Multimodal Perception and Interaction with Transformers

Tasks T1.2, T2.2, T2.3, T2.4 and T3.1

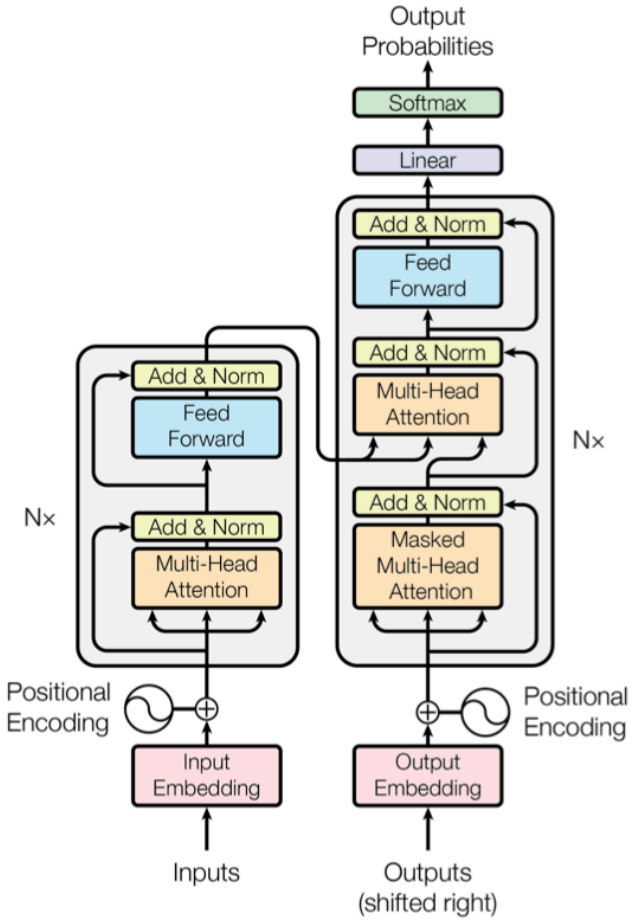
1 April 2021 to 31 Sept 2021 (6 months)

INRIA - James Crowley and Yangtao Wang
Eotvos Lorand University (ELTE) - Andras Lorincz
Univ Grenoble Alpes, (LIG) - Dominique Vaufreydaz, Fabien Ringeval
Univ Paris Saclay (LISN CNRS) - Camille Guinaudeau, Marc Evrard
Jozef Stefan Institut (JSI) - Marko Grobelnik
Charles University - Pavel Pecina

Objective:

Provide tools, data sets and performance metrics to demonstrate the potential of transformers for multimodal perception and multimodal interaction.

Transformers and Self-attention for Multimodal Interaction



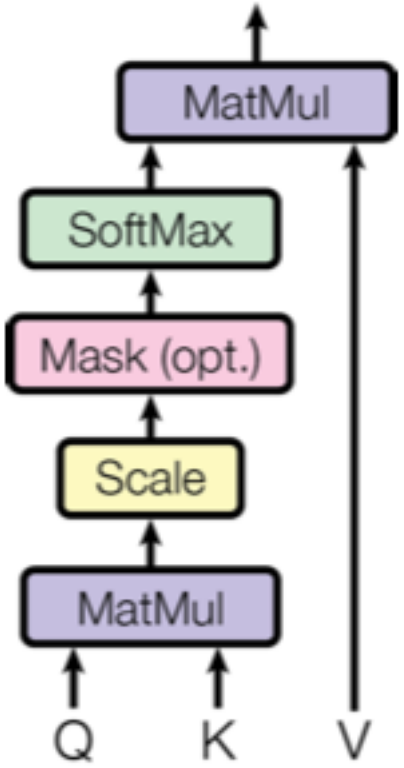
Transformers and self-attention replace deep learning with multiple layer of encoder-decoders using self-attention to encode signals at multiple levels of abstractions. (words, sequences, sentences, contexts…).

