ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

Survey on Attention Neural Network Models for Natural Language Processing

Ms Swati Meshram

Assistant Professor, P.G. Department of Computer Science, SNDT Women's Unviersity, Mumbai India.

ABSTRACT: Natural language processing(NLP) task like machine translation, sentence summarization,sentence pair modeling, paraphrase identification, natural language inference, question answering etc typically learn the sentence embedding for representation of context and further to transform the text for providing the solution using neural network encoder decoder architecture. The models typically used are Long Short Term Memory(i.e.LSTM), Convolution Neural Network (i.e.CNN), RNN Recurrent Neural Natwork (i.e RNN) etc. These neural network models performance degrades if length of the sentence is over 30 words is fed as input for NLP task in general. Attention models have shown better performance for large sentences. Attention mechanisms have recently attracted wide interest in NLP community due to its ability to perform parallel computation, requires significantly less training time, and flexibility in modeling context dependencies. This paper is a comprehensive survey of attention models for natural language processing task.

KEYWORDS: NLP, sentence encoding, attention, machine translation, classification, inference.

I. INTRODUCTION

Natural language processing(NLP) is interaction of computers and natural languages for processing large amount of natural language data. An artificial neural network model (NN) are typically employed for representation of natural language models. NN is a computational model inspired by human brain ability to learn to perform task like classification, prediction, pattern recognition, regression, clustering, machine translation(MT) etc.

Typically, Sentence encoding is learning the context aware representation using various neural network models like RNN, CNN, LSTM etc. RNN sentence encoding is achieved by reading words in sequence to represent the context of all prior inputs through hidden state vector. It utilizes the last hidden state vector as a proxy for the entire sequence. The longer the input sequence, the same vector is updated. And higher the chance of losing the earlier input. The sequential operations leads to difficulty in parallelize and take more time to train. CNN targets on positional invariant dependencies still does not perform well in many of other tasks[1]. The problem with the long sequences is that the performance of the neural network degrades as it has to memorize long sequence.

To provide solution to the above stated problem in the recent year, the research community has taken vast interest in attention neural network as it can model the context dependencies, trained faster and employs parallelizable computation. Attention neural networks has seen successful applied in handwriting synthesis [2], digit classification [3], machine translation [4], image captioning[15], speech recognition [16]and sentence summarization[17], to geometric reasoning [18] etc.

II. GENERAL FRAMEWORK

Attention mechanism was first introduced in [4] for machine translation of longer sequences. In the encoder decoder architecture {x1,x2,..xn} is the input sequence fed to encoder to generate a fixed length context vector {h1,h2,..hn} which represents the hidden states. The decoder generates a target sequence {y1,y2,..ym}

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

As the name suggest, the attention model concentrates on a part of sentence at a time for NLP task. It does part by part translation. The attention neural network architecture utilizes all the hidden states without lessening the context. The layers of interconnected hidden units could be bi-LSTM/bi-GRU/bi-RNN. These units being bilateral learn from last word to first word and other layer from first word to last word. It creates rich set of features of the current sample and surrounding sentence.

In the machine translation process, first piece has to be translated by focusing on first few words and not on ending or last words of the same sentence. The attention model computes attention weights αij of hidden states for each piece of information to generate the context. The decoder predict with the conditional probability based on previous predicted word yi-1,hidden state si, and the context vector ci for each target word yi. It combines all hidden units weighted by the score. An attention operation for the decoder output represents every hidden state with varying weights to different encoder.

2.1 **Types of attention**

Global attention[19] considers all the hidden states and hence is expensive vs **Local attention**[19] is to attend while concentrating few hidden states falling within the window.

Soft attention[4] is to calculate the context vector as a weighted sum of the encoder hidden states vs in **Hard attention**[15], contrary to weighted average of all hidden states, employs attention scores to select a single hidden state.

The main contribution of the present survey is to provide a detail study on application of attention mechanism using neural network for natural language processing tasks.

