Few-shot learning in NLP final presentation - PRO2

Dawid Przybyliński, Aleksander Podsiad, Piotr Sieńko

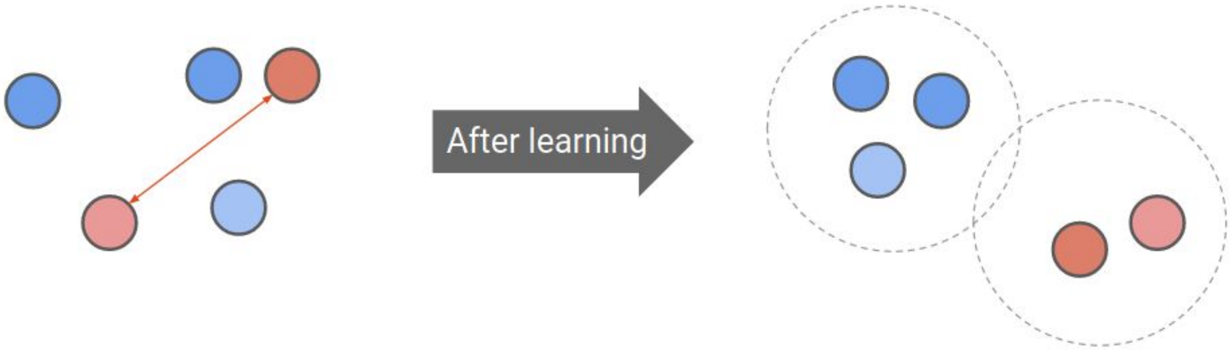
Few-shot learning

- Deep learning requires huge amount of data to train
- Labelling the data is costly and time consuming
- Efficient techniques for undersized samples are needed

Few-Shot Methods

Contrastive learning

- Having positive and negative examples, create such an embedding, that distance between instances from the same class is minimized and distance to different classes is maximized



Source: Lilian Weng, Jong Wook Kim, NeurIPS 2021

Phase I – experiments summary

Model	avg Accuracy	avg F1	n train
BERT Contrastive + voting	77.0%	77.4%	100
RoBERTa large + soft-triple loss	**83.8%**	**83.7%**	100
SimCSE embeddings + voting	74.5%	70.0%	100
Bag of Words + MLP	66.3%	71.6%	100

- Most of the tests were done with semi cross-validation having 2 or 3 folds
- Most testing datasets had 9000 samples (depending on model)
- Experiment were additionally carried out multiple times

Phase I – plans for next project

- Expanding on the idea of voting with contrastive loss function
- Usage of data augmentation
- Tests for more loss functions
- Transfer learning with new data sources
    - IMDb: Large Movie Review Dataset - 25,000 training, 25,000 test

    - Amazon Review Polarity Dataset - 1,800,000 training, 200,000 test
    - Yelp Review Polarity Dataset - 280,000 training, 19,000 test

Dataset – IMDb

- IMDb: Large Movie Review Dataset
- 25,000 training, 25,000 test
- Review examples:

  “If you like original gut wrenching laughter you will like this movie…” - positive

  “So im not a big fan of Boll's work but then again not many are…” - negative

Dataset – Yelp Reviews

- Yelp Review Polarity Dataset
- 280,000 training, 19,000 test
- Review examples:

  “Been going to Dr. Goldberg for over 10 years. I think I was one of his 1st patients when he started at MHMG. He's been great over the years…” - positive

  “Unfortunately, the frustration of being Dr. Goldberg's patient is a repeat of the experience I've had with so many other doctors…” - negative

Dataset – Amazon Reviews

- Amazon Review Polarity Dataset
- 1,800,000 training, 200,000 test
- Review examples:

“Stunning even for the non-gamer This soundtrack was beautiful! …” - positive

“Did not fit 2004 Kia Optima LX I purchased this for my 2004 Kia Optima LX…” - negative

Data augmentation

- using Python package NLPAUG
- replacing some percentage of the words in review with their synonyms from NLTK and WordNet
- replacing some percentage of words based on the embeddings from the BERT-base uncased model
- back-translation of the reviews - translating from English to German and from German to English
- text summarisation - summarising the text with the use of t5-small model

