Investigating ways to improve text classification performance when dealing with limited amounts of labeled data

Audun Ljone Henriksen and Morten Blørstad

INF368A - Project Report

Abstract

In real world text classification tasks high quality labeled data is costly to obtain. Models such as GAN-BERT and LAMBADA have been shown to improve sentence classification performance when dealing with limited amounts of labeled data. We investigate if the results can be extended to document classification problems. The GAN-BERT and LAMBADA methods are applied to the IMDB and medical text data sets using different subset sizes. The result of the models are compared to a baseline classifier based on BERT.

1 Introduction

Text classification model often requires lots of labeled data in order to perform well, and in real world tasks high quality labeled data is costly to obtain, both in time and monetary cost.

In the paper by D. Croce et al., 2019, they showed that the amount of labeled data can be drastically reduced and still obtain good performance in several sentence classification tasks, using a semi-supervised generative adversarial network method that extends BERT called GAN-BERT.[2]

The GAN-BERT model requires unlabeled data in order to work, and the paper by A. Anaby-Tavor et al., 2019 showed how text classifiers performance could be improved on a variety of sentence classification tasks with limited amounts of labeled data without the need of unlabeled data using a method called LAMBADA.[1] LAMBADA synthesizes new labeled data by fine-tuning a state-of-the-art language generator.

Both GAN-BERT and LAMBADA improved the performance on a variety of sentence classification tasks with limited amounts of labeled data. In our project, we are interested in comparing the two methods on the same data sets, but instead of applying them to sentence classification tasks we will compare the methods on document classification tasks. The performance of the methods will also be compared to a baseline classifier based on BERT.[1]

2 Related Work

GAN-BERT [2] shows that by utilizing unlabeled data in poor training conditions with little labeled data, performance can be significantly improved. This is done at no extra cost during test time as the generative part of the model is thrown away. With GAN-BERT they find that with less than 200 annotated examples, it is possible to achieve results that are comparable to those of a fully supervised setting, and they find that their model always improves over BERT.

LAMBADA [1] has shown that generating fake samples with language models to use in training data can improve performance of various text classification methods when labeled data is scarce. This is done by using a form of transfer learning where the language model is pre-trained on a large corpus, and then fined tuned on the relevant data set. With LAMBADA they achieve state of the art results over all their data sets and all classifiers, even beating methods using unlabeled data. Notably this is sentence classification.

3 Preliminaries

In this section we will present the three methods we use in this project, and look at the difference between sentence and document classification. BERT stands for Bi-directional Encoder Representations from Transformers and is a language model based on the transformer architecture. It outputs a vector representation for each word in the input sequence.

GAN-BERT, where the GAN part stands for Generative Adversarial Network, is a generative model that uses a discriminator and a generator in order to utilize unlabeled data to help with text classification with scarce labeled data.

LAMBADA stands for Language Model Based Data Augmentation, and is a data augmentation method

[1] The code for the project is available at https://github.com/ALjone/INF368-Final-Project

that utilizes the capabilities of GPT-2 in order to synthesize new samples and artificially increase the size of the training data.[3]

Sentence classification is the task of identifying a single sentence as belonging to a class.

Document classification is the task of identifying an entire document as belonging to a class.

A document consists of one or more sentences, and sentence classification is therefore considered a harder task than document classification.[1]

4 Experimental setup

4.1 Data

For the experiment the IMDB movie reviews[2] and the medical data set[3] are used.

Data set	Classes
IMDB	2
Medical text	5

Table 1: The data sets used in the project

The IMDB data set consists 50 000 highly polar movie reviews with two classes.

The medical text data set consists of 28 880 medical text describing the current conditions of a patient. A medical text belongs to one of five classes.

Each data set is modified to fit our experiment. The data sets are divided into a training set and a test set. The training set is divided into 4 subsets including 5, 10, 25 and 50 samples per class. A unlabeled data set consisting of 5000 samples is also constructed using the training set by removing the labels. The test set consists of 500 samples. All the sets are constructed to be balanced, i.e., each label is equally represented in each set.

4.2 Method

4.2.1 BERT

We use a baseline classifier based on BERT. BERT is a bi-directional transformer, using self-attention to weight the importance of words. We attach a linear-layer to BERT and train that layer in order to make it a classifier.

[2] http://ai.stanford.edu/~amaas/data/sentiment/

[3] https://www.kaggle.com/chaitanyakck/medical-text

4.2.2 GAN-BERT

The first model we test is GAN-BERT. This is a model that utilizes unlabeled data in order to increase performance when there is not enough labeled data.