Transformers have become the dominant approach for natural language processing (NLP). Recent results indicate that transformers are well suited for multi-modal perception.

Our objective is to **build a research community** around the use of transformers and self-attention for multimodal perception and multi-modal interaction.

*Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.

Attention is all you need



Self-attention associates entities with mutual relevance using learned matrices for Query-Key-Value.

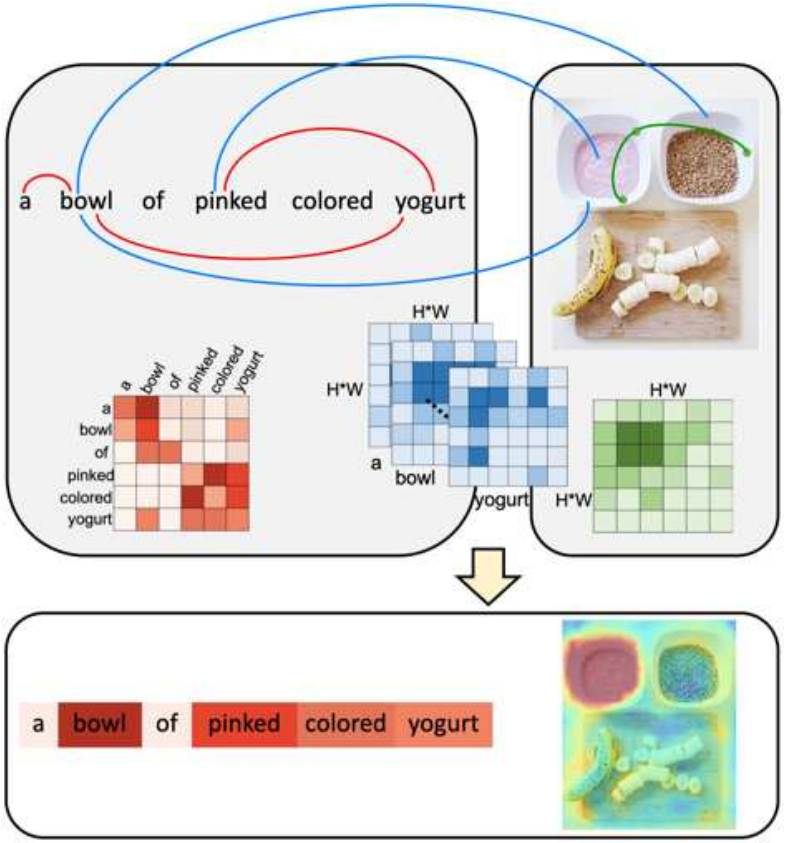
Each layer uses multiple Self-Attention Heads to associate multiple mutually relevant entities to be interpreted at that next level.

We can extend self attention to multiple modalities using concatenation.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

*Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.

Transformers for Multimodal Interaction



Each layer uses multiple Self-Attention Heads to associate multiple mutually relevant entities to be interpreted at that next level.

Transformers can be trained to complete missing data with multiple modalities, to anticipate events and to predict best actions for situations.

However extension to new modalities requires new forms of encoding and new training data.

From: Ye et. Al. "Cross-Modal Self-Attention Network for Referring Image Segmentation", CVPR 2019, IEEE Conf. on Computer Vision and Pattern Recognition, June, 2019.

What will we do?

1. Explore new forms of embedding for image, video sequences, prosody and spoken language.
2. Explore audio-visual narration, and multimodal perception of emotion, engagement and attention.
3. Define performance metrics and benchmark data for research challenges.
4. Provide base-line techniques for multimodal perception with transformers and self-attention.

Tangible Results

1. Benchmark data sets
2. Research challenges with Performance Metrics
3. Workshop Reports
4. Baseline demonstrations for NLP, Vision, Prosody, spoken language

Ultimately, define the tools for a new scientific community devoted to use of transformers for multimodal perception and multimodal interaction.