Is attention really all you need?

Kyle Roth
Brigham Young University
kylrth@gmail.com

Contents

Abstract

This is a wonderful abstract.

1 Introduction

The attention mechanism is an exciting development in the artificial intelligence community. The concept is simple, and rooted in our understanding of attention in biological intelligence [Larochelle and Hinton, 2010; Hassabis *et al.*, 2017]. Attention mechanisms also lend themselves to more intuitive interpretation, an attractive feature as concerns about model interpretability gain traction. In this review, we observe the trajectory that current research is taking with regard to the theory and application of attention, and suggest where future work with attention may yield fruits. [STRENGTHEN THIS THESIS]

1.1 Background on neural machine learning

Machine learning is a branch of artificial intelligence that deals with training a model to perform tasks by providing the model with data. There are various types of machine learning algorithms, applicable to different tasks. One of the most popular algorithms is the neural network. A neural network is a system of connections between simple functions (e.g. multiplication, summation, ReLU [Nair and Hinton, 2010]), which together approximate a larger, more complex function. It has been proven that sufficiently large neural networks can approximate arbitrary continuous functions [Cybenko, 1989]. Theoretically, this means that for every reasonable mapping there exists a neural network that replicates it. Some example mappings are listed in Table 1.

Input	Output
images	the objects the image contains
audio recordings	a text transcription
English text	a French translation

Table 1: The caption of the table

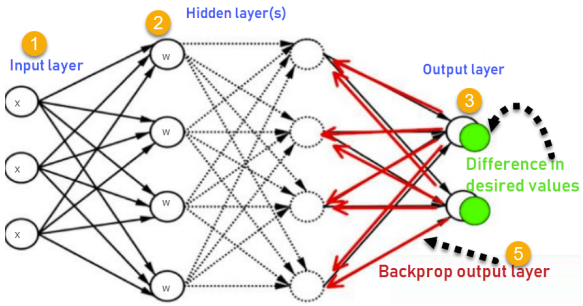

Figure 1: A visualization showing that backpropagation uses the difference between the desired and actual output to train the weights of the neural network. Image source [Guru99, 2019].

Most commonly, neural networks undergo supervised training, meaning that they are shown the correct output for each training example and made to improve in some way. Neural networks generally use backpropagation and gradient descent to train weights that are present in the simple functions of the network. Backpropagation uses the difference between the correct example and the network's output to determine how the weights in the network need to change. Figure 1 provides a visualization of backpropagation in a neural network.

While the theoretical limitations of neural networks are quite permissive, it is relatively difficult to train them. Back-

propagation is sensitive to "noisy" or incorrect data, meaning that data selection is at least as important as architecture selection. The training process is also relatively expensive in terms of computation time and memory usage; neural networks are commonly trained for days or even months [Vaswani *et al.*, 2017].

Another real issue is that backpropagation is unlikely to find the "global optimum", the absolutely optimal set of weights that approximate the function. This is due to the fact that the algorithm only adjusts the weights incrementally, and can get stuck in a "local optimum", a condition where no slight change would improve performance but performance is still not as good as at the global optimum.

These difficulties with training neural networks are the reason that so much effort is given to developing better network architectures. Certain architectures and data representations lend themselves to a smoother search space for the weights, increasing the likelihood that backpropagation reaches a better optimum.

1.2 Description of attention

Attention appears to provide that smoother search space for training the network because its output can explicitly depend on each input given to the network, while also giving an intuitive view of the inputs that the network learns to "attend" to.

Bahdanau *et al.* [2015] first described attention as an "alignment model" for words in a sentence to be translated. They used an RNN encoder-decoder model as others had before [Cho *et al.*, 2014], but for every position in the target sentence, the alignment model produced a "soft alignment"[1] between the target location and every word in the source sentence. See Figure 2 for more details on the model.

The key development here was the use of a soft alignment, meaning that each pair of source and target words was given an alignment score between 0 and 1. This made the alignment calculation differentiable, allowing the alignment to be performed by a feed-forward neural network trained along with the rest of the model.