Data augmentation

Dataset	Augmentation	avg Accuracy	avg F1	n train
Yelp	None	**84.5%**	**84.4%**	100
Amazon		84.6%	84.6%	100
IMDB		75.6%	75.3%	100
Yelp	Copying data	81.8%	81.7%	200
Amazon		**85.9%**	**85.9%**	200
IMDB		76.9%	76.7%	200

model: RoBERTa-base with softriple loss

Data augmentation

Dataset	Augmentation	avg Accuracy	avg F1	n train
Yelp	Synonym replacement	84.3%	84.2%	200
Amazon		83.4%	83.3%	200
IMDB		74.9%	74.8%	200
Yelp	Embedding replacement	80.7%	80.3%	200
Amazon		84.3%	84.0%	200
IMDB		**77.1%**	**76.9%**	200

model: RoBERTa-base with softriple loss

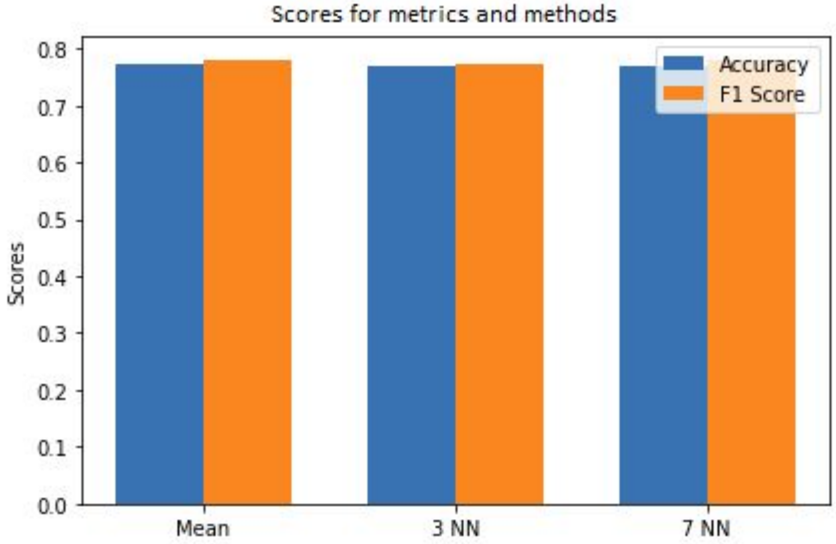
Data augmentation

Dataset	Augmentation	avg Accuracy	avg F1	n train
Yelp	Back-translation	80.3%	80.1%	200
Amazon		83.5%	83.3%	200
IMDB		73.8%	73.3%	200
Yelp	Text summarisation	84.2%	84.2%	200
Amazon		84.9%	84.8%	200
IMDB		76.8%	76.6%	200

model: RoBERTa-base with softriple loss

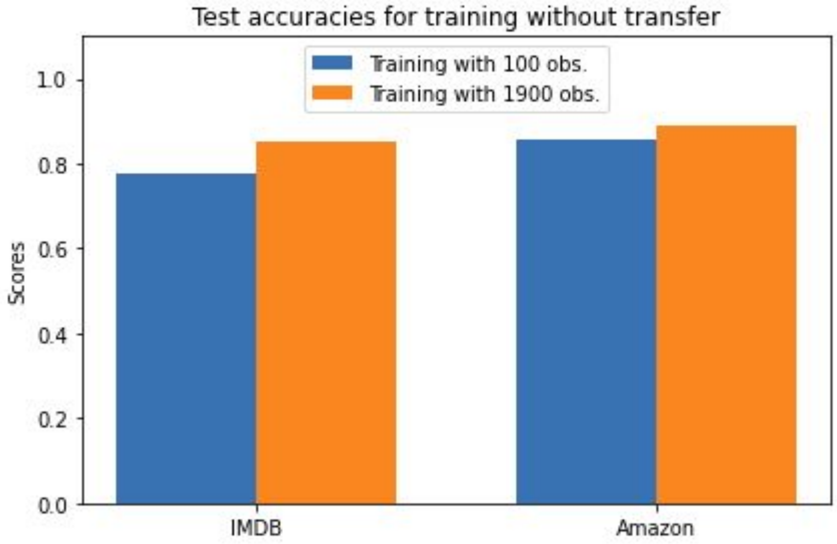
Voting with contrastive loss function

- While experimenting with contrastive learning embeddings, we noticed that while looking for most similar samples, voting with a few of the closest neighbors gives significantly higher results than calculating average similarity to classes
- We checked whether such modification could improve our Siamese Networks performance

