling (Lecture Notes in Computer Science)*, Björn Þór Jónsson, Cathal Gurrin, Minh-Triet Tran, Duc-Tien Dang-Nguyen, Anita Min-Chun Hu, Binh Huynh Thi Thanh, and Benoit Huet (Eds.). Springer International Publishing, Cham, 84–96. https://doi.org/10.1007/978-3-030-98355-0_8
- Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z. Chen. 2017. Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation. In *MICCAI*.
- Huihan Yao, Ying Chen, Qinyuan Ye, Xisen Jin, and Xiang Ren. 2021. Refining Language Models with Compositional Explanations. *Advances in Neural Information Processing Systems* 34 (2021).
- Qinyuan Ye, Xiao Huang, Elizabeth Boschee, and Xiang Ren. 2020. Teaching Machine Comprehension with Compositional Explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 1599–1615.
- Jan Zacharias, Michael Barz, and Daniel Sonntag. 2018. A survey on deep learning toolkits and libraries for intelligent user interfaces. *arXiv preprint arXiv:1803.04818* (2018).

Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z. Li. 2020. Context-Aware Attention Network for Image-Text Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.