Reversing Gradients in Adversarial Domain Adaptation for Question Deduplication and Textual Entailment Tasks

Anusha Kamath
Carnegie Mellon University
Pittsburgh, PA
akamath1@andrew.cmu.edu

Sparsh Gupta
University of California San Diego
San Diego, CA
spg005@ucsd.edu

Vitor Carvalho
Intuit AI
San Diego, CA
vitor_carvalho@intuit.com

Abstract

Adversarial domain adaptation has been recently introduced as an effective technique for textual matching tasks, such as question deduplication (Shah et al., 2018). Here we investigate the use of gradient reversal on adversarial domain adaptation to explicitly learn both shared and unshared (domain specific) representations between two textual domains. In doing so, gradient reversal learns features that explicitly compensate for domain mismatch, while still distilling domain specific knowledge that can improve target domain accuracy. We evaluate reversing gradients for adversarial adaptation on multiple domains, and demonstrate that it significantly outperforms other methods on question deduplication as well as on recognizing textual entailment (RTE) tasks, achieving up to 7% absolute boost in base model accuracy on some datasets.

1 Introduction

Domain adaptation is a flexible machine learning approach that allows the transfer of category independent information between domains. Through domain adaptation we can leverage source task representations to bring the source and target distributions closer in a learned joint feature space. In this paper we are focused only on semi-supervised domain adaptation — when knowledge from a large labeled dataset in a source domain can be somewhat transferred to help improve the same task on a target domain, which typically has a significantly smaller number of labels. In particular, this paper focuses on domain adaptation for the detection of question duplicates in community question answering forums (Shah et al., 2018; Hoogeveen et al., 2015), as well as for RTE tasks (Dagan et al., 2005; Zhao et al., 2017).

Generally speaking, the effectiveness of domain adaptation depends essentially on two factors: the similarity between source and target domains, and representation strategy to transfer the source domain knowledge. Long *et al.* showed transferring features across domains becomes increasingly difficult as domain discrepancy increases (Long et al., 2017), since the features learned by models gradually transition from general to highly domain specific as training progresses. Recent domain adaptation strategies attempt to counter this issue by making certain features invariant across source and target domains using distribution matching (Cao et al., 2018) or minimizing distance metrics between the representations (Sohn et al., 2019).

The idea of generating domain invariant features was further enhanced by the use of adversarial learning methods. Recent work has advocated for tuning networks using a loss functions that reduce the mismatch between source and target data distributions (Sankaranarayanan et al., 2018; Tzeng et al., 2017). Others have proposed a domain discriminator that maximizes the domain classification loss between source and target domains (Cohen et al., 2018; Shah et al., 2018). One particular limitation of these approaches is that they are restricted to using only the shared domain invariant features and hence can't benefit from target domain specific information. Small amounts of labeled target domain data could in principle be used to fine-tune learned shared representations and improve the target task, however this could also lead to overfitting (Sener et al., 2016).

To address this issue, Qiu *et al.* used both shared domain invariant and domain specific features: while the shared features are learned by maximizing domain discriminator loss, the domain specific features are learned by jointly minimizing the task loss and the domain classification loss by domain specific discriminators (Qiu et al., 2018). Similar ideas were put forth by

Peng *et al* for cross-domain sentiment classification where they demonstrate the effectiveness of using both domain specific and domain invariant features (Peng et al., 2018). Moreover, Bousmalis *et al* have made similar observations in domain adaptation for image classification and related vision tasks (Bousmalis et al., 2016). All these studies follow similar approach of learning shared feature space by maximizing domain classification loss.

In contrast, our work here enhances the ideas from from Qiu *et al.* by utilizing a Gradient Reversal Layer (GRL) (Ganin and Lempitsky, 2015) to train the domain discriminator in a minimax game manner, and show that it results in significantly better transfer performance to multiple target domains. The use of gradient reversal layer is further advocated by works of Elazar *et al* (Elazar and Goldberg, 2018) and Fu *et al* (Fu et al., 2017) for removal of demographic attributes from text, and relation extraction from text, respectively. To the best of our knowledge, the use of Gradient Reversal in textual matching tasks, such as question deduplication and RTE, is novel and may trigger further applications of this approach in other language tasks.

To summarize our contributions, **(1)** we propose a novel approach for adversarial domain adaptation that uses gradient reversal layers to discover shared representations between source and target domains on textual matching tasks, and elegantly combines domain specific and domain invariant shared features. **(2)** We apply it to question deduplication tasks and empirically confirm that it outperforms all other strong baselines and feature sets on five different domains, with absolute accuracy gains of up to 4.5%. **(3)** We further apply the same approach to two different textual entailment domains, where it again outperforms other baselines by as much as 7% absolute accuracy points.

2 Approaches

2.1 Base Model:BiMPM

Wang et al. (Wang et al., 2017) proposed the Bilateral Multi-Perspective Matching model for many language tasks, including question duplicate detection and RTE. This model takes in the two candidate sentences as inputs to a Bi-LSTM layer that generates hidden representations for both of them. These representations are passed on to a **multi-perspective matching block** that uses four different matching mechanisms - full matching, max-pooling matching, attentive matching and max attentive matching to generate matched representations of all words of both the sentences. This matching takes place in both the directions, i.e. if P and Q are the two input sentences, then representations for all words of P are computed by matching with words of Q, and same is done for all words of Q by matching with all words of P. These representations are then fed into an aggregation layer followed by fully connected layers for classification. In our experiments, we modified this architecture by replacing the aggregation LSTM in the aggregation layer by an aggregating attention layer, and replacing the following fully connected layers by a bilinear layer.

