Supervised Mover's Distance: A simple model for sentence comparison

Muktabh Mayank Srivastava

ParallelDots, Inc.
muktabh@paralleldots.com,
website: https://paralleldots.xyz

Abstract. We propose a simple Neural Network model which can learn relation between sentences by modeling the task as Earth Mover's Distance(EMD) calculation. Underlying hypothesis is that a neural module can learn to approximate the flow optimization in EMD calculation for sentence comparison. Our model is simple to implement, light in terms of parameters and works across multiple supervised sentence comparison tasks. We show good results for the model on two datasets.

Keywords: Supervised Mover's Distance, Sentence Comparison, Paraphrase Detection, Natural Language Inference

1 Introduction

Sentence Comparison is a common NLP task which comes up in multiple domains. We propose a simple architecture for sentence comparison that works across domains. Sentence comparison measure might be needed to check redundant data [1] or check sentences for paraphrases [2]. Our model tries to model sentence comparison task as a Earth Movers' distance (EMD) calculation on representations obtained by LSTM layers. The representations from LSTM model are sent to a neural module (called Relational Layers) which approximates the EMD. We show good results for the model on both the tasks aforementioned. The contributions of this work are: 1. Model the task of supervised sentence comparison as EMD task. 2. Propose a simple module which can approximate EMD Calculation.

2 Previous Work

Earth Mover's Distance (EMD) is a measure of distance between two probability distributions imagined as dust piles in a region which represents the minimum cost of turning over piles of one distribution into another. Word Mover's Distance (WMD) [3] is a method to calculate similarity between two sentences by comparing combinations of pretrained embeddings of words in each. The WMD, however, cannot be tuned for supervised sentence comparison. A supervised version for WMD was proposed to calculate document similarity [4], which classified

documents by k-Nearest Neighbors. The supervised version has a complex loss function optimization procedure and works on datasets with larger documents (average length of documents is greater than 40 for almost all datasets used). Our work is on comparing sentences and we propose a simpler architecture. Different sentence comparison tasks have been studied extensively previously. The two most common supervised sentence comparison tasks are: 1. Similarity Detection [1] and 2. Paraphrase Detection[2]. As Supervised Mover's Distance, we propose a baseline that generalizes well across different tasks. Our network combines LSTM layers [5] with a neural module to learn EMD over hidden layers, hence modeling semantic relationship between the sentences. The module to learn EMD has similar architecture to RelationNets' relational learning module. [6]

3 Method

The neural network architecture we propose is trained on pair of sentences to predict one of various classes the pair might fall into. For redundancy detection and paraphrase detection the labels are positive or negative, but might be different for any other tasks. The architecture has two basic parts: 1. LSTM layers and 2. Relational Layers. The LSTM layers can consist of one or more LSTM layers which take both sentences as input individually and produce hidden layers as output for each of the words in the sentences. This would yield two series of output hidden states, one hidden for each time step of each sentence. To clarify again, there is one common LSTM which runs on both sentences separately. In the relational layer, all possible pairs of hidden states across both sentences are taken as concatenated vectors and passed through a fully connected (or Dense) layer. This yields an embedding for each possible pair of hidden state outputs from the LSTM. These embeddings are averaged and passed through another fully connected layer to predict the output. Please note that there is just one common fully connected layer that takes each pair as input individually and returns the corresponding embedding. By taking all pairs of hidden states and using them to model sentence comparison task, we hypothesize that the relational layer is able to model Earth movers' Distance between the sentences. While [4] models the flow optimization as a complex optical flow model, we try to model it by simple Relational Layers.

We illustrate the architecture in 1. Our model is light in terms of parameters as it has only a LSTM layer and two dense (fully connected) layers in Relational Layers. A limiting case of the architecture can be when the number of LSTM layers is zero, and word embeddings are passed as inputs directly to Relational Layers.

The network is trained with common hyperparameters for all the tasks. Pretrained word embeddings are used to initialize the word embedding layer which are finetuned by backpropogation. We use the publicly available 6 Billion token 100 dimensional version of GloVe embeddings [7]. The hidden state output from the LSTM is 100 dimensions and the size of embedding generated in the rela-

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26847v1 | CC BY 4.0 Open Access | rec: 12 Apr 2018, publ: 12 Apr 2018

