Université de Montréal

Faculté des arts et des sciences
Département d'**informatique et de recherche opérationnelle**

Transformer model for question answering, a review

Gao Yinghan, Wei Wei, Frederic Boileau

January 27, 2021

“The dominant approach to date encodes the input sequence with a se- ries of bi-directional recurrent neural networks (RNN) and generates a variable length output with another set of de- coder RNNs, both of which interface via a *soft-attention mechanism* (Bahdanau et al., 2014; Luong et al., 2015).”
“Compared to recurrent layers, convolutions create representations for fixed size contexts, however, the effective context size of the network can easily be made larger by stacking several layers on top of each other. This allows to precisely control the maximum length of dependencies to be modeled.”
“Convolutional networks do not depend on the computations of the previous time step and therefore allow parallelization over every element in a sequence. This contrasts with RNNs which maintain a hidden state of the entire past that prevents parallel computation within a sequence.”
“Multi-layer convolutional neural networks create hierarchi- cal representations over the input sequence in which nearby input elements interact at lower layers while distant ele- ments interact at higher layers. Hierarchical structure pro- vides a shorter path to capture long-range dependencies compared to the chain structure modeled by recurrent networks”[1]

References

[1] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning, 2017.