NLP project Proof of Concept Report
Few-shot Learning: Training Deep Learning Classifiers with Little Labeled Data - NaturAI

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 the second project for NLP Winter Course 2022 is to expand and improve the methods and algorithms investigated in the first project. Experiments with a new contrastive loss function and comparison to previously used method were conducted. Also replacement of voting with kNN method was verified. In order to further extend the performance of the models without using more significant amounts of data, we employed the data augmentation methods such as replacing words with synonyms.

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 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. Since the results obtained during the first phase project were promising, we decided to continue this topic. Below we present three modifications which aimed to improve the model performance.

2 New contrastive loss function

Previously, we have used the most basic contrastive loss having the form:

$$L(x_i, xj) = \mathbb{1}[y_i = y_j]|f(x_i) - f(x_j)|^2 + \mathbb{1}[y_i \neq y_j]max(0, m - |f(x_i) - f(x_j)|)^2 \quad (1)$$

Which gave relatively good and stable results. Nevertheless, it was decided to evaluate a variant of InfoNCE function adjusted to our Siamese architecture. The final formula used in the loss calculation is as follows:

$$L = -\log \frac{\exp(sim(x, x^+))/\tau}{\exp(sim(x, x^-))/\tau} \quad (2)$$

Where $sim(x, x^+)$ and $sim(x, x^-)$ are the sums of cosine similarities between all batch pairs within one group and pairs of observations from different classes respectively. A τ is a temperature hyper-parameter, in our experiments, it was set to 0.05. Both functions were applied to the same network and training/test sets.

Function	avg F1 score
Simple contrastive loss	77.4%
New contrastive loss	78.2%

Table 1: Contrastive voting F1-score for different loss functions (n: train=100, test: 9000)

The obtained F1-score values showed that the new contrastive function improved model performance. The metric increased by 0.8%. The same enhancement can be observed for accuracy.

Function	avg Accuracy
Simple contrastive loss	77.0%
New contrastive loss	77.8%

Table 2: Contrastive voting accuracy for different loss functions (n: train=100, test: 9000)

3 New voting approach

During tests with SimCSE embeddings (Gao, 2021) conducted within the first project we observed that voting realized with k Nearest Neighbors Approach achieved significantly higher results that calculating the mean similarity over all training samples. That motivated us to verify whether using such approach would improve the performance of the siamese networks from the previous project where only the mean similarity was taken into account. We conducted the experiment where results aggregated as mean where compared to outcomes of using the kNN approach. Accuracy and F1 scores are presented in Figure 1. As can be noticed results are very comparable, which means that in case of using the siamese networks the mean similarity is as valuable for prediction as only the few closest observations. All experiments were conducted with training set of size 100.

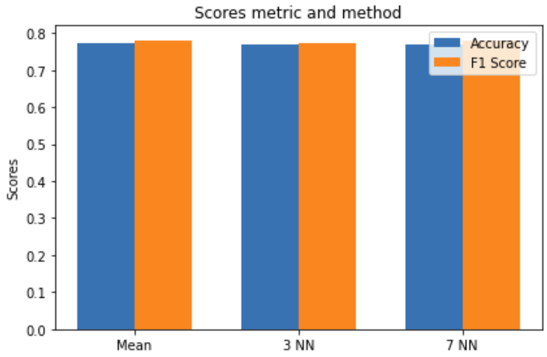

Figure 1: Voting approaches scores for Siamese networks on testing datasets

4 Data augmentation

The other approach we tested was data augmentation. This enables us to have bigger training set without additional annotations and is done automatically. We have tried 3 different augmentation methods: synonym replacement, context word embedding with replacement and back-translation. The tests were conducted on the Yelp Review Polarity Dataset. In the synonym replacement each review had a percentage of all words replaced with the synonyms from the wordnet base. In the context embedding method some words were replaced so that the meaning of the sentence and context are as close as possible to the original. This method was performed based on the BERT-base model embeddings. The last method was back-translation. The augmented sample was created by translating original review form English to German and back from German to English. This was done with the use of facebook wmt19 models from the Hugging Face library.

Augmentation method	Accuracy
none	82.6%
synonym replacement	83.4%
context embedding	81.9%
back-translation	82.5%

Table 3: RoBERTa-base Accuracy for different augmentation methods (n: train=100, test: 9000)

Augmentation method	F1-score
none	82.6%
synonym replacement	83.3%
context embedding	81.6%
back-translation	82.6%

Table 4: RoBERTa-base F1-score for different augmentation methods (n: train=100, test: 9000)

The initial tests were done on the RoBERTa-base model. The size of the training dataset was 100 samples. After each augmentation method the size was increased to 150 samples. In each of the replacement methods the replacement ratio was equal to $1/3$. The size of the test set was 9000 samples.

The comparison of accuracy for the synonym replacement method for different replacement ratios can be seen in the Figure 2.

The comparison of F1-score for the synonym replacement method for different replacement ratios can be seen in the Figure 3.

The performance of the model is higher when using data with 0 augmented words, but of the same train set size. However augmenting data effectively makes the train set larger so it can still be beneficial.

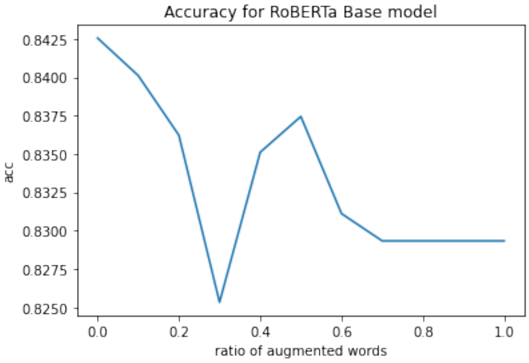

Figure 2: RoBERTa-base Accuracy for different synonym augmentation ratios

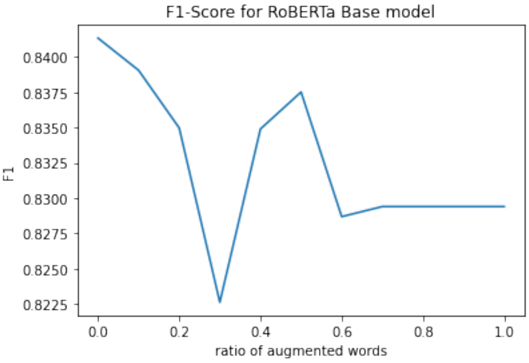

Figure 3: RoBERTa-base F1-score for different synonym augmentation ratios

References

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

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)

Ian Goodfellow et al. 2015. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*.

Khorram, Soheil et al. 2022. *Contrastive Siamese Network for Semi-supervised Speech Recognition*. Google Inc.

Tianyu Gao and Xingcheng Yao and Danqi Chen 2021. *SimCSE: Simple Contrastive Learning of Sentence Embeddings*.

Oord, Aaron van den and Li, Yazhe and Vinyals, Oriol 2018. *Representation Learning with Contrastive Predictive Coding*.

Feng, Steven Y., Gangal, Varun, Wei, Jason, Chandar, Sarath, Vosoughi, Soroush, Mitamura, Teruko, and Eduard Hovy. *A Survey of Data Augmentation Approaches for NLP* arXiv, (2021). https://doi.org/10.48550/arXiv.2105.03075.