Annotated Bibliography

Kyle Roth

December 5, 2019

In this paper I will discuss the technical applications of attention mechanisms used in neural networks. Attention was first invented as a way to solve long-range dependencies when doing sequence to sequence models, like machine translation. The application of attention has since broadened to include sequence comparison, image generation, image description, and a host of other tasks. The theoretical breadth of attention has also increased, as observed in (Vaswani et al., 2017) and (Kim, Denton, Hoang, & Rush, 2017). But there is evidence to suggest that attention is just a weakened version of active memory (Kaiser & Bengio, 2016), meaning that further investigation of attention could bring better approaches to light.

Studying the body of work on the subject of attention mechanisms will help me develop my personal understanding of attention, so that I can implement it myself when creating my own models. This study will give me a greater depth to understand what has been done with attention and what open questions have been left unanswered. Perhaps I will find that attention is now defunct because a newer approach is more versatile. In that case, I want to understand why that is.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *CoRR*, *abs/1409.0473*. Retrieved Nov 21, 2019, from https://arxiv.org/pdf/1409.0473.pdf

This paper was foundational to attention mechanisms. Up to this point, RNNs for machine translation relied on an encoding vector of fixed length to represent the entire sentence. This time, the contributors allowed the decoder to rely on encoded representations of each word in the sentence, allowing it to learn the (sometimes long range) dependencies between individual words. This paper coined the phrase "attention mechanism", defined it as a learned mechanism (instead of a latent variable), and demonstrated that models with attention achieved higher performance on translations of long sentences. They produced the first neural machine translation model with accuracy comparable to that of statistical models.

Chan, W., Jaitly, N., Le, Q. V., & Vinyals, O. (2015). Listen, attend, and spell. *CoRR*, *abs/1508.01211*. Retrieved Nov 21, 2019, from http://arxiv.org/abs/1508.01211

The same year the original attention paper was published, this group performed end-to-end speech transcription with an RNN+attention model. They used a time-dimension reducing BLSTM network to convert raw audio into a compressed representation, and then an RNN with attention to decode that representation into

character sequences. This work demonstrated that attention could be used to allow for greater dependence between outputs, because without those dependencies a character-level output would not work. This paper does not contribute in terms of model architecture, but does demonstrate attention's application for character-level decoding.

Chen, H., Ding, G., Lin, Z., Zhao, S., & Han, J. (2018, 7). Show, observe and tell: Attribute-driven attention model for image captioning. In *Proceedings of the twenty-seventh international joint conference on artificial intelligence, IJCAI-18* (pp. 606–612). International Joint Conferences on Artificial Intelligence Organization. Retrieved Nov 21, 2019, from `https://doi.org/10.24963/ijcai.2018/84` **doi: 10.24963/ijcai.2018/84**

This paper implemented a CNN-RNN attention architecture for image captioning, achieving state-of-the-art performance on the task. The attention module received both region-based and attribute-based features from the image. This gave the model the ability to adaptively attend to objects in the image, taking into account the position and attributes of the objects. This paper demonstrates the versatility of the attention mechanism with regard to input feature type.

Cho, K., Courville, A. C., & Bengio, Y. (2015). Describing multimedia content using attention-based encoder-decoder networks. *CoRR*, *abs/1507.01053*. Retrieved Nov 21, 2019, from `http://arxiv.org/abs/1507.01053`

This article offers a valuable introduction to the structure of an attention network, as used in the first paper on attention. It also reviews work on using attention for images and videos, which lends itself to tasks where object identification and location are both important, such as content descriptions. The paper provides a good understanding of how convolutions and recurrent layers can both be used in networks with attention mechanisms.

Cybenko, G. (1989, Dec 01). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314. Retrieved from `https://doi.org/10.1007/BF02551274` **doi: 10.1007/BF02551274**

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing machines. *CoRR*, *abs/1410.5401*. Retrieved Dec 4, 2019, from `http://arxiv.org/abs/1410.5401`

Gregor, K., Danihelka, I., Graves, A., & Wierstra, D. (2015). DRAW: A recurrent neural network for image generation. *CoRR*, *abs/1502.04623*. Retrieved Nov 21, 2019, from `http://arxiv.org/abs/1502.04623`

These authors combine variational autoencoders (VAEs) with recurrent networks and attention to create a model that iteratively improves a generated image from an encoded input vector. At each iteration, attention is used both to choose which part of the encoding to consider, and to choose which part of the output drawing to update. This interesting application of attention is very natural to the way humans understand image generation, and also presents promising models for image classification.

