Contribution ID: **21**

Type: **not specified**

Towards end-to-end speech-to-text summarisation

Wednesday 29 June 2022 09:48 (11 minutes)

Deep neural networks research has thrived since AlexNet (Krizhevsky et al., 2012) outperformed classical computer vision techniques in the ImageNet Large Scale Visual Recognition Challenge in 2012. The latter used tailored feature extraction algorithms, whereas AlexNet only resorted to the depth of the model to achieve such high performance. Since then, deep neural networks have been broadly used in Natural Language Processing (NLP), from which transformers and language models account for some of its most recent developments. Under a paradigm-changing title, "Attention Is All We Need", the transformer architecture was initially introduced by Vaswani et al. (2017) to implement end-to-end translation and relies on a mechanism of self-attention to encode each input word considering its global context. The transformer architecture then became prevalent in most NLP tasks: machine translation, parsing, text summarisation, to name a few. Unsurprisingly, the same methodologies have been applied to speech processing (Dong et al., 2018), particularly Automatic Speech Recognition (ASR). Motivated by this progress, more advanced methods that implement end-to-end speech-to-text translation have arisen (Vila et al., 2018), in opposition to cascade models, which require an intermediate step for transcription. Following the same reasoning, the goal of this master thesis is to develop a model that implements end-to-end speech-to-text summarisation. The model will be based on the transformers architecture and shall use the speech audio as input to directly produce a summary of its content without a transcription step.

Author: POMBO MONTEIRO, Raul

Presenter: POMBO MONTEIRO, Raul