Few-shot Learning: Training Deep Learning Classifiers with Little Labeled Data
Project Proposal for NLP Course, Winter 2022

D. Przybyliński, A. Podsiad, P. Sieńko
Warsaw University of Technology
piotr.sienko.stud@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

The purpose of this project for NLP Winter Course 2022 is to investigate methods and algorithms for few-shot learning scenario with deep learning models. The proposed approach will utilize the Bag-of-words method or pre-trained transformer language model such as *BERT* to create sentence embeddings for performing sentiment classification task. Various classification techniques and learning scenarios will be analysed, such as contrastive learning, transfer learning and semi-supervised learning. We will check the influence of various loss functions and data augmentation. Results will be analysed with respect to the amount of data used for training. Finally, we are going to look for modifications that might be able to increase our previous models performance.

1 Introduction

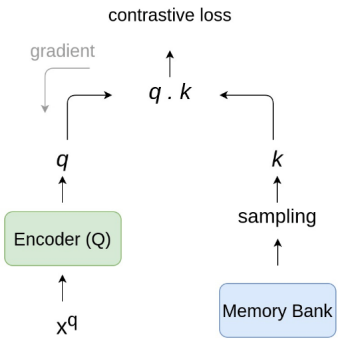
Although the size of the available text data is growing rapidly nowadays, the problems and costs related with obtaining labelled datasets still hinder a potential of current NLP models. The few-shot approach is a relatively new method in the machine learning, which aims to train a model on a limited number of labelled instances (supervised learning) and then use obtained model on much more extensive unlabelled data. The purpose of this approach is to utilize currently available huge amounts of unclassified data, which labelling would be time-consuming and costly. One popular solution for this is a contrastive learning approach, where a model learns representations (an image, text) by comparing positive and negative pairs of examples. The objective is to obtain such embeddings that similar examples are close to each other in the representation space and the unrelated ones are far from positive observations. In the context of the few-shot learning, this method enables to fine tune the model on the very limited data, because each instance is used multiple times in different training pairs. In this project we will also focus on the influence of data augmentation and unique loss functions that can be used in the sentiment classification task.

2 Related Works

2.1 Approaches and Learning Scenarios

2.1.1 Contrastive learning

Mapping input data to the target space where similar examples are close to each other and different ones are separated was firstly introduced in the image recognition tasks (Chopra, 2005). In the context of NLP, the first implementation was for unsupervised traning objective (Smith, 2005). Also the famous *Word2vec* model (Mikolov, 2013) applies the approach with the vector space where contextually similar words are located close to each other.



(b) Memory Bank

Figure 1: Standard memory-bank pipeline for contrastive learning (Jaiswal, 2020).

From the few-shot learning perspective, particularly interesting approach is applied in the

SimCSE model (Gao, 2021). Namely, in the unsupervised mode, a positive observation is used multiple times with a different dropout mask, which works as a very simple, yet highly effective data augmentation method. The rest of examples are treated as a negative sample.

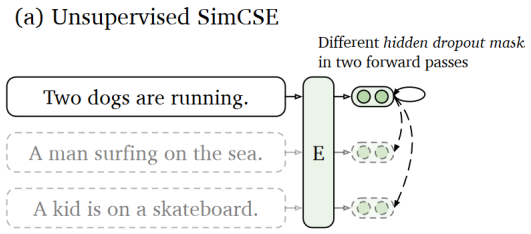


Figure 2: Dropout mask as a data augmentation (Gao, 2021).

Contrastive learning could be applied to increase the amount of training data by changing the classification task. Instead of predicting the label of a single observation, model estimates whether a pair of observation has the same label or not. This way number of observations for training with respect to the modified objective is the number of observation pairs for the original problem.

This approach can be used also for multi label classification (Gunel, 2020), where two observations are assumed to be always positive or negative to each other. However, contrastive learning can be enriched by additional classification component, e.g. *kNN* model, that takes into account an information from the neighbourhood of a particular instance (Wang, 2022). In such way, similarities between labels and correlations between them can be included in the final estimation.

2.1.2 Transfer Learning

Transfer Learning is a method of utilizing knowledge (such as data and possibly model associated with it) used to solve a problem and apply it to a new, similar task (Bozinovski, 1976). As an example deep learning model fine-tuned to recognise pictures of a certain type of animal might need little data to learn how to recognise different types of animals with high accuracy. This approach could improve the model in various ways: higher starting performance, higher pace of learning and higher final performance asymptote, as described in literature (Torrey, 2010); also in the domain of Natural Language Processing (Ruder, 2019).

