Improving Sentence-BERT using multi-head attention model-based pooling and uniformity metrics

Garvit Goyal
ggoyal9@gatech.edu

Gunjan Gupta
ggupta68@gatech.edu

Tusheet Goli
tgoli3@gatech.edu

Abstract

The architecture of the SBERT model was explained in the paper, *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (Reimers and Gurevych, 2019). SBERT model has a siamese network architecture and uses similarity measures such as cosine similarity/Euclidean/Manhattan distance to derive fixed-size vector embeddings for sentence models. It has been shown that SBERT outperformed other state-of-art sentence transformer models on common STS tasks. For this project, we intend to alter the SBERT model by changing the pooling layer and similarity metrics.

1 Research Paper Summary

1.1 Background and Summary

BERT for sentence classification and pair regression tasks uses sentence pairs to predict the target value. However, this technique requires $\mathbb{O}(n^2)$ computations and is unable to compute independent sentence embeddings. To overcome this, models were proposed that used the average of the BERT output layer or [CLS] token embedding to represent sentences as fixed-size vectors. In general, this leads to poor performance, worse than averaging GloVe embeddings. Compared to the other neural sentence embedding models such as SkipThought, InferSent, and Universal Sentence Encoder, the training time of SBERT is lower as it uses a pretrained BERT network and fine-tunes the model to get sentence embeddings.

1.2 Bibliographical Information

Title: Sentence-BERT Sentence Embeddings using Siamese BERT-Networks
Authors: Nils Reimers and Iryna Gurevych
Publication: arXiv
Publication Year: 2019
URL: https://arxiv.org/abs/1908.10084

1.3 Summary of Contributions

The model pools the output of word embeddings from the BERT model to derive sentence embeddings. It uses a siamese or triplet network to finetune the BERT model such that semantically similar sentences are embedded close to each other in the vector space. The input can be informed of sentence pairs or triplets. The objective depends on the dataset available and can be of the following types - classification objective, regression objective, and triplet objective. In this model, Cosine/Euclidean/Manhattan distances can be used as a measure of similarity between vector embeddings. The contrastive loss function introduced by Hadsell et al., 2006 is used here, works to reduce the distance between similar pairs and increase the distance between dissimilar ones. The authors evaluate the model performance on supervised, and unsupervised STS tasks, and SentEval tasks and show that it is able to outperform other sentence embedding models. It is also computationally efficient than InferSent and Universal Sentence Encoder.

1.4 Limitations and Discussion

The paper presents a computationally efficient way of learning sentence embeddings through contrastive learning. The model in general outperforms other sentence embedding models, however, it appears to depend significantly on the training corpus. For instance, SBERT does not perform as well on the TREC dataset compared to Universal Sentence Encoder which was trained on question-answer type data. The SBERT model is also not suited for transfer learning for other STS tasks, which makes it restrictive in the application.

1.5 Why this Paper

This paper describes the detailed implementation of the SBERT model and provides a fundamental understanding of the model we want to implement. By gaining a deeper understanding of the model described in this paper, we can make our own siamese network to improve sentence embedding models.

1.6 Wider Research Context

Semantic textual similarity is the task of measuring closeness of semantic meaning of two text snippets. We train the model to get vector embeddings such that semantically similar sentences are closer in the vector space, thereby deriving meaningful sentence embeddings. This has further applications in other tasks such as semantic search, clustering, paraphrase mining, question answering (retrieve and re-rank) and image search. Having a reliable and computationally efficient sentence embedding model can help to improve the above tasks as well.

2 Project Description

2.1 Goals

The SBERT model is the state-of-art sentence embedding model that is a computationally efficient way to get sentence embeddings. We expect that such a model is able to cluster semantically similar sentences together in vector space. Below is an example of the output from the all-miniLM-L6-v2 (SBERT model). We compare the cosine similarity of three sentences with the reference sentence in Table 1.

Reference Sentence: That is a happy person.

Sentence	Similarity Score
That is a very happy person	0.943
That is a happy dog	0.665
That is a happy girl	0.685

Table 1: Example of semantic similarity between sentences from SBERT model

The model does not significantly differentiate between a happy girl and happy dog. Our goal in this project is to see if we can improve on the SBERT model by 1) changing the pooling layer to go from BERT word embeddings to sentence embeddings and 2) exploring other similarity metrics than the distance that might be more suited for the task.

2.2 NLP Tasks

The NLP task we would be exploring in this project would be semantic textual similarity, where the objective is to quantitatively assess the semantic similarity between two text parts. An example is shown above.

2.3 Data

Similar to our main research paper on SBERT (Sentence-BERT), we would like to evaluate our proposed model on the same datasets used in the SBERT paper ((Reimers and Gurevych, 2019)). The first dataset is the SNLI (Stanford NLI) dataset (Bowman et al 2015) which contains 570,000 sentence pairs annotated with the labels entailment, contradiction, and neutral. This is a publicly available dataset https://nlp.stanford.edu/projects/snli/ that offers the corpus as both JSON lines (jsonl) and a tab-separated text file. It is about 100MB and can be downloaded from https://nlp.stanford.edu/projects/snli/snli_1.0.zip. An example of this dataset can be seen in Table 2.

The second dataset we will use is the Multi-Genre NLI (Williams et al., 2018) which contains 433,000 sentence pairs that are annotated with textual entailment information. This corpus is very similar to the SNLI corpus, but it also covers a range of genres of spoken and written text while supporting cross-genre generalization evaluation. This is also a publicly available dataset https://cims.nyu.edu/~sbowman/multinli/ that offers the corpus as both JSON lines (jsonl) and a tab-separated text file. It is 227MB and can be downloaded from https://cims.nyu.edu/~sbowman/multinli/multinli_0.9.zip. An example of this dataset can be seen in Table 3.

