Humane AI Net Micro-project Proposal

Multimodal Perception and Interaction with Transformers

Transformers and self-attention (Vaswani et al., 2017), have become the dominant approach for natural language processing (NLP) with systems such as BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020) rapidly displacing more established RNN and CNN structures with an architecture composed of stacked encoder-decoder modules using self-attention. While most results with transformers have addressed problems in NLP, recent results have shown that transformers are well suited for multi-modal perception combining language and computer vision (Sun et al, 2019). However, adaptations to domains other than natural language are far from straight-forward as the most appropriate encoder-decoder techniques, embeddings and related hyper-parameters must be determined for each new modality.

This micro-project will provide tools and data sets for experiments and a first initial demonstration of the potential of transformers for multimodal perception and multimodal interactions. We will define research challenges, benchmark data sets and performance metrics for multimodal perception and interaction tasks such as (1) audio-visual narration of scenes, cooking actions and activities, (2) audio-video recordings of lectures and TV programs (3) audio-visual deictic (pointing) gestures, and (4) perception and evocation of engagement, attention, and emotion.

Humane AI Net WP Tasks Involved
T1.2 Learning with and about narratives
T2.2 Multimodal perception and modeling of actions, activities and tasks
T2.3 Multimodal perception of awareness, emotions, and attitudes
T2.7 Assembling benchmark datasets
T3.6 Language-based and Multilingual Interaction

Partners and Effort Charged to Humane AI Net

INRIA - James Crowley and Yangtao Wang (4 PM)
Eotvos Lorand University (ELTE) - Andras Lorincz (3 PM)
Univ Grenoble Alpes, LIG (Dominique Vaufreydaz, Fabien Ringeval (2 PM)
Uni Paris Saclay (CNRS), LISN - Camille Guinaudeau, Marc Evrard (2 PM)
JSI (Jozef Stefan Institut)- Marko Grobelnik (2 PM)
Charles University - Pavel Pecina (1 PM)

Associated Partners (participate in meetings):

Aalto Univ, Helsinki - Antti Oulasvirta (0 PM)

Start: 1 April 2021
Duration 6 months
Total Charged Effort 14 PMs

Tangible Output
Tangible results will include:
1) Benchmark data and performance targets for a phased set of research challenges of increasing difficulty.
2) Tools for experiments to explore use of embeddings, encoder-decoders, self-attention architectures and related problems associated with applying transformers to different modalities.
3) Concept demonstrations for simple examples of multimodal perception.

Bibliography:

1) 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
2) Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
3) Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., and Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
4) Sun, C., Myers, A., Vondrick, C., Murphy, K., and Schmid, C. (2019). Videobert: A joint model for video and language representation learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7464-7473)