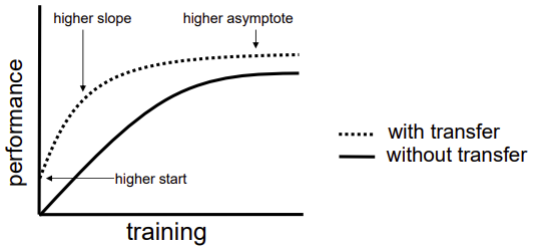


Figure 3: Possible improvements of using Transfer Learning (Torrey, 2010).

We assume that transfer learning might be particularly beneficial for text classification as this problem generalizes easily: for instance movie reviews data might be used for sentiment analysis in majority of other domains where positive/negative emotions need to be distinguished. Feature spaces can be also very similar if not the same, which is often hard to achieve with other machine learning challenges. Therefore, we might consider treating and analysing this approach separately or try to find data that is not as similar to the target problem.

2.1.3 Semi-supervised Learning

Semi-supervised learning is an approach involving a limited, small number of labeled observations and a large number of unlabeled ones (Ouali, 2006). This technique combines both supervised and unsupervised learning. Sufficient amount of unlabeled data allows model to learn the data structure in more comprehensive way (Van Engelen, 2020).

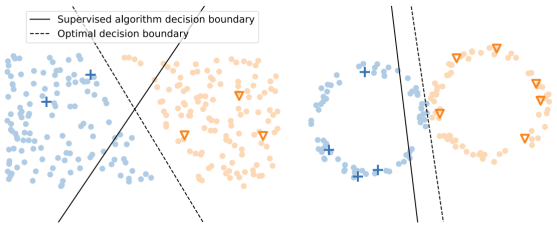


Figure 4: Examples of unlabeled data improving decision boundary (Van Engelen, 2020).

2.1.4 Loss Function

Loss Function choice for training and fine-tuning is vital for the whole learning process and influences the results obtained (Li, 2003). During

the project, we plan to analyse unique loss functions are test whether they can improve the methods described above, such as TripleEntropy (Sosnowski, 2021). As claimed by the authors, the loss function improves the results by about (0.02% - 2.29%). In case on small datasets, the gain is the largest (0.78% on average).

2.2 Data

We plan to use the datasets for the binary classification task of sentiment analysis.

2.2.1 Datasets

The first dataset will be the IMDb: Large Movie Review Dataset (Maas, 2011). It is a dataset for binary sentiment classification containing a set of 25,000 highly polar movie reviews for training, and 25,000 for testing. There is also an additional unlabeled data for unsupervised learning purposes. The additional training process will be mostly done with the unlabeled data. Another data source will be the Amazon Review Polarity dataset (Zhang, 2015). It also focuses on the binary classification. It contains 1,800,000 training samples and 200,000 testing samples in each polarity sentiment. Lastly we are going to consider the Yelp Review Polarity dataset (Zhang, 2015). The Yelp reviews dataset is obtained from the Yelp Dataset Challenge in 2015. This dataset contains 280,000 training samples and 19,000 test samples in each polarity.

3 Planned Solution

The proposed implementation will be split into two phases. In the first phase (Project 1), Bag-of-Words combined with contrastive learning and several loss functions will be evaluated. In the phase two (Project 2), created solution will be extended by the usage of BERT model embedding instead of Bag-of-Words approach. In case of unsatisfactory results, additional data augmentation will be also performed in both phases.

3.1 Bag-of-Words

For the chosen dataset, the bag-of-words model will be constructed by selecting a number of most frequent words from the training subset. This will be the first solution test to see how much performance we can expect from this traditional model. In case of the model not meeting our expectations, we are going to try the advanced sentence embedding methods.

3.2 BERT

BERT stands for Bidirectional Encoder Representations from Transformers. BERT is a method of pre-training language representations, meaning that we train a general-purpose ”language understanding” model on a large text corpus, and then use that model for downstream NLP tasks that we care about. BERT outperforms previous methods because it is the first unsupervised, deeply bidirectional system for pre-training NLP. For our task the vector representation obtained from BERT will allow us to have smaller and less complicated classification module. The features obtained with already pre-trained BERT model already take into account general language properties and make it easier to explore similarities of sentences in the sentiment analysis context.