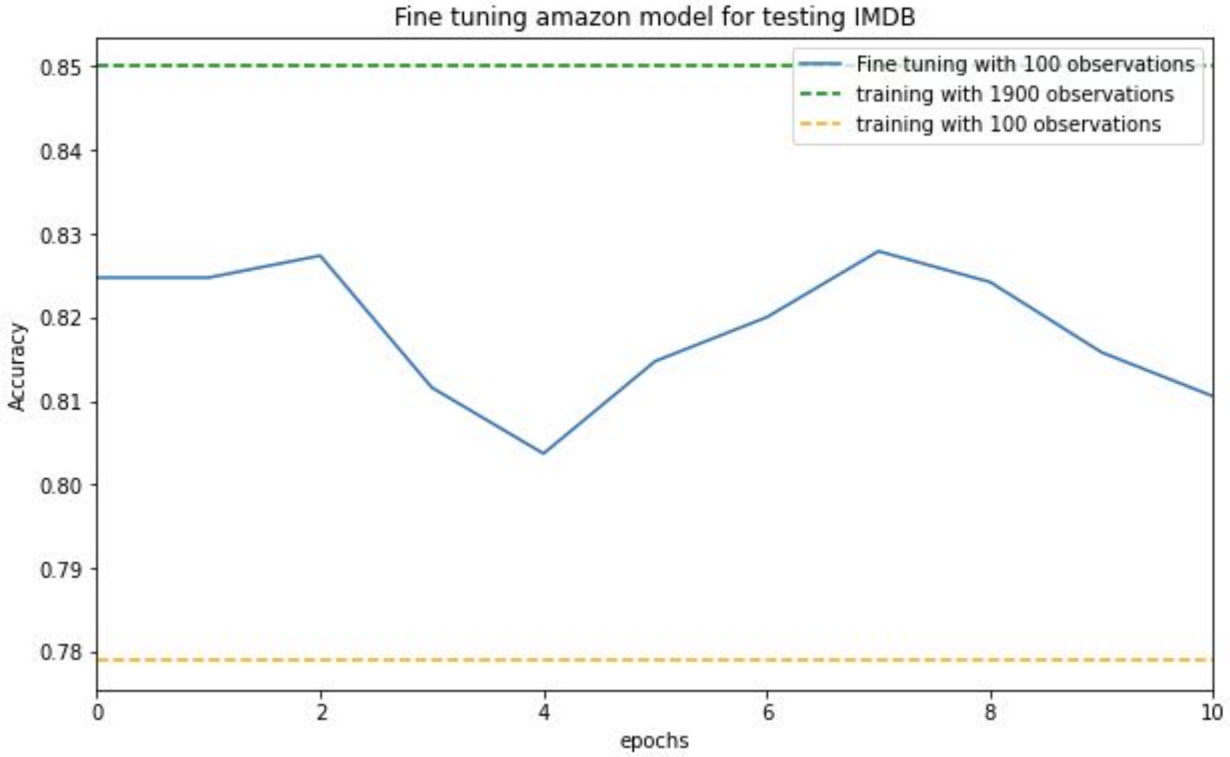


Transfer learning with new data sources

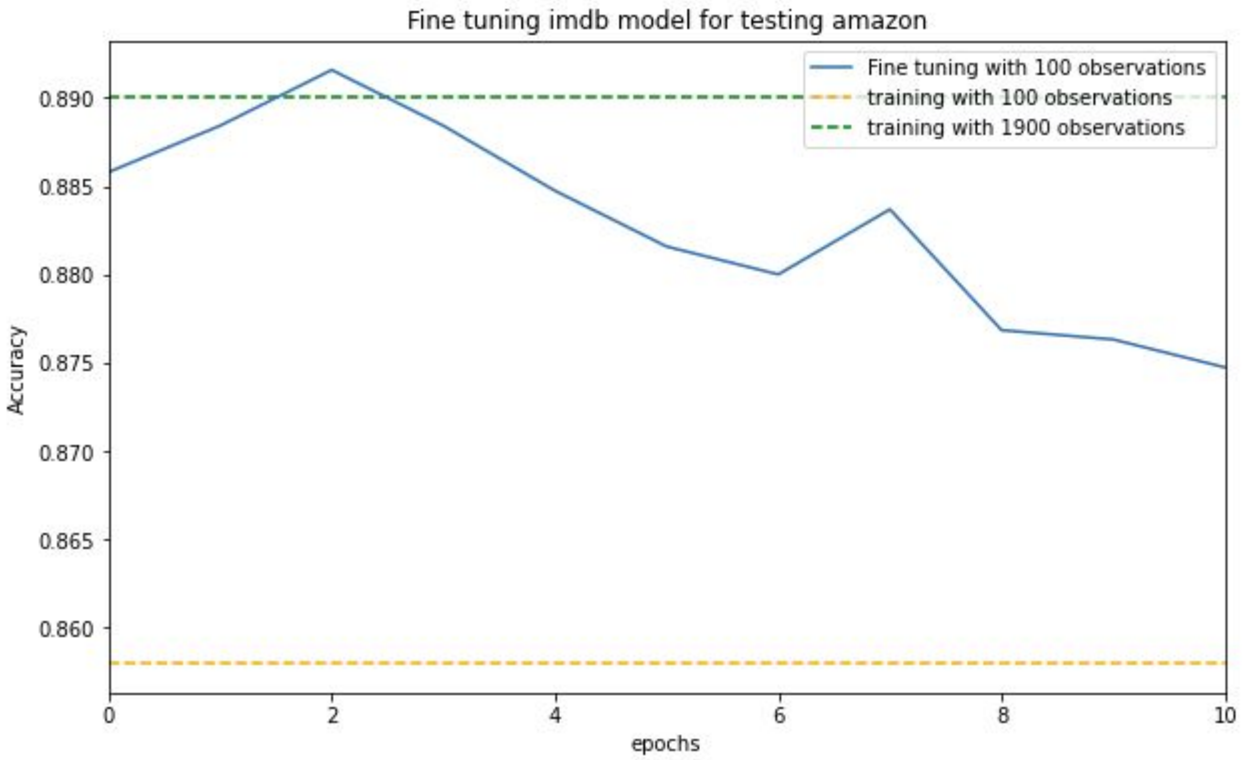
- Amazon Review Polarity Dataset: contains reviews about Amazon products
- To what extent learning sentiment from movie reviews helps in analysing sentiment for various products?