As seen in Figure 1 GAN-BERT consists of BERT, a generator $\mathcal{G}$, and a discriminator $\mathcal{D}$. The generator $\mathcal{G}$ generates fake samples from a noise vector drawn from a normal distribution. Fake samples, as well as labeled and unlabeled data that is passed through BERT, are then passed to the discriminator $\mathcal{D}$. It will then try to classify it either into one of the k classes from the data set, or as k+1, meaning that it is a fake sample. For the fake samples and the unlabeled samples, the loss only considers at whether it is able to correctly classify whether it is real or not, and for labeled samples the loss also considers if it is correctly classified. After training, the generator $\mathcal{G}$ is discarded.

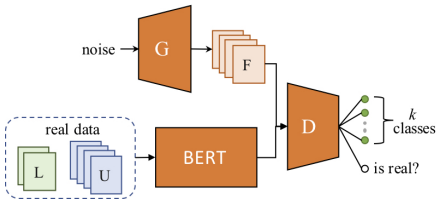


Figure 1: $\mathcal{G}$ generates fake samples give a random distribution. These are used as the input for the discriminator $\mathcal{D}$ along with labeled (L) and unlabeled (U) real data. [4]

4.2.3 LAMBADA

The second model we test is LAMBADA. LAMBADA creates weak-labeled synthesized data samples that are added to the original data set before training a classifier on it. The data is synthesized by GPT-2, after fine tuning it on the data set. The classifier used, both the final and the one for filtration, are based on the exact same architecture as in the BERT-classifier used as a baseline. When generating new sentences, we generated ten times as many sentences are we wanted to keep. Only the 10% best were kept during step 4. of Algorithm 1

[4] https://github.com/crux82/ganbert

Algorithm 1 LAMBADA

Input: Training data set D_{train}
Classification algorithm $\mathcal{A}$
Language model $\mathcal{G}$
Number to synthesize per class $N_1, ..., N_q$

1: Train a baseline classifier h from D_{train} using $\mathcal{A}$
2: Fine-tune $\mathcal{G}$ using D_{train} to obtain G_{tuned}
3: Synthesize a set of labeled sentences D^* using G_{tuned}
4: Filter D^* using classifier h to obtain $D_{synthesized}$

5 Experimental Evaluation

BERT, GAN-BERT and LAMBADA were trained on each of the training subsets and evaluated on the test set. Each model was trained using 5 epochs and a learning rate of $5e-5$. BERT and LAMBADA used a batch size of 4, whereas GAN-BERT used a batch size of 64. When generating examples with LAMBADA, we used an amount of synthesized data equal to the number of annotated examples, i.e., we doubled the amount of data in the training set.

Table 2 displays the results. Panel A shows that GAN-BERT performs better than the baseline classifier, BERT, only for the case 25 samples per label on the IMDB. LAMBADA performs better than BERT, for 5, 25 and 50 samples per label. For 10 samples per label, BERT performs better than both GAN-BERT and LAMBADA. GAN-BERT performs better than LAMBADA for 10 and 25 samples per label, whereas LAMBADA performs better than GAN-BERT for 5 and 50 samples per label.

Panel B shows GAN-BERT performs better than the baseline classifier, BERT, for the case 10 and 50 samples per label on the Medical text. LAMBADA performs better than BERT, for 25 and 50 samples per label. For 5 samples per label, BERT performs better than both GAN-BERT and LAMBADA. GAN-BERT performs better than LAMBADA for 10 samples per label, whereas LAMBADA performs better than GAN-BERT for 5, 25 and 50 samples per label.

The Figure 2 is a visualization of the results. Panel A shows that for IMDB, GAN-BERT has the greatest increase in performance when increasing samples per label from 5 to 25, but then it stagnates. From 10 to 50 samples per label the performance of both BERT and LAMBADA increase, but LAMBADA with a slightly greater rate.

Panel B shows that for Medical text the performance fluctuates for all models, especially for BERT and GAN-BERT. LAMBADA seems to be more stable and from 10 and on-wards the performance increases, thought at a decreasing rate.

	BERT	GAN-BERT	LAMBADA
Panel A: IMDB			
5	0.582	0.456	**0.588**
10	**0.580**	0.536	0.528
25	0.588	**0.779**	0.600
50	0.784	0.777	**0.816**
Panel B: Medical			
5	**0.600**	0.272	0.598
10	0.274	**0.460**	0.256
25	0.562	0.375	**0.586**
50	0.512	0.549	**0.620**

Table 2: Performance comparison
The table compares the accuracy of BERT, GAN-BERT and LAMBADA on the subset 5, 10, 25, 50 samples per class. Panel A shows the accuracy comparison on the IMDB data set. Panel B shows the performance comparison on the Medical text data set.

