# TECHNICAL REVIEW ON ATTENTION BASED APPROACHES TO TEXT CLASSIFICATION

**Abhishek Jain |aj26 |aj26@illinois.edu**
**CS410: Text Information Systems UIUC**

## Introduction

Text classification is one of the fundamental problems in Natural Language Processing area and it has wide applications like topic labeling, sentiment analysis, spam identification, intent classification etc. Text documents which are classified can be short texts like tweets, messages & review comments, medium size like blogs & news articles and can be large texts like reports, research papers, recorded conversations etc. There is no one-size-fit-all solution which exists and can be used for these different text types & classification problem complexities. With time text classification approaches has evolved from rule-based systems, conventional machine learning, transfer learning-based to deep learning implementations aka neural NLP solutions.

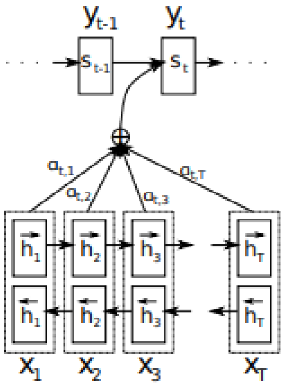
Conventional machine learning based approaches utilizes bag of words, TFIDF & word embedding techniques to represent word & documents as vectors and then a classifier for identifying the document class. These conventional approaches have resulted into considerable results but only for simple problems. These approaches do not capture word sequence ordering, contextual details & relationships.

With deep learning approach using RNN architectures this problem of ordering is solved to some extent but due to vanishing/exploding gradient problem, RNN's faced performance challenges for medium to long sentence sequences. Further to address this issue LSTM based approaches were followed where LSTM memory cells enabled network to remember prior words even for longer sentences. LSTM based *Encoder-Decoder* [1] or *Seq2Seq* [2] architecture, has achieved excellent results on a wide range of NLP problems. Nevertheless, it suffered from the constraint that all input sequences are forced to be encoded to a fixed length internal vector. This is believed to limit the performance of these networks especially when considering long input sequences.

In recent days with breakthrough in next generation NLP architectures, *Attention Mechanism* has evolved as a breakthrough principle in solving above mentioned challenges and this has revolutionized building state-of-the-art NLP solutions. In this technical review I will further detail about attention mechanism and evolution of new era of NLP architecture with emphasis on text classification problem.

## Attention Mechanism

The Attention mechanism is a based on principle of paying attention towards most relevant factors when processing the documents. It was originated from the paper by Bahdanau, et al. on *Neural Machine Translation by Jointly Learning to Align and Translate* [3] which is an extension to Encoder-Decoder or Seq2Seq architecture.



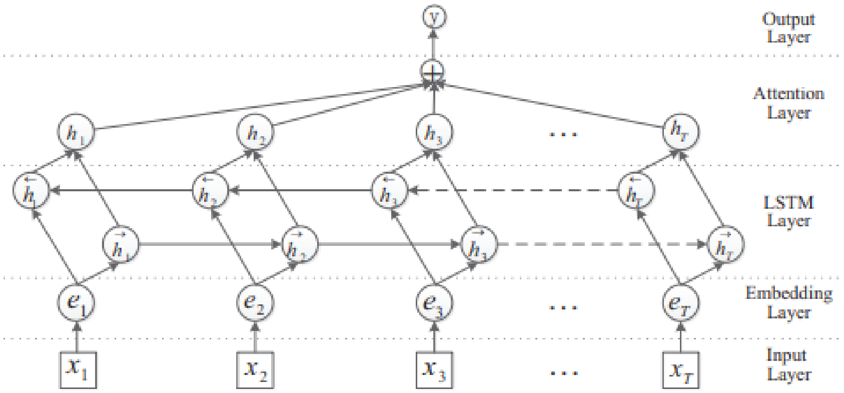
According to this paper attention mechanism emphasizes on two steps alignment and translation. Alignment focuses on identification of part of sequences which are of high relevance and translation consumes the high relevance information and generates the output. This align and translate information is encoded into context vectors for each output time step as opposed to a single fixed length context vector in Encoder-Decoder and Seq2Seq model.

## Attention based Text Classification approaches

Attention Mechanism has gained popularity over time and this has resulted into several solution architecture styles which can be used to solve Text Classification problem. The principle in all the solution is to transform input text sentences to a vector representation with emphasis on important words and their position. Below are some of these solution architecture and details on setup details for Text Classification problem:

### Bidirectional LSTMs with Attentions (AttBLSTM)

This architecture style was proposed in paper proposed by Zhou, et al as novel neural network AttBLSTM [4] for relation classification. This model utilizes neural attention mechanism with Bidirectional Long Short-Term Memory Networks (BLSTM) to capture the most important semantic information in a sentence.