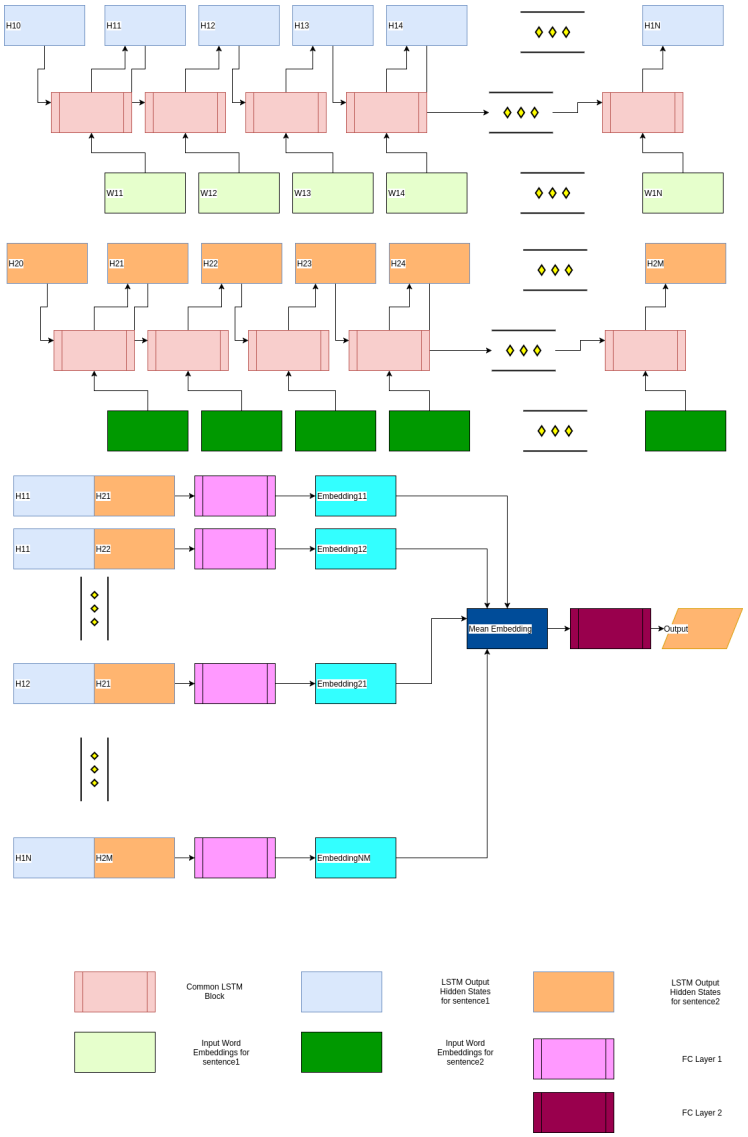

Fig. 1. Illustration of the Neural Network architecture for Supervised Mover's Distance between sentences (user might need to zoom-in to view)

tional layers is 100 dimensions too. The network is trained with simple Stochastic Gradient Descent with momentum (common values for training across both datasets, learning rate = 0.001, momentum=0.9).

4 Results

As stated we test our model on two datasets. Model is compared to state of the art methods and baselines for each dataset in this section.

Microsoft Research Paraphrase Corpus Microsoft paraphrase corpus [2] is a corpus of sentence pairs classified as paraphrases or non-paraphrases. The model has 4076 sentences in training set and 1725 sentences in test set. Our model was trained on the training set with the standard set of hyper parameters mentioned above and evaluated on the test set. The accuracy numbers of different models were taken from this url[1]. Our model gets an accuracy of 80.2% on the dataset as compared to state of the art accuracy of 80.4% [8].

Quora Questions' Pair Dataset Quora Questions' Pair Dataset contains question pairs from the Q&A website[2] tagged as similar or not. A random 90%-10% train-test split is performed as is customary for other methods and the model is trained on the train set and evaluated on the test set. As in case of other datasets, the hyperparameters are fixed as the standard values specified earlier while training. Our model gets an accuracy of 81.2% on the dataset. List of state of the art models on the dataset is available on this url[3]. The best accuracy a model gets on the dataset is 88% [9]. Although our model doesn't get results as good as the state of the art, it is competitive to baselines like siamese Convolutional Neural Networks (79.6%) and siamese LSTMs(82.58%).

It should be noted that in both models, dataset specific hyperparameter tuning was not performed.

5 Discussion

We propose a new method which uses a new and simple neural network model to compare sentences. The model tries to calculate approximate Earth Movers' Distance(EMD) between sentences with the hypothesis that EMD can be used to approximate distance between sentences. Results on two datasets are captured and presented.

[1] https://aclweb.org/aclwiki/Paraphrase_Identification_(State_of_the_art)
[2] quora.com
[3] https://github.com/bradleypallen/keras-quora-question-pairs

References

1. Shankar Iyar, N.D., Csernai., K.: First quora dataset release: Question pairs (January 2016)
2. Dolan, B., Brockett, C.: Automatically constructing a corpus of sentential paraphrases. In: Third International Workshop on Paraphrasing (IWP2005), Asia Federation of Natural Language Processing (January 2005)
3. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From word embeddings to document distances. In: Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37. ICML'15, JMLR.org (2015) 957–966
4. Huang, G., Guo, C., Kusner, M.J., Sun, Y., Sha, F., Weinberger, K.Q.: Supervised word mover's distance. In: Advances in Neural Information Processing Systems. (2016) 4862–4870
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8) (1997) 1735–1780
6. Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T.: A simple neural network module for relational reasoning. In: Advances in neural information processing systems. (2017) 4974–4983
7. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). (2014) 1532–1543
8. Ji, Y., Eisenstein, J.: Discriminative improvements to distributional sentence similarity. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. (2013) 891–896
9. Wang, Z., Hamza, W., Florian, R.: Bilateral multi-perspective matching for natural language sentences. arXiv preprint arXiv:1702.03814 (2017)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26847v1 | CC BY 4.0 Open Access | rec: 12 Apr 2018, publ: 12 Apr 2018