2 Theoretical generalization

Ironically, attention mechanisms have received lots of attention since their conception in 2015. Initially formulated as a word alignment mechanism for text translation [Bahdanau *et al.*, 2015], the concept was expanded [BY WHOM? SEE WHO CITES THESE GUYS]

[Chan *et al.*, 2015]

Unsure: [Chen *et al.*, 2018], [Gregor *et al.*, 2015]

2.1 Self-attention

[Lin *et al.*, 2017]

2.2 Attention between existing pairs

[Parikh *et al.*, 2016]

[1] In machine translation, alignment refers to the pairing of words or phrases in the source sentence with words or phrases in the target sentence.

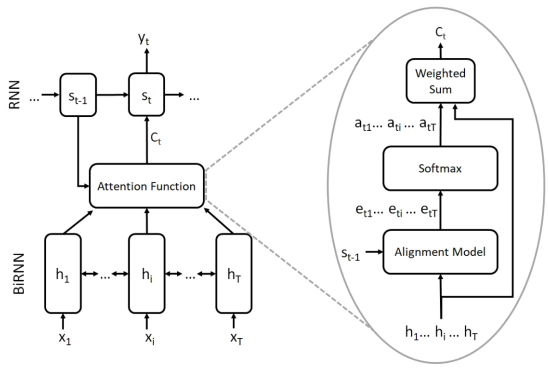

Figure 2: The model used by [Bahdanau *et al.*, 2015]. They used a bidirectional RNN to encode each word x_i into a fixed-length vector h_i. Then, for every position t in the target sentence, the alignment model produced an alignment vector a_t, a measure of alignment between the target location and every word in the source sentence. a_t was used to find the weighted sum $c_t = \sum_{i=1}^{T} a_{ti} h_i$, which was used as the input to the decoder (a unidirectional RNN). Note that the alignment model also received the previous decoder RNN state s_{t-1} as input. Image source [Galassi *et al.*, 2019].

2.3 Attention on arbitrary graphs

[Kim *et al.*, 2017]

2.4 Attention as a lookup

2.5 Active memory

[Kaiser and Bengio, 2016]

Neural memory resources aren't the same thing [Graves *et al.*, 2014]?

3 Application

3.1 Attention without recurrent layers

[Vaswani *et al.*, 2017]

4 Interpretation

Can you use attention to do feature importance?

5 Conclusion

References

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2015.

[Chan *et al.*, 2015] William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. Listen, attend, and spell. *CoRR*, abs/1508.01211, 2015.

[Chen *et al.*, 2018] Hui Chen, Guiguang Ding, Zijia Lin, Sicheng Zhao, and Jungong Han. 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*, pages 606–612. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax,Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, oct 2014. Association for Computational Linguistics.

[Cybenko, 1989] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, Dec 1989.

[Galassi *et al.*, 2019] Andrea Galassi, Marco Lippi, and Paolo Torroni. Attention, please! A critical review of neural attention models in natural language processing. *CoRR*, abs/1902.02181, 2019.

[Graves *et al.*, 2014] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural Turing machines. *CoRR*, abs/1410.5401, 2014.

[Gregor *et al.*, 2015] Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. DRAW: A recurrent neural network for image generation. *CoRR*, abs/1502.04623, 2015.

[Guru99, 2019] Guru99. Back propagation neural network: Explained with simple example. https://www.guru99.com/backpropogation-neural-network.html, 2019.

[Hassabis *et al.*, 2017] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95:245–258, 2017.

[Kaiser and Bengio, 2016] Lukasz Kaiser and Samy Bengio. Can active memory replace attention? *CoRR*, abs/1610.08613, 2016.

[Kim *et al.*, 2017] Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. Structured attention networks. *CoRR*, abs/1702.00887, 2017.

[Larochelle and Hinton, 2010] Hugo Larochelle and Geoffrey Hinton. 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*, NIPS'10, pages 1243–1251, USA, 2010. Curran Associates Inc.

[Lin *et al.*, 2017] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130, 2017.

[Nair and Hinton, 2010] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[Parikh *et al.*, 2016] Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *CoRR*, abs/1606.01933, 2016.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.