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping.
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

Table 2: Example of the SNLI dataset.

2.4 Methods

2.4.1 Pooling

Pooling is an essential component of a range of different sentence representations and embedding models. A variety of neural network models incorporate pooling techniques like mean or max pooling to obtain sentence embeddings. In the paper “Enhancing Sentence Embedding with Generalized Pooling”, Chen and Ling explore several generalized pooling techniques to improve sentence embedding. In particular, they analyze an extension of the scalar self-attention models previously proposed by Lin et al. (2017) to vector-based multi-head attention models that incorporate other traditional pooling techniques like max pooling, mean pooling, and scalar self attentions.

By utilizing a multi-head attention model, the vectors enhance the expressiveness of the attention mechanism far better than traditional techniques and the incorporation of a penalization term helps to reduce redundancy in multi-head attention techniques. The multi-head approach discussed in Chen and Ling’s paper enables the extraction of the different aspects of a sentence into several vector representations while the vector-based attention mechanism enables better focus on the different interpretations of the words encoded in the context vec-

Premise	Label	Hypothesis
Fiction: The Old One always comforted Ca'daan, except today.	neutral	Ca'daan knew the Old One very well.
Telephone Speech: yes now you know if if everybody like in August when everybody's on vacation or something we can dress a little more casual or	contradiction	August is a black out month for vacations in the company.
9/11 Report: At the other end of Pennsylvania Avenue, people began to line up for a White House tour.	entailment	People formed a line at the end of Pennsylvania Avenue.

Table 3: Example of the Multi-NLI dataset.

tors. This approach produced significant improvements over other sentence-encoding-based methods when evaluated over four benchmark datasets. Two of four these datasets happen to be the SNLI and Multi-NLI datasets that we plan on using as our baseline evalution datasets. As the paper mentions, this proposed approach can be incorporated into other models and be used to improve the performance for more general problems. Thus, we plan on using this approach to pooling to potentially increase the performance of our model.

2.4.2 Loss functions

Contrastive representation learning usually constrains the output feature vectors to be of unit l_2-norm, thus embedding the feature space on a unit hypersphere. Unsupervised contrastive learning methods, specifically, optimize the contrastive loss, which is furthermore seen as a lower bound for information gain associated with two positive pairs sampled from the data.

Wang and Isola, 2020 introduce two metrics associated with alignment (minimizing distance between embeddings associated with similar samples, i.e., positive pair) and uniformity (feature vectors being uniformly distributed on the unit hypersphere, preserving maximal information). They show, empirically and theoretically, that the contrastive loss optimizes alignment and uniformity asymptotically in the limit of infinite negative samples. The alignment metric computes the expected distance between positive pairs. On the other hand, the uniformity metric computes the logarithm of the average pairwise Gaussian potential using the Gaussian potential kernel. The latter has been further demonstrated to be similar to the uniform distribution on the hypersphere and universally optimal point configurations. The two metrics thus introduced are as follows:

Let $p_{data}(\cdot)$ be the data distribution over $\mathbb{R}^n$ and $p_{pos}(\cdot,\cdot)$ the distribution of positive pairs over $\mathbb{R}^n \times \mathbb{R}^n$.

$$\mathcal{L}_{align}(f;\alpha) \triangleq \mathbb{E}_{(x,y)\sim p_{pos}}[\|f(x)-f(y)\|_2^\alpha]$$

$$\mathcal{L}_{uniform}(f;t) \triangleq \log \mathbb{E}_{(x,y)\overset{\text{i.i.d.}}{\sim} p_{data}}[\exp(-t\|f(x)-f(y)\|_2^2)]$$

where $\alpha > 0, t > 0$, the latter being the parameter of the Gaussian kernel.

2.5 Baseline

Since the alignment metric proposed in Wang and Isola, 2020 minimizes the distance between positive pairs, we choose the SBERT-WikiSec-base as our baseline since the latter is trained on the Wikipedia section distinction task using the triplet loss function. The pretrained model has been publicly released at https://www.sbert.net/docs/pretrained-models/wikipedia-sections-models.html. Further, we will use the published results for the baseline in Reimers and Gurevych, 2019.

2.6 Evaluation

As discussed in methods section, this work has two contributions: 1) multi-head attention model-based pooling layer, and 2) two new metrics focusing on uniformity and alignment of the embeddings on the unit hypersphere. Thus, we evaluate the proposed model against the baseline SBERT-WikiSec-base (pretrained model released at https://www.sbert.net/docs/pretrained-models/wikipedia-sections-models.html) on the Wikipedia section distinction task, as given in section 4.4 of Reimers and Gurevych, 2019. We will be using the accuracy metric as defined in Reimers and Gurevych, 2019. Specifically, after training the proposed model on the dataset from Dor et al., the accuracy obtained with the proposed approach will be compared against the published results.

2.7 Source/Version Control

We created a private repository on Gatech GitHub to store our data, code, analysis, and other documents. We have added all our team member as collaborators to the repository and will add the TAs and Professor if required.

This is the link to our repository - https://github.gatech.edu/tgoli3/gt-nlp-f22

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Qian Chen, Zhen-Hua Ling, and Xiaodan Zhu. 2018. Enhancing sentence embedding with generalized pooling. *CoRR*, abs/1806.09828.

Liat Ein Dor, Yosi Mass, Alon Halfon, Elad Venezian, Ilya Shnayderman, Ranit Aharonov, and Noam Slonim. 2018. Learning thematic similarity metric from article sections using triplet networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Melbourne, Australia. Association for Computational Linguistics.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Maximilian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. *CoRR*, abs/1705.08039.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.