III. SURVEY

[5] proposed a matching aggregation framework which is a multiway attention network (MwAN). Their model applies the attention mechanism at the word level which leads to improve the matching between words in two sentences. This information is further aggregated to the sentence level for making the decision. Their framework implement the word level interaction. They have used four attention functions to match words in corresponding sentences viz concat, bilinear, dot, minus. The contextual word representation is learnt for two sentences using a bi-directional RNN. The activation function used is gated recurrent unit. The dataset is Quora Question Pairs with over 400,000 question pairs for paraphrase identification. Their experiments demonstrate 89.12% model accuracy. The other natural language inference dataset used were SNLI, MultiNLI and SQuAD a reading comprehension dataset for answer sentence selection.

[6] The authors have proposed a recurrent neural network model to read two sentences to recognize entailment using long short-term memory units. The word-by-word neural attention mechanism support reasoning over entailments of pairs of words and phrases. LSTM recurrent neural networks reads pairs of sentences to produce a final representation from which a simple classifier predicts entailment. They use concatenated attention function. Their LSTM model has achieved an accuracy of 80.9% on SNLI dataset.

[7] Their model ESIM –enhanced sequential inference model is based on BiLSTMs and TreeLSTMs, determines a natural language hypothesis h is inferred from a natural language premise p. Local inference modeling is implemented through soft alignment layer attention with difference and dot product functions. Incorporating inference modeling and inference composition and syntactic parsing information , their model has performed with 88.6% accuracy on the SNLI dataset.

[8] This paper presents a “compare-aggregate” framework to perform word-level matching and aggregation using Convolution Neural Networks. Their model first compare vector representations of smaller units such as words from these sequences and then aggregate these comparison results to make the final decision. Their model is four layers, first layer preprocesses each word of the sequence into word embedding vector to obtain contextual information other than

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

the word itself. Second layer, attention mechanism applies element-wise dot product to obtain attention weighted sum of vectors. Third layer, comparison is a neural network layer that applies linear transformation followed by a non-linear activation function. A comparison function based on element-wise subtraction and multiplication works the best as per their experiments. Lastly the aggregation is implemented using one layer CNN which is for final classification. The experiments were conducted on SNLI, InsuranceQA, WikiQA and MovieQA datasets. With highest accuracy of 86.8 % on SNLI dataset.

[9] The transformer model is a work by Google researchers. An encoder-decoder structure, which is six layers of stacked multi-head attention and point-wise, fully connected layers of feed forward network and layer normalization for the encoder. It takes every single word and compare with other words to obtain the similarity with every other word in order to determine the context and encodes it. They use positional encoding, to visualize the sentence as a wavelength. The output of the embedding layer is 512 dimensions. The use of multi-head attention is to represent a sentence into multiple semantic subspaces and later apply self attention in every subspace. The attention scores help to visualize what is learned. The results are then concatenated and probability is predicted for the similarity with other words. Here output of one token is a sample; it does not employ multi step back propagation as in RNN. At the decoder, the multi head attention has a query Q which is searched in the keys K with value V. On the WMT 2014 English-to-German translation task, the big transformer model, has a BLEU score of 28.4. WMT 2014 English-to-French translation task, their model achieves a BLEU score of 41.0

[10] The authors proposed a hierarchical attention network for document classification which employs word and sentence level attention to represent content when constructing the document representation. Their intuition is documents have hierarchical structure, likewise document representation is constructed by building important word vectors aggregated to form sentence vectors on and then aggregating those to obtain document representation. The context of the token decides the importance of token in a sequence i.e context dependent word importance. Their model has attended 63.6 % accuracy on Amazon reviews and 75.8% accuracy on Yahoo answers

[11] According to the authors, in their model every sentence is segregated into multiple phrases based on syntactic parse tree, further performing self attention at the phrase level instead of the sentence level, thus reducing the memory consumption as the number of phrases increases. Further engaging Gated memory to refine each word representation hierarchy to integrate long term dependencies.