Below are the components of architecture:

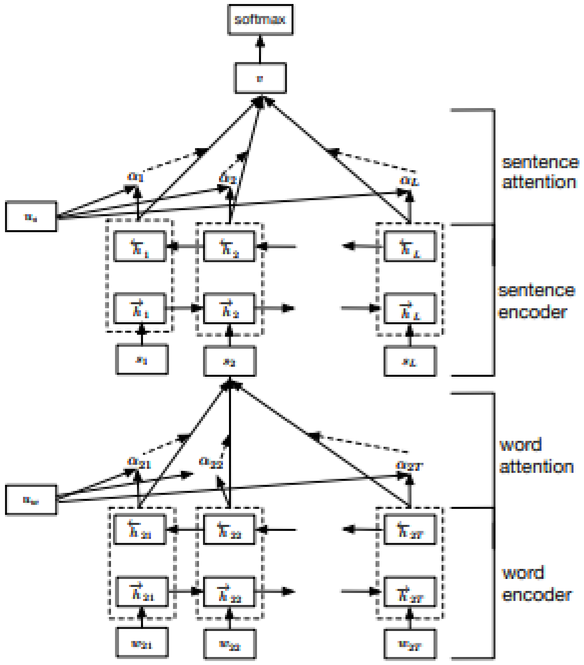
1. **Input layer**: input sentence to this model
2. **Embedding layer**: map each word into a low dimension vector.
3. **LSTM layer**: utilize BLSTM to get high level features from step.
4. **Attention layer**: produce a weight vector and merge word-level features from each time step into a sentence-level feature vector, by multiplying the weight vector
5. **Output layer**: the sentence-level feature vector is finally used for relation classification

## Hierarchical Attention Networks (HAN)

A hierarchical attention network [5] for document classification constitutes two distinctive characteristics:

1. It has a hierarchical structure that mirrors the hierarchical structure of documents.
2. It has two levels of attention mechanisms applied at the word and sentence-level, enabling it to attend differentially to more and less important content when constructing the document representation.

Hierarchical attention architecture extensively uses GRU based encoding scheme which track the state of sequence. It further creates hierarchy following document structure which is further divided into below 4 components:



a) **Word Encoder**: Bi-directional GRU over embedded word vectors which results in contextual information over word.

b) **Word Attention**: Word attention mechanism extract important words and aggregate the representation of those informative words to form a sentence vector context.

c) **Sentence Encoder**: Bi-directional GRU to generate contextual information over sentence

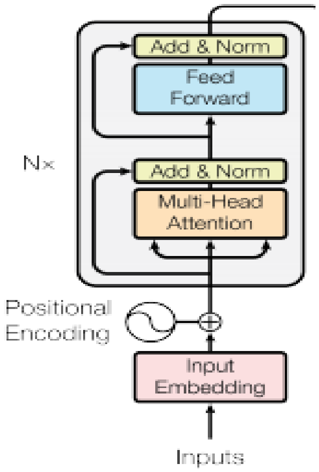
d) **Sentence Attention**: Document level vector context building process based on rewarding approach to sentences that are clues to classification.

e) **Classifier**: Softmax based classifier for classes identification.

## Self-Attentions – Transformers

Self-attention also sometimes called as intra-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence. Transformer network architecture proposed by Vaswani, et. al [6] based solely on self-attention mechanisms, dispensing with recurrence and convolutions entirely. Benefit of using self-attention style is it reduces over all network complexity per layer and parallel computation capabilities compared to RNN & CNN based implementation.

For Text Classification encoder component of Transformer is extensively important, but for other NLP tasks like language translation both encoder & decoder-based architecture are considered. Below are some details of encoder component of Transformer



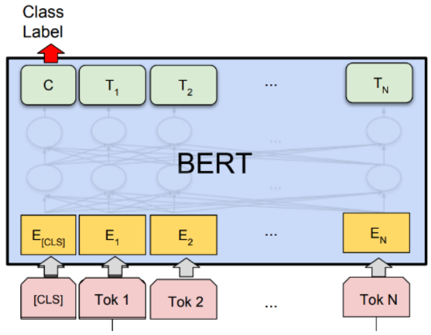
a) **Input Embedding**: Word embedding layer to translate word into vectors.
b) **Positional Encoding**: Captures the order of sequence to get absolute & relative position of tokens.
c) **Multi-Head Attention**: Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions.
d) **Feed Forward Net**: for additional processing of output to be consumed by next layers.
e) **Classifier**: Prepend a special token class and use the hidden state for the special token class as input to your classifier.

