ACAI 2021
The advanced course on AI on Human Centered AI

Tutorial Outline

Multimodal Perception and Interaction with Transformers
Francois Yvon, Univ Paris Saclay, James Crowley, INRIA and Grenoble Institut Polytechnique

Transformers and self-attention have become the dominant approach for natural language processing with systems such as BERT and GPT-3 rapidly displacing more established RNN and CNN structures. Recent results have shown that Transformers are also well suited for multi-modal perception and multi-modal interaction.

In this advanced tutorial, we review the emergence of attention for bilingual language models, and show how this led to the Transformer architecture composed of stacked encoder and decoder layers using multi-headed attention. We discuss techniques for token and position embeddings for natural language, and show how these can be trained using a masked language model. We describe how this approach can be extended to other individual modalities (speech, images) and to multiple modalities by concatenating encodings of modalities and discuss problems and approaches for adapting transforms for use with computer vision and spoken language interaction. We conclude with a review of current research challenges, performance evaluation metrics, and benchmark data sets, followed by a discussion of potential applications such as multimodal sentiment analysis, affective interaction, and narrative description of activities.

Introduction: Multimodal Perception and Interaction (James Crowley - 10-15 mins)

1. Multimodal Perception and Multimodal Interaction
2. Why multimodal Interaction is important for Human Centered AI
3. Transformers as a possible rupture technology for Multimodal Interaction

Transformers in Natural Language Processing (François Yvon – 1h30)

1. Text classification and language models
    a. The simplest of all tasks: spam filtering
    b. Processing sequential events: language models
    c. Feed forward and recurrent neural networks for language modeling
    d. Neural networks as feature extractors
2. The Transformer architecture

. Internal layers: multi-headed attention and feed-forward
a. Input and output layers
b. The computations of a Transformer
c. Causal transformers for language modeling
d. Transformers as feature extractors

3. Encoder-Decoder architecture for Neural Machine translation

. Basic NMT: a conditional language model

a. Multilingual NMT and Multilingual representations
b. Monolingual machine translation

[break]

Transformers in Speech (Marc Evrard – 30-45 minutes)

1. Speech representation

a. Classical speech features and modeling
i.Spectrum
ii.MFCC
iii.HMM
b. Speech representation in NN
.Encoding/Decoding
i.CTC loss

2. Speech Transformer

. Attention
.2D Attention Mechanism
i.Positional encoding
ii.Wav2Vec
a. Application of Transformers
.Speech Recognition (ASR)
i.Spoken Language Understanding (SLU)
ii.Emotion Recognition (AER)

3. Speech Recognition Transformers

. Conformer
a. Speech recognition with Wav2vec2

Transformers in Vision (Camille Guinaudeau – 30-45 minutes)

1. From CNN to Vision Transformer

a. Features extraction in Image
b. Self-attention layers

2. Vision Transformers

. Explicit positional encoding
a. Implicit positional encoding
b. Introducing Convolutions to Vision Transformers

3. Multi-Modal Transformer and Temporal encoding

Conclusions (James Crowley - 10-15 minutes)

1. Open Problems, Data Sets and Research Challenges
2. Research Communities and Publication Venues

Biographical Information of Tutors

Francois Yvon is a senior researcher in the Spoken Language Processing Group of the CNRS LISN Laboratory (Laboratoire Interdisciplinaire des Sciences du Numérique) at the Univ. Paris Saclay. He currently focuses mainly on machine translation using statistical and neural methods - and more generally on machine learning applied to both written and vocal multilingual language data. He is a member of the executive board of the European Meta-NET network, as well as the French contact point for the European CEF-ELRC programme.

Marc Evrard is an associate professor (Maitre de Conference) at Univ. Paris Saclay and works as a researcher at the Spoken Language Processing Group of the CNRS LISN Laboratory (Laboratoire Interdisciplinaire des Sciences du Numérique). He received his doctorate in Computer Science from the Univ. Paris-Sud (now Univ. Paris Saclay) in 2015, with a thesis on Expressive Text-to-Speech Synthesis. His research now focuses on speech processing and natural language processing in the context of digital humanities.

Camille Guinaudeau is an associate professor (Maitre de Conference) at Université Paris Saclay. She works as a researcher at the CNRS LISN Laboratory (Laboratoire Interdisciplinaire des Sciences du Numérique) in the Spoken Language Processing group. She received her doctorate in 2011 with a thesis on the automatic structuring of TV streams. Her current research concerns spoken language processing, information retrieval and the structure of multimedia documents.

James L. Crowley is a Professor at the Univ. Grenoble Alpes, where he teaches courses in Computer Vision, Machine Learning, and Artificial Intelligence at Grenoble Institut Polytechnique at the Univ Grenoble Alpes. Over the last 35 years, professor Crowley has made a number of fundamental contributions to computer vision, robotics, and multi-modal interaction. Professor Crowley has recently been named to the Chair on Intelligent Collaborative Systems at the MAIA AI Institute at the University Grenoble Alpes.