# CMSC742 Final Project Report: Avoiding Negative Transfer Using Adversarial Network on NLP

**Anonymous Authors**[1]

## Abstract

Transfer learning has been widely applied to many tasks when the target data is difficult to obtain. However, when the source data and the target data are less related, the negative transfer can happen, where the source data may be useless or even harmful to the performance of the network. Although the negative transfer has gained more attention recently, it is less studied on complex tasks in natural language processing (NLP). In this paper, we explore the negative transfer in NLP with Question Classification (QC) data and analyze the impact of the ratio of labeled data. The result shows that the lower the ratio of labeled data, the severer the negative transfer. Furthermore, to avoid the negative transfer in the complex NLP tasks, we propose to adopt the adversarial networks for fine-tuning and add a gate for reducing the weight of unlabeled data.

## 1. Introduction

When the target data is scarce or expensive to obtain, transferring knowledge from a related bountiful source or domain can help a model to learn the target task. Transfer learning has become maturer and is utilized in several fields, turning out to be powerful. However, it may lead to the so-called negative transfer, i.e., the knowledge from the previous task may hinder learning of the current task. So far, there have been limited works that discuss this topic, and the existing approaches are conducted on rather simple settings and tasks. For more practical applications, analyzing and solving negative transfer on complex settings and tasks remains to be done.

Transfer Learning has been widely used in Natural Language Processing (NLP). Being a powerful tool, transfer learning has outperformed the standard supervised learning in many cases, as it takes benefit from the pre-learned knowledge. However, negative transfer can happen when the source and target domain are not closely related, which hurts the performance and needs to be avoided. Some researchers resort to carefully selecting the data for transfer learning, but the effectiveness of data selection methods are task dependent and cannot be generalized to other tasks efficiently. Other researchers tried to avoid negative transfer with the application of adversarial networks, but they only investigate easy tasks, which may not be applicable to complex tasks in NLP.

Unlike many previous papers on negative transfer analysis that worked on simple tasks in computer vision, we focus on negative transfer of complex tasks in NLP and come up with a method to avoid it with the application of adversarial networks. First, we reviewed the definition of negative transfer. Then, we analyze transfer learning in NLP with the BERT pretrained model. Next, with the application of adversarial network, we try to alleviate the negative transfer in NLP by utilizing generative adversarial network (GAN) to fine tune the pretrained BERT model. Finally, borrowing the idea of GATE from XXX, we try to add the gate to the loss function in GAN to further avoid negative transfer.

We explore the negative transfer in NLP with the Question Classification (QC) data on the UIUC dataset. Furthermore, we also investigate the potential factors of negative transfer in our experiments. By adjusting the ratio of labeled data in the whole dataset, we find that the size of unlabeled data in the dataset affects the severity of negative transfer.

In the rest of the paper, we first briefly summarize the recent works on negative transfer in §2, and then detailed our method design in §3. We present the experiments and results in §4 and conclude in §5.

## 2. Related Work

In this section, we briefly introduce several existing works on negative transfer.

### 2.1. Supervised Local Weight

In many real-world applications, there could could be multiple relevant domains where knowledge can be transferred. Hence Multiple Source Transfer Learning was introduced and attracted a lot of attention. Multiple Source Transfer Learning assumed that: 1. these domains where the knowledge comes from are relevant to target domain, and

2. There is a balanced distribution of classes. Yet, in the extreme cases where both pre-requisites fail, negative transfer would occur. Hence Liang Ge et al. proposed a two-phase framework to ensure the success of transfer learning when there exists irrelevant domain and imbalanced distribution of classes. In the first phase, a Supervised Local Weight scheme is employed to assign a proper weight to each source domain's classifier. In this step, they used spectral clustering algorithm to partition the target data into clusters and label propagation to approximate groundtruth of the target labels. In the second phase will take the weight from the first phase and the target training data to learn a classifier. The significance of this classifier is that it shall ensure the performance of the proposed method will be no worse than the prediction using target labeled data alone. Their results demonstrated that the proposed method has better performance than existing MSTL approaches.

### 2.2. Batch Spectral Shrinkage

In the context of deep learning, transfer learning faces two issues that could hamper the hamper the generalization performance of the models. One is catastrophic forgetting, which means that the models could lose previous learnt knowledge during finetuning, leads to overfitting; the other is negative transfer. The authors put their focus on spectral components. Spectral components in high layers with small singular values are not transferred, and Xinyang Chen et al. noticed that these spectral components tend to be ignored by models during finetuning phase when there are sufficient training data. Inspired by this phenomenon, Chen et al. proposed Batch Spectral Shrinkage to suppress spectral components with small singular values that could impede finetuning and induce negative transfer. The proposed architecture is end-to-end trainable with differentiable single value decomposition. And the result shows BSS would yield significant performance gains.