Guru99. (n.d.). Back propagation neural network: Explained with simple example. `https://www.guru99.com/backpropogation-neural-network.html`.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*, 245–258. Retrieved Dec 4, 2019, from `https://www.cell.com/neuron/fulltext/S0896-6273(17)30509-3`

Kaiser, L., & Bengio, S. (2016). Can active memory replace attention? *CoRR*, *abs/1610.08613*. Retrieved Nov 21, 2019, from `http://arxiv.org/abs/1610.08613`

Soft-decision attention functions much like memory, in that values are stored for future use and read from somewhat selectively. This paper discusses the benefits of using active memory as a replacement for attention, and finds that in nearly all cases attention performs worse than active memory. This has broad implications for the use we find in the attention mechanism, and future work with attention mechanisms will need to determine whether attention's usefulness is found simply in attention as memory.

Kim, Y., Denton, C., Hoang, L., & Rush, A. M. (2017). Structured attention networks. *CoRR*, *abs/1702.00887*. Retrieved Nov 21, 2019, from `http://arxiv.org/abs/1702.00887`

This paper generalizes the idea of attention beyond its application to sequences, like sentences or audio samples, to arbitrarily-structured graphs. Models like conditional random fields can be used to model structural dependencies during end-to-end training of neural networks. Allowing for tree-like structure lets attention models attend to subsequences, and this accomplishes near-state-of-the-art results on machine translation and on natural language inference. This generalization of attention mechanisms deserves further study.

Larochelle, H., & Hinton, G. (2010). Learning to combine foveal glimpses with a third-order boltzmann machine. In *Proceedings of the 23rd international conference on neural information processing systems - volume 1* (pp. 1243–1251). USA: Curran Associates Inc. Retrieved from `http://dl.acm.org/citation.cfm?id=2997189.2997328`

Lin, Z., Feng, M., dos Santos, C. N., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. *CoRR*, *abs/1703.03130*. Retrieved Nov 21, 2019, from `http://arxiv.org/abs/1703.03130`

This work uses attention to create an embedding for sentences that reflects the semantic value of each sentence. In this case, each hidden vector in the LSTM is used as input to a self-attention mechanism, which compiles a single vector that tends to focus on a specific component of the sentence. This use of attention is distinct because it focuses on creating representations that are semantically significant without any further processing (like in translation or transcription).

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (icml-10)* (pp. 807–814).

Parikh, A. P., Täckström, O., Das, D., & Uszkoreit, J. (2016). A decomposable attention model for natural language inference. *CoRR*, *abs/1606.01933*. Retrieved Nov 21, 2019, from `http://arxiv.org/abs/1606.01933`

Other articles referenced here use attention as a way to describe relationships between elements in a sequence. This paper extends attention to learning the relationship between elements in two sequences, in this case allowing the model to determine whether one sentence entails or contradicts the other. (Intra-sequence attention is also used to provide the model with the "compositional relationships" between words in the sentence.) The authors' model achieved accuracy that slightly outperformed state-of-the-art models that were much larger. This interesting use of attention points to further uses of attention beyond what's currently considered.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008. Retrieved Nov 19, 2019, from https://papers.nips.cc/paper/7181-attention-is-all-you-need

This work did away with the recurrent layers used with attention up until this point, simply training the attention mechanism to perform the task of text translation. Recurrent models have to be run sequentially, so removing them allows training and evaluation to run faster. They achieved state-of-the-art performance on English-to-German and English-to-French, while being able to train the models in significantly less time. In this work the researchers were careful to analyze what each component of the model contributed, allowing them to draw interesting conclusions. Another contribution is the definition of attention as a learned function of queries, keys, and values. The source of each of these inputs determines the kind of attention used.