Transfer learning for predicting IMDB



Transfer learning for predicting Amazon



Tests for more loss functions

- Basic Contrastive Loss:

$$\mathcal{L}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boxed{\mathbb{1}[(y_i = y_j)] \| f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j) \|^2} + \boxed{\mathbb{1}[(y_i \neq y_j)] \max(0, m - \| f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j) \|)^2}$$

minimizes the embedding distance when they are from the same class

maximizes the embedding distance when they are from the same class

Source: R. Zhang et al, NAACL 2022

- Modified InfoNCE Loss:

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

sim(x,y): cosine similarity

τ: temperature hyper-parameter

Tests for more loss functions

Model	avg Accuracy	avg F1	n train
BERT Contrastive + voting	77.0%	77.4%	100
BERT Contrastive + voting + InfoNCE	77.8%	78.2%	100

- With the same architecture and samples, modified InfoNCE function achieved better results than simple contrastive loss

Splitting reviews into sentences

- Create more training examples by splitting a review into a set of sentences
- Keep sentences that have more than 50 characters
- From 50 to 464 training examples
- The test dataset was also split into a list of examples
- The final review sentiment estimated by an average sentiment of review's sentences (additional voting)

Model	avg Accuracy	avg F1	n train
BERT Contrastive + voting	77.0%	77.4%	100
BERT Contrastive + voting + split sentence	70.7%	70.5%	50

References

1. Generalizing from a Few Examples: A Survey on Few-Shot Learning (Wang, Yao, Kwok, Ni, 2020) https://arxiv.org/abs/1904.05046
2. Distributed Representations of Words and Phrases and their Compositionality (Mikolov, Sutskever, Chen, Corrado, Dean) http://arxiv.org/abs/1310.4546
3. X. Zhang and J. Zhao and Y. LeCun 2015. Character-level Convolutional Networks for Text Classification, https://doi.org/10.48550/arXiv.1509.0162
4. Lilian Weng, Jong Wook Kim, Self-Supervised Learning: Self-Prediction and Contrastive Learning, NeurIPS 2021, https://neurips.cc/media/neurips-2021/Slides/21895.pdf
5. Zhang et al. Contrastive Data and Learning for Natural Language Processing, NAACL 2022, https://contrastive-nlp-tutorial.github.io/files/contrastive_nlp_tutorial.pdf
6. A survey on semi-supervised learning (Van Engelen, Jesper E and Hoos, Holger H) https://doi.org/10.1007/s10994-019-05855-6
7. Realistic Evaluation of Deep Semi-Supervised Learning Algorithms (Oliver, Odena, Raffel, Ekin D. Cubuk & Ian J. Goodfellow) https://arxiv.org/abs/1804.09170
8. Chen et al. A Simple Framework for Contrastive Learning of Visual Representations, 2022
9. Applying SoftTriple Loss for Supervised Language Model Fine Tuning (Sosnowski, Wróblewska, Gawrysiak, 2021) https://arxiv.org/abs/2112.08462.
10. https://kevinmusgrave.github.io/pytorch-metric-learning/losses/#softtripleloss