xPrEs – Crossmodal Perception Trace Embedding

Prof. Dr. Achim Rettinger & Simon Werner, M.A.

Representation learning techniques are key to the success of recent machine learning models. A key component are (self-) attention mechanisms which learn to judge the relevance of elements of the input (e.g. words in a sentence) for contextualizing other elements of the input. Interestingly, this can be related to human perception models, which are researched in media studies or psychology.

When interpreting eye-tracking experiments, it becomes apparent that humans exploit essential cues, that are key to extracting abstract meaning from multimedia documents: (1) Humans construct meaning from a document's content by moving their attention between text and image as guided by layout and design elements; (2) Not all parts of the document are perceived equally: e.g., words are skipped, while others are revisited several times.

Since multimedia documents are made by humans for humans, we argue that to better represent their content for computational processing, they should be parsed in a human-like manner. xPres intends to investigate the potential of human-like perception models for representation learning techniques. The core idea is a novel content representation paradigm that represents multimedia documents similar to the way they are perceived by humans: A sequence of shifts of attention across different modalities, like words and image regions arranged in a multimedia document. With such a perception-based document representation, xPres will enable to investigate two fundamental research questions:

(1) What are similarities and differences between human perception and modern representation learning techniques, in respect to how they attend to the content of multimedia documents? (2) If augmented with an inductive perception bias can representation learning techniques be improved?

xPres will produce:

(1) techniques for the automated extraction of perception traces from multimedia documents;

(2) more data-efficient representation learning techniques for multimedia documents;

(3) insights into similarities and differences between human perception and patterns learned by representation learning techniques;

(4) superior performance on representation learning benchmarks;

(5) open and freely usable human eye-tracking recordings on multimedia documents, crossmodal perception-trace based representation learning models and context dependent cross-modal perception-trace based embeddings.

Motivated by human perception, xPres' representation learning approach is the first to attempt to capture meaning encoded in a multimedia document beyond the sum of its single-modal parts. Since no perception-based ML-approaches have been proposed yet, xPres has the potential to become seminal for many important lines of current research in Artificial Intelligence (AI) like explainable AI, AI alignment and learning from limited data.