[1] In their paper, authors have proposed a directional self-attention model with temporal order encoded using multi-head scaled dot-product attention to represent each word by weighted summation of all words in the sentence. They have also incorporated multi-dimensional attention to perform a feature-wise selection on the input sequence and the directional self-attention uses the positional masks to produce the context-aware representations. The model was fed SNLI dataset to obtain the accuracy of 85.62%

[12] Authors have built Gated Convolutional Neural Network (GCNN) architecture to simulate n-gram with a self-attention mechanism to understand sentiment polarities from Amazon reviews. The convolution layer generates the input feature vector and the gated layer is implemented to select the feature with highest probability. They have compared their model with linear support vector machine, GCNN, where their model gives accuracy of 90%

[13] Their model is a description-strengthened and fusion attention recurrent neural network .It aims to retrieve relevant person instances with natural language description query in a large-scale person dataset, to facilitate human accessible solutions for suspect identification. Instead of employing person image and attribute, searching person with comprehensive natural language description collected from eye witness rather than be restricted by few attributes. It is a text-based image retrieval task. Here discriminative words in the description are given paid more attentions. Their model DSFA-Net has achieved 20.33% in top-1 accuracy, 42.4% in top-5 accuracy, 54.90% in top-10 accuracy.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

[14]Graph neural network applies deep learning on the graphs where the input features vectors for every node is given and hidden representation for each node is computed. The purpose is to use this hidden representation in machine learning algorithm for classification. Graph attention network deploys multiple convolution layer and inductive learning. The graph attention network uses convolution layer for aggregation of neighboring node features and applying a non linear function to generate the output features. Here the attention is given to the node and its neighbors which are at distance of 1 or are directly connected. Their features are feed to convolution of a node and the output is only a feature vector for the current node. For every edge in the graph whose two neighboring nodes hi and hj vectors are multiplied by the weights W, the result is concatenated and dot product on it with vector a is taken where a is attention with a∈ R 2XF' dimensions. Further nonlinearity is applied with LeakyReLU function for every edge eij which is given to the softmax function to obtain the self attention head αij. The aggregated features of each head is concatenated and averaged to obtain the output $\vec{h}'_i$. Experiments were performed on Cora, Citeseer, Pubmed, PPI graph dataset where the model GAT could attain classification accuracy of 83%.

Table 1: An overview on the characteristic of various attention models.

Work	Sentence Encoder	Attention/ attention function	Classification	Resource /Dataset	Accuracy/ BLEU %
[Tan et al, 2018]	Bi-RNN	Word-word /concat, bilinear, dot, minus	bi-directional GRU+MLP	Quora SNLI, MultiNLI, SquAD	89.12
[Rocktaschel ¨ et al., 2015]	RNN with LSTM	Word-word attention, two-way attention functionconcat	softmax layer	SNLI	83.5
[Chen et al., 2017]	BiLSTM, TreeLSTM	Difference,dot product, soft alignment	Pooling, MLP	SNLI	88.6
Wang and Jiang, 2017a	LSTM/GRU	softmax attention +dot product	One-layer CNN	SNLI, InsuranceQA, WikiQA, MovieQA	86.8 % on SNLI dataset
Vaswani et al, 2017	Postionwise fully connected feed-forward network.	Scaled dot product,Multi headed self attention	Positional FFN, softmax layer	WMT 2014 English-to-German and English-to-French translation	BLEU – 28.4,41.0
[Yang et al 2016] HAN	Bidirectional GRU	Hierarchical attention	Softmax layer	Yahoo answers	75.8%
[Wu et al 2018]PSAN	Gated memory	multi-dimensional attention	Two layer feedforward network, softmax layer	SNLI	86.1%
[Shen et al 2017] DiSAN	GloVe	multi-dimensional self-attention, directional self-attention	Fusion gate	SNLI	85.62%
[Yanagimto et al, 2018]	Convolution layer	Self Attention	Gated layer	Video Games reviews from Amazon	90%
[Z. ji et al, 2018] DSFANet	LSTM	Word-word attention	softmax layer	CUHK-PEDES	0.20%,top1 0.42%in top-5 0.54% in top-10