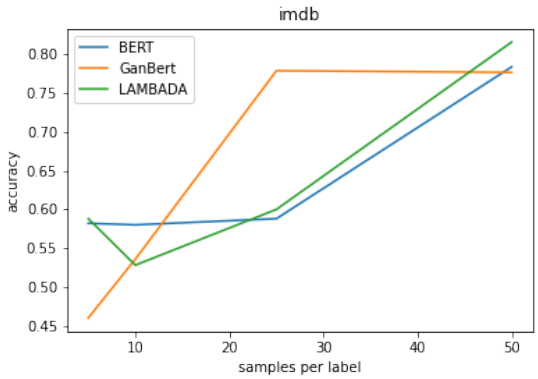
Our results for document classification does not show the same clear results as reported in the GAN-BERT and LAMBADA for sentence classification.

Interestingly, the times GAN-BERT does outperform both LAMBADA and BERT, it has a large performance gain over both. However, in most cases GAN-BERT fails to improve on BERT, and even does worse in most cases.

The results for LAMBADA seem promising for 25 and 50 samples per label and outperforms BERT for both data sets. However, we see some very volatile results for BERT and LAMBADA on the medical text data set, which makes conclusions hard to draw. Notably the drop in the accuracy of BERT from 25 samples per class to 50 samples per class for the medical text data set is worrying. Nonetheless LAMBADA outperforms BERT in almost every case, thought the improvements in most cases are small.

In both the GAN-BERT paper and the LAMBADA paper they showed that the models greatly outperformed BERT for low number of samples per label, but when increasing the number of labeled data, BERT approximates the performance of GAN-BERT

Panel A : IMDB



Panel B : Medical text

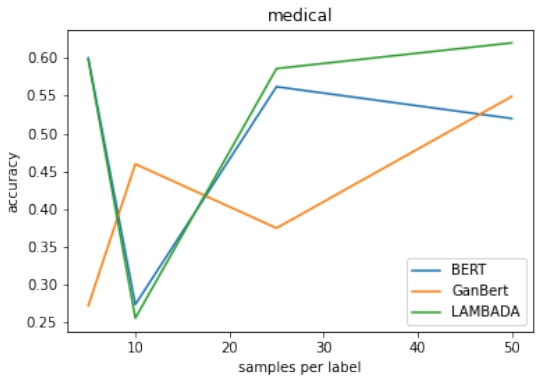


Figure 2: Performance comparison

The figure plots the change in accuracy for `BERT`, `GAN-BERT` and `LAMBADA` when increasing the number of sample per label. Panel A shows the change in performance on the IMDB. Panel B shows the change in performance on the Medical text.

and `LAMBADA` [1, 2].[5] In both papers they primarily used data sets with a large number of classes, which means that the total amount of labeled data is larger compared to our data set as we use data sets with a small number of classes.[6] This could indicate there might be a minimum amount of samples required for the methods to be effective. They also mentioned that classification tasks with a large number of classes are more difficult than classification tasks with smaller number of classes and sentence classification is harder than document classification. A possible explanation for our results could be these methods aren't suitable for these data sets, as they might be too easy. Easy classification tasks might have less potential performance gain, as less data is needed to acquire good results, and the effect of synthesized data or unlabeled data might therefore be less significant.

6 Conclusion

In this project we extended the experiments from the `GAN-BERT` and `LAMBADA` papers. We applied these methods to two document classification tasks. We were unable to reproduce the results for both methods consistently. `LAMBADA` showed promising result for sample size of 25 and 50 per label. However it was too inconsistent for sample size of 5 and 10 per label to draw a clear conclusion. `GAN-BERT` performed better than BERT for 25 labels on the IMDB dataset and 10 and 50 on medical text, but the inconsistency of its performance makes it hard to draw conclusions. This might indicate that the methods we investigated are more suitable for sentence classification with a high number of classes than they are for document classification with a low number of classes.

Further extension of this project would be to investigate document classification task with larger number of classes. Both `GAN-BERT` and `LAMBADA` have large number of hyper-parameters. A hyper-parameter search for these model could improve our results. Furthermore, investigating on more data set with various characteristics would help finding where the methods are applicable or not.

[5] Low number of samples per label meaning less than 10% labeled data and 5 samples per label for `GAN-BERT` and `LAMBADA`, respectively.

[6] The data sets we used have 5 or less classes

References

[1] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, and N. Zwerdling. Not enough data? deep learning to the rescue! *CoRR*, abs/1911.03118, 2019.

[2] D. Croce, G. Castellucci, and R. Basili. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119, Online, July 2020. Association for Computational Linguistics.

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2018.