## BERT – Bidirectional Encoder Representations from Transformer

An extension to transformer architecture discussed above, BERT[7] stands for Bidirectional Encoder Representations from Transformer. It is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks without substantial task specific architecture modifications.

BERT is treats best-in-class below NLP techniques as the foundation to build a state-of-art pretrained tool for NLP tasks:

a) **ELMo**: Used for word embedding. ELMo looks at the entire sentence before assigning each word in it an embedding.
b) **ULMFiT**: ULM-FiT introduces a language model based process to effectively fine-tune that language model for various NLP tasks.
c) **Transformers**: To deal with long-term dependencies better than LSTMs as discussed in previous sections. Encoder-Decoder structure of the transformer made it perfect for machine translation.



For text classification implementation setup using BERT is straight forward. Input features are entered into BERT and output is further propagated through a feed forward neural network with a softmax.

Below are two more variants of BERT which can be considered to handle long document classification:

a) **Recurrence over BERT (RoBERT)** [9]: BERT is limited to a particular input length for long document is splitted into segments of a fixed size with overlap. These segment outputs are propagated through a LSTM network & then softmax based classification is applied.
b) **Transformer over BERT (ToBERT)** [9]: Given that Transformers' edge over recurrent networks is their ability to effectively capture long distance relationships between words in a sequence a transformer based variant is also considered for long document classification implementation.

### Generative Language Model

Recently OpenAI's GPT-n models have created a series of generative transformers based pretrained models over huge corpus. The quality of the text generated by GPT-3 is so high that it is difficult to distinguish from that written by a human. Puri et. al [10] has proposed a methodology to use Zero-Shot Technique for Text Classification using third generation generative language models.

This method reformulates text classification problems as multiple choice question answering. To enable this model to generalize to new classification tasks, we provide the model with a multiple choice question description containing each class in natural language, and train it to generate the correct answer, also in natural language, from the provided description. Also, with zero shot case where there is even no demonstration is allowed and only multi choice natural language description is enough for GPT3/2 models to predict using their commonsense reasoning techniques.

## Conclusion

Attention mechanism proposed by Bahdanau et. al in 2015 has come a long way and today it is at core of every third-generation state-of-art NLP tool. Architectures around attention has evolved from a conventional training-based methodology to a general purpose pre trained NLP tools which require Zero to Few shots for training. In this review we have also seen how implementations of attention-based architecture has gradually evolved from having extensive dependencies on RNNs & LSTMs (BLSTMAtt) to zero neural nets (Transformers or Self-Attentions) based implementation.

We have discussed how all these different techniques can be further architected in context of a Text Classification problem. Text Classification problems have different level of complexities ranging from the size of texts, number of classes, and problem nature. Below are some conclusion remarks on individual techniques which shall help in selecting one model over other:

- Based on studies Hierarchical Attention Networks has proved to perform better over Neural Network based architecture on medium size documents like review comments.
- Att-BLSTM for relation classification model does not rely on NLP tools or lexical resources. It uses raw text with position indicators as input for classification. In case text to be classified are structured in away where NLP tools cannot be applied though information is sequential this technique can be a choice.
- BERT is state-of-art NLP tool & can be a defacto choice for documents with size less than 512 words. For larger documents, the other two variants RoBERT or ToBERT can be preferred.
- GPT-2 & BERT variants have shown comparable performances, recommendation would be to use both these model variants and observe which works best for individual use case.

## References:

1. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation
   http://emnlp2014.org/papers/pdf/EMNLP2014179.pdf
2. Sequence to Sequence Learning with Neural Networks
   https://papers.nips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf
3. Neural Machine Translation by jointly learning to Align and Translate
   https://arxiv.org/pdf/1409.0473.pdf
4. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification
   https://www.aclweb.org/anthology/P16-2034.pdf
5. Hierarchical Attention Networks for Document Classification
   https://www.aclweb.org/anthology/N16-1174/
6. Attention Is All You Need
   https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
7. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
   https://arxiv.org/pdf/1810.04805.pdf
8. BERT Explained
   http://jalammar.github.io/illustrated-bert/
9. Hierarchical Transformers for Long Document Classification
   https://arxiv.org/pdf/1910.10781.pdf
10. Zero-shot Text Classification With Generative Language Models
    https://arxiv.org/pdf/1912.10165.pdf