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

IV. CONCLUSION

Attention model has drawn interest from researchers as it has potential to not lower the performance of neural network models for longer sequences and it has been successfully applied for wide range of applications. This survey gives an overview on various approaches of attention mechanism for various natural language processing task based on neural network models.

REFERENCES

[1] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang." Disan: Directional self-attention network for rnn/cnn-free language understanding." CoRR, abs/1709.04696,2017
[2] A.Graves, " Generating sequences with recurrent neural networks," Technical report, arXiv preprint arXiv:1308.0850, 2013.
[3] V. Mnih, N. Heess, A. Graves, et al. "Recurrent models of visual attention," In NIPS, pages 2204–2212, 2014. (Conference proceedings)
[4] D.Bahdanau, K.Cho, and Y.Bengio,"Neural machine translation by jointly learning to align and translate," ICLR 2014 September 2014. (Conference proceedings)
[5] C. Tan, F. Wei, W. Wang,W.Lv, M. Zhou,"Multiway Attention Networks for Modeling Sentence Pairs," Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18) (Conference proceedings)
[6] T. Rocktaschel, E. Grefenstette, K.M.Hermann, T.Kocisky, and P. Blunsom,"Reasoning about entailment with neural attention," CoRR, 2015. (Conference proceedings)
[7] Q. Chen, X. Zhu, Z. Ling, S. Wei, H. Jiang, and D. Inkpen."Enhanced lstm for natural language inference." In ACL, pages 1657–1668, Vancouver, Canada, July 2017. Association for Computational Linguistics. (Conference proceedings)
[8] S. Wang and J. Jiang,"A compare-aggregate model for matching text sequences,", In ICLR, 2017. (Conference proceedings)
[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. "Attention is all you need." In Advances in Neural Information Processing Systems, pages 6000–6010. 2017.
[10] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. 2016."Hierarchical attention networks for document classification", In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies
[11] W.Wu,H.Wang ,T.Liu and S.Ma,"Phrase-level Self-Attention Networks for Universal Sentence Encoding", Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3729–3738, Association for Computational Linguistics, Belgium November 2018.
[12] H. Yanagimto, K. Hashimoto, M. Okada "Attention Visualization of Gated Convolutional Neural Networks with Self Attention in Sentiment Analysis" 2018 International Conference on Machine Learning and Data Engineering (iCMLDE)
[13] Z. Ji, S. Li, Y. Pang, "Fusion-Attention Network for person search with free-form natural language" Pattern Recognition Letters 116 (2018) 205–211
[14] P.Velickovic,G. Cucurull,A. Casanova, A.Romero,P. Lio, and Y.Bengio (2018). Graph attention networks. In Proceedings of the International Conference on Learning Representations (ICLR).
[15] K.Xu, J.Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. "Show, attend and tell: Neural image caption generation with visual attention," In ICML, 2015.
[16] J.Chorowski, D. Bahdanau, D. Serdyuk, K.Cho, and Y. Bengio," Attention-based models for speech recognition," In NIPS, 2015
[17] A.Rush,S.Chopra, and J.Weston,"A neural attention model for abstractive sentence summarization," In EMNLP, 2015.
[18] I. Sutskever, O. Vinyals, and Q. Le. "Sequence to sequence learning with neural networks". In NIPS, pages 3104–3112, 2014.
[19] T.Luong, H.Pham, and C.D.Manning. Effective approaches to attention-based neural machine translation. In EMNLP, pages 1412–1421, Lisbon, Portugal, September 2015. ACL.