Reading 1

(1) Title, author and where it was published

- Visualizing A Neural Machine Translation Model (Mechanics of Seq2seq Models With Attention)
- Sequence to Sequence Learning with Neural Networks
- Authors: Jay Alammar; Ilya Sutskever, Oriol Vinyals, Quoc V. Le
- Published: Github Blog Post

(2) Summary of the content

This blog post starts by introducing the basic concepts about the sequence-to-sequence model, and then extends the introduction to the attention model. The author explains the mechanism behind an attention model in detail using the example of a machine translation task throughout its discussion. By using animations, the author provides visualizations of the learning process of the model introduced in the second paper. This model consists of a multilayered LSTM model as an encoder that takes in input sequences and another deep LSTM as a decoder to output the result. The authors in the second paper tested the effectiveness of such a structure using the WMT-14 dataset, and concluded that LSTM is capable of learning long sentences especially by reversing the words in the source sentences. In the second paper, the authors provide more details on the setup of the model, including maximizing the log probability on a correct translation as the objective function and a left-to-right beam search decoder to find the most likely translation.

(3) Strengths of the paper

The step-by-step animations in the first paper provides the reader with a clear understanding of the mechanism of an attention model. The author first simplifies the sequence-to-sequence model by using an encoder-decoder chain, and explains the role of context in this model. While an attention model is similar to the basic setup of a sequence-to-sequence model, its decoders take all the hidden states vectors from the encoders, and apply a soft-maxed score to each vector to output the "most-relevant" one, and concatenate this information with the context vector, and apply this procedure recurrently on each attention layer. In this way, hidden states with high scores are amplified, and states with low scores will be drowned in the process. The reader could really understand how the sentence in French is being translated into English in a context-related but not word-by-word structure.

The authors in the second paper offer convincing evidence on the effectiveness of their LSTM translation model including the two-dimensional PCA projection of the hidden states on the processed phrases and the relative BLEU scores compared with the baseline approach. We could see that the LSTM model performs generally better than the baseline model, but it is outperformed on words with lower frequency or sentences with extremes in long length. These visualizations give the readers a better understanding of the authors' model.

(4) Major critiques

The proposed LSTM model seems feasible on general sentences and phrases, but not so reliable on extreme cases. One improvement might be to apply the attention model idea from the first paper and see if feeding the entire hidden states vectors from each encoder will produce better translation results for longer sentences. Also, English and French belong to the same language system, yet translations from different language systems are needed in most cases. Therefore, it is necessary to test this approach on varying languages, like from Chinese to English, to better improve the robustness of this model.

Reading 2

(1) Title, author and where it was published

- Title: The Illustrated Transformer
- Author: Jay Alammar
- Github blog post

(2) Summary of the content

This blog post builds upon the Attention model blog and talks about the setups and learning process of a Transformer in a simplified way with detailed graphics and visualizations. It is basically explaining the ideas in the paper "Attention Is All You Need". Transformer boosts the attention models' learning speed through parallelization. Like an attention model, a transformer is commonly used in machine translation tasks, and it could be broken down into two parts – the encoders and the decoders.

The encoders are identical in structure and each of them has a self-attention layer and a feed-forward neural network layer, whereas the decoders have an extra layer of Encoder-Decoder Attention in-between the two layers. The author then explains the typical NLP approach (embedding algorithm) of translating a word into a vector; together with a positional encoding vector that accounts for the current order of words in the sentences, these two vectors are then passed into the self-attention layer. It follows by a detailed walk-through on the calculations happening in the self-attention layer, and it goes deeper into the "multi-headed" version of an attention model where a Transformer itself typically has eight heads, or eight sets of query/key/value matrices.

The operations in the decoder side are similar to the encoder ones, and there is a final softmax layer outputting the predicted word with the highest probability.

(3) Strengths of the paper

This paper provides a detailed description in complex concepts like self-attention, which make the paper more read-friendly. The author breaks the self-attention calculation into six steps using a vector-version explanation rather than matrix-version for simplicity. In the first step, a query vector, a key vector, and a value vector is created for each encoding; these vectors are created by multiplying the embedding by three matrices that we trained during the training process. Then a score measuring how much focus to place on other parts of the input sentence is calculated first, the soft-maxed score will be multiplied by the value vector, and the last step is to sum up this weighted vector and passed the output as an input for the feed-forward neural network layer. Essentially, by multiplying with the value matrix, this layer learns how to amplify relevant information and drown out irrelevant ones, bringing "attention" to the correct phrases.

(4) Major critiques

One improvement that I would suggest on this kind of introductory paper is to have a consistent learning example throughout the explanation so that the readers would have a better sense of what stage they are currently in. However, the video in the beginning does help connect the procedures of each phrase, so overall this paper does a very good job in explaining this complicated concept.

Reading 3

(1) Title, author and where it was published

- Title: Generative Adversarial Text to Image Synthesis
- Authors: Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, Honglak Lee
- Published: University of Michigan

(2) Summary of the content

This paper introduces a GAN architecture and training strategy that aims to convert words and characters into image pixels. One critical problem in this kind of learning is that such tasks are highly multimodal, where multiple configurations of pixels are applicable to a certain sequence of text. The authors first mentions a wide range of literature reviews targeting on tasks like multimodal autoencoders, generative adversarial networks, deep convolutional encoder and decoder, as well as recurrent neural networks, briefly explaining the role played by each technique and making it clear that their GAN architecture is built upon these findings. Basically, their approach is to train a deep convolutional generative adversarial network (DC-GAN) conditioned on text features encoded by a hybrid character-level convolutional recurrent neural network. After a walkthrough of the model's high-level architecture, the authors present some variations they developed to improve the performance, including a Matching-aware discriminator (GAN-CLS) and manifold interpolated text embeddings (GAN-INT). The variant GAN-CLS takes into account extra error information than the normal GAN, which is the real images with mismatched text. Then, the results of their experiment are discussed along with the conclusion that the model can synthesize many plausible visual interpretations of a given text caption.

(3) Strengths of the paper

Instead of writing out the network's architecture word by word, the authors choose to illustrate its setup using a clear and simplified high-level visualization of the generator network and the discriminator network accompanied with the writeup of specific layer setup inside each network, giving more spaces for the detailed explanation. It demonstrated that the text encodings for the original descriptions are passed to both networks for further stages of convolutional processings.

Furthermore, the paper presents an extensive number of comparisons among the images generated by different variations (GAN, GAN-CLS, GAN-INT, and GAN-INT-CLS) across two datasets on five examples, providing the most straightforward representation of the effectiveness of each method and thus convincing support for the statement that interpolation regularizer was needed to reliably achieve visually-plausible results.

(4) Major critiques

Many of the literatures and methods the model based upon are relatively old and potentially outdated techniques, so there might be improvements on the small components inside the network. Moreover, the effectiveness of the GAN architecture on long sentences/paragraphs is not tested or not mentioned in this paper. Inspired by the above readings, we might be able to use the attention network with reversed sentence input to process longer descriptions and generate more precise images based on these text inputs. Lastly, birds and flowers tend to have many detailed descriptions like color and size in it, whereas other objects might have more variations in their descriptions. For example, clothes tend to have a wide range of shapes, sizes, and even textures. Further research might be needed to testify the applicability of this model on other datasets.