3.3 Few-shot Learning

After obtaining vector representations of text observations, we are going to analyse and compare the described methods, while following the few-shot scenario. We will examine the performance with various amounts of labeled and unlabeled data and try to combine techniques when possible (for instance using a method with addition of data augmentation and particular loss function). After comprehensive analysis of State of The Art methods, we will try to find interesting and promising modifications and approaches that hopefully may compete with analysed procedures.

3.4 Data Augmentation

In addition to learning a model from representations and embeddings we are going to leverage them in using raw data without annotations. With that approach we will be able to continue training based on similarity of the data samples without annotations and samples from training part of the dataset. Additionally we will use linguistic database (e.g. WordNet) and synonym replacement approach to create more examples. This will effectively increase the size of our training annotated corpus data. While for the bag-of-words approach this method may not prove as effective as for advanced pre-trained embedding method such as BERT, we will examine the overall improvements and possibilities of this method in contrast to learning on the different sizes of annotated data.

3.5 Loss Functions

For the sentiment classification task we will try a range of loss functions including our own modifications in order to ensure that chosen data representation methods are utilized to their full potential. Firstly we are going to use the classic loss functions such as softmax. In addition we will use the TripleEntropy loss (Sosnowski, 2021) and other modifications if needed.

References

Wang, Yaqing and Yao, Quanming and Kwok, James and Ni, Lionel M. 2019. *Generalizing from a Few Examples: A Survey on Few-Shot Learning*. https://arxiv.org/abs/1904.05046

Chopra, S. and Hadsell, R. and LeCun, Y. 2005. *Learning a similarity metric discriminatively, with application to face verification*. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)

Smith, Noah A. and Eisner, Jason 2005. *Contrastive Estimation: Training Log-Linear Models on Unlabeled Data*. Association for Computational Linguistics, 354–362

Tomas Mikolov and Ilya Sutskever and Kai Chen and Greg Corrado and Jeffrey Dean 2013. *Distributed Representations of Words and Phrases and their Compositionality*. http://arxiv.org/abs/1310.4546

Ashish Jaiswal and Ashwin Ramesh Babu and Mohammad Zaki Zadeh and Debapriya Banerjee and Fillia Makedon 2020. *A Survey on Contrastive Self-supervised Learning*. https://arxiv.org/abs/2011.00362

Tianyu Gao and Xingcheng Yao and Danqi Chen 2021. *SimCSE: Simple Contrastive Learning of Sentence Embeddings*. https://arxiv.org/abs/2104.08821

Beliz Gunel and Jingfei Du and Alexis Conneau and Ves Stoyanov 2020. *Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning*. https://arxiv.org/abs/2011.01403

Wang, Ran and Dai, Xinyu et al. 2022. *Contrastive Learning-Enhanced Nearest Neighbor Mechanism for Multi-Label Text Classification*. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)

Bozinovski, Fulgosi 1976. *The influence of pattern similarity and transfer learning upon training of a base perceptron b2*.

Torrey, Lisa and Shavlik, Jude 2010. *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*.

Ruder, Sebastian and Peters, Matthew E. and Swayamdipta, Swabha and Wolf, Thoma 2019. *Transfer Learning in Natural Language Processing*.

Yassine Ouali and Céline Hudelot and Myriam Tami 2006. *An Overview of Deep Semi-Supervised Learning*.

Van Engelen, Jesper E and Hoos, Holger H 2006. *A survey on semi-supervised learning*.

Li, Fan and Yang, Yiming 2003. *A loss function analysis for classification methods in text categorization*.

Witold Sosnowski and Anna Wróblewska and Piotr Gawrysiak 2021. *Applying SoftTriple Loss for Supervised Language Model Fine Tuning*.

Maas, Andrew L. and Daly, Raymond E. and Pham, Peter T. and Huang, Dan and Ng, Andrew Y. and Potts, Christopher 2011. *Learning Word Vectors for Sentiment Analysis*, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, http://www.aclweb.org/anthology/P11-1015

X. Zhang and J. Zhao and Y. LeCun 2015. *Character-level Convolutional Networks for Text Classification*, https://doi.org/10.48550/arXiv.1509.01626