### 2.3. Loss-Balanced Task Weighting

In multi-task settings, there are two scenarios where negative transfer could occur. First, when all tasks are irrelevant to each other. Second, one group of related task dominates the training process. The authors introduce Loss-Balanced Task Weighting, which combined Reinforced Multi-Task Learning and GrandNorm, to reduce negative transfer. The result shows that Loss-Balanced Task Weighting have successfully reduced negative transfer on the PubChem BioAssay chemistry dataset. However, the proposed version of Loss-Balanced Task Weighting uses uniform task weights when making prediction, which limited their application.

## 3. System Design

In this section, we first introduce the concept and the definition (we use in this work) of the negative transfer. Second, we introduce the concept of the current famous model: Bidirectional Encoder Representations from Transformers (BERT) since we aim to investigate the negative transfer in NLP. Next, we introduce the adversarial network based on BERT as our framework for negative transfer investigation and finally we address our method/idea for mitigating the negative transfer based on BERT with the adversarial network.

### 3.1. Negative transfer

Intuitively, negatively transfer occurs when the knowledge contained in the source data is irrelevant to the data in the target domain. This could be detrimental to the model's performance during fine-tuning. In this work, we are adopting definition of negative transfer from Wang et al.'s work. In particular, we are using three premises from their work: 1. Negative transfer should be algorithm specific. That is, negative transfer is meaningful when it is based on comparison between same algorithms, one with source-domain data and the other with target domain data. 2. Negative transfer is induced by the divergence between the joint distributions. In other words, negative transfer would occur when the distribution of the source domain shares no similarity with the distribution of the target domain. In an extreme case where the distribution of the source domain is uniform, i.e., data from the source domain contains no useful knowledge, the algorithm should perform better using target data only. 3. Negative transfer largely depends on the size of the labeled target data. These three premises will be the guideline of our work and direct our experiments. Though the negative transfer has been investigated in several works, there is not a classic definition yet. We follow the definition in (Wang et al., 2019) to characterize the negative transfer as below: the negative transfer gap is

$$R_{P_T}(A(\mathcal{S},\mathcal{T})) - R_{P_T}(A(\emptyset,\mathcal{T})),$$

where $R_{P_T}(A(\mathcal{S},\mathcal{T}))$ is the expected risk of using source data and target data as inputs and $R_{P_T}(A(\emptyset,\mathcal{T}))$ is the expected risk of using only target data for training and $\mathcal{T}$ represents the task. Thus, the negative transfer occurs if the negative transfer gap is positive.

### 3.2. Bidirectional Encoder Representations from Transformers (BERT)

BERT is one of most commonly used models in NLP. It is one of the variants of Transformer, and it is pretrained on large unlabeled dataset. One of the advantages of BERT and also the reason we consider using it is that it not pretrained to fit one particular task, but to extract features about language,

which allows the pretrained instances of BERT to handle various NLP tasks after finetuning. Though it might be a better approach for our research to train a BERT from scratch, we do not have enough computation resource to handle such task. The instance we are going to use in this work is pretrained by Google.

### 3.3. Adversarial Network

Generative adversarial network (GAN) was designed by Ian Goodfellow and his group members and has gained considerable attention on many topics. The core idea of the adversarial network is based on the "indirect" tranining through the so-called discriminator and the generator aims to fool the discriminator. The method has been widely applied to many fields for unsupervised learning, semi-supervise learning, supervised learning, and reinforcement learning.

In transfer learning, we typically are unable to obtain bountiful labeled target data while getting unlabeled data can be obtained more easily. Therefore, we focus on incorporating adversarial network into BERT for NLP to do the transfer learning with few labeled target data and bountiful unlabeled target data (semi-supervised learning).

Though the overhead of getting labeled data can be largely reduced with semi-supervised learning, the negative transfer may occur due to the divergence between the source and target data. (Wang et al., 2019) proposed to reweight each source sample in some proper manner instead of assuming equal weight among the source samples to avoid negative transfer on adversarial network. The authors exploit a GAN discriminator to perform the density ratio estimation and concludes at any point, the optimal discriminator is given by

$$D(x, y) = \frac{P_T(x, y)}{P_T(x, y) + P_S(x, y)}, \tag{1}$$

where $P_T(x, y)$, $P_T(x, y)$ are the joint distribution in the source and the target domain, where $X$ is the input random variable, and $Y$ is the output. By incorporating the density ratio in the loss function, the authors show that the negative transfer is indeed reduced. Since the density ratio acts as a gating function, the authors name the method as discriminator gate.

### 3.4. GAN-BERT

In many real scenarios, obtaining high quality annotated data is expensive and time consuming; in contrast, unlabeled examples characterizing the target task can be, in general, easily collected, whose goal is consistent with transfer learning. the authors (Croce et al., 2020) proposed to combine semi-supervised generative adversarial learning (GAN) and bert that extends the fine-tuning of BERT-like architectures with unlabeled data in a generative adversarial setting. Using the task-specific layer and fine-tune the entire architecture with generator and discriminator layers on annotated data to realize sentence labeling.

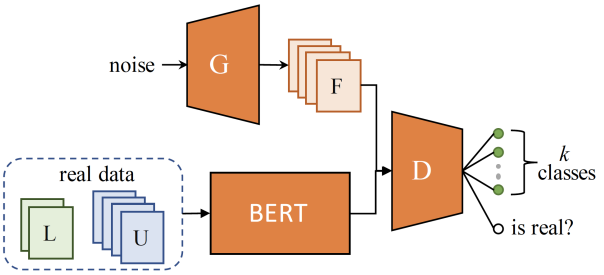


*Figure 1.* Network Structure

As in the figure 1, generator takes in noise vectors and generate adversarial examples to the feature extractor, which are approximating the statistic of real data as much as possible. On the other side, real data is pre-trained by bert and both of them are inputs of discriminator in order to be classified into k classes or fake. We denote loss function of discriminator as $L_D = L_{D_{sup.}} + L_{D_{unsup.}}$ where

$$L_{D_{sup}} = -E_{x,y\sim p_d} log[p_m(\hat{y} = y|x, y \in (1, ..., k))]$$

measures the error in assigning the wrong class to a real example among the original $k$ categories.

$$L_{D_{unsup}} = -E_{x_d} log[1 - p_m(\hat{y} = y|x, y = k + 1] \tag{2}$$
$$- E_{x\sim G} log[p_m(\hat{y} = y|x, y = k + 1)] \tag{3}$$

measures the error in incorrectly recognizing a real unlabeled example as fake and not recognizing a fake example.

The loss function of generator is defined as $L_G = L_{G_{featuremapping}} + L_{D_{unsup.}}$ with $f(x)$ the activation on an intermediate layer of discriminator,

$$L_{G_{featuremapping}} = ||E_{x\sim p_d} f(x) - E_{x\sim G} f(x)||_2^2$$

$$L_{G_{unsup}} = -E_{x\sim G} log[1 - p_m(\hat{y} = y|x, y = k + 1)]$$

considers the error induced by fake examples correctly identified by discriminator.

### 3.5. GAN-BERT With Gate

With GAN-BERT, we aim to avoid the negative transfer by borrowing the idea of gating function proposed by (Wang et al., 2019). By reweight the source data and the target data, we expect to reduce or avoid the negative transfer phenomenon. However, due to limited time, we are still not able to figure out how to modify the loss function and add the gating function into the loss function.

### 3.6. Overall Idea of The Project

We focus on the powerful and hit transfer learning model, BERT, on NLP. Although BERT has been widely applied and prove empirically to be useful and powerful. The drawback of BERT is that we have to gain labeled data as the input for the fine-tuning procedure. However, often times, we are not able to obtain enough or many labeled target data to fine-tune the model on BERT. GAN-BERT (Croce et al., 2020) was proposed to take unlabelled data (much more than the labeled data) as well as the input of the fine-tuning on BERT, which largely reduces the need for labeled data as the input. Thus, we use GAN-BERT-based model to do the second stage transfer learning by adding the target data and source data as the input, With GAN-BERT, we are able to take unlabeled targe data as the input to do the semi-supervised learning. However, since the labeled target data is scarse, and it is very likely to be biased by the source data if the source data is divergent from the target data, causing the negative transfer. We borrowed the idea of gating function (Wang et al., 2019) to reweight the target and source dataset to avoid the negative transfer, expecting to reduce or avoid the negative transfer in transfer learning.

## 4. Experiment and Result

In this section, we introduce the experiment setups and analyze the result of our experiments.

### 4.1. Experiment setup

In an ideal scenario, to measure the negative transfer, we would need to compare the performance of two instance of one same model: one pretrained with source-domain data then finetune with target domain data; the other trained with target domain data only. However, due to limitation of the amount of data we have and the purpose of our experiment, we are permuting the source and target domain dataset. Our dataset is based on Question Classification (QC) on the UIUC dataset (Li and Roth, 2006). The questions that are related to history are selected to be target data and others are used as source data. The target data is labeled as HIST (history), while the sub-categories are "volume size", "weight", "human", "location", "number", "entity", and "description".

For GAN-BERT model, we use dataset that contains both labeled and unlabeled data as training set. As we want to investigate the relation between the percentage of labeled data and the severity of negative transfer, we alter the amount of labeled data in the training set in each of the four scenarios while keeping the size of training set constant. We represent this change with L%, which is the ratio of labeled data in the whole training dataset. To set up the comparison, with each of the 4 different L%, we have 2 sub-scenarios: one uses mixture of source data and target data, one uses purely target data. For the BERT model, we use the same labeled data as those in the training set of GAN-BERT in each of the four scenarios and their sub-scenarios.

### 4.2. Result

The result of our experiment is shown in Table 1. From Table 1, we can find that the testing accuracy of the network which is trained by target only data is higher than that of the network which is trained by both source and target data. Therefore, we can observe the negative transfer in NLP when the source data and the target data are less related. Furthermore, when the L% increases and is greater than 15%, the difference between the mixed accuracy and the target-only accuracy decreases. This illustrates that the severity of negative transfer decreases when the labeled data ratio increases. Thus, decreases the unlabeled data ratio can be helpful to avoid the negative transfer.

*Table 1.* Testing Accuracy

| L% | MIXED ACCURACY | TARGET-ONLY ACCURACY |
|---|---|---|
| 25% | 0.74 | 0.96 |
| 20% | 0.64 | 0.95 |
| 15% | 0.62 | 0.92 |
| 10% | 0.31 | 0.45 |

## 5. Conclusion and Future Work

We applied generative adversarial network (GAN-BERT) to sentence classification task in natural language processing to investigate the negative transfer phenomena. From the experiment, we indeed observe negative transfer effects when the unlabeled dataset is incorporated. We also attempt to incorporate the gating function to reweight the source and target data to further avoid the negative transfer, but due to limited-time, we are not yet unable to figure out and fully implement it. This can be the future work to see if the negative transfer can be mitigated.

## References


Arase, Y. and Tsujii, J. Transfer fine-tuning: A BERT case study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5393–5404, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1542. URL https://aclanthology.org/D19-1542.

Chen, X., Wang, S., Fu, B., Long, M., and Wang, J. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. Advances in Neural Information Processing Systems 32, 2019.

Croce, D., Castellucci, G., and Basili, R. 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*, pp. 2114–2119, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.191. URL https://aclanthology.org/2020.acl-main.191.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Ganin, and Yaroslav, E. U., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*.

Ge, L., Gao, J., Ngo, H. Q., Li, K., and Zhang, A. On handling negative transfer and imbalanced distributions in multiple source transfer learning. *Stat. Anal. Data Min.*, 7:254–271, 2014.

Liu, S., Liang, Y., and Gitter, A. Loss-balanced task weighting to reduce negative transfer in multi-task learning. pp. 9977–9978. Proceedings of the AAAI Conference on Artificial Intelligence, 2019. doi: 10.1609/aaai.v33i01.33019977.

Meftah, S., Semmar, N., Tamaazousti, Y., Essafi, H., and Sadat, F. On the hidden negative transfer in sequential transfer learning for domain adaptation from news to tweets. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pp. 140–145, Kyiv, Ukraine, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.adaptnlp-1.14.

Pruksachatkun, Y., Phang, J., Liu, H., Htut, P. M., Zhang, X., Pang, R. Y., Vania, C., Kann, K., and Bowman, S. R. Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work? *ArXiv*, abs/2005.00628, 2020.

Ruder, S. and Plank, B. Learning to select data for transfer learning with Bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 372–382, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1038. URL https://aclanthology.org/D17-1038.

Wang, Z., Dai, Z., P´oczos, B., and Carbonell, J. Characterizing and avoiding negative transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019. IEEE.

Weiss, K. R., Khoshgoftaar, T. M., and Wang, D. A survey of transfer learning. *Journal of Big Data*, 3:1–40, 2016.

Yang, Z., Hu, Z., Dyer, C., Xing, E. P., and Berg-Kirkpatrick, T. Unsupervised text style transfer using language models as discriminators. In *NeurIPS*, 2018.

Zhang, W. and Lingfei Deng, Lei Zhang, D. W. Overcoming negative transfer: A survey. *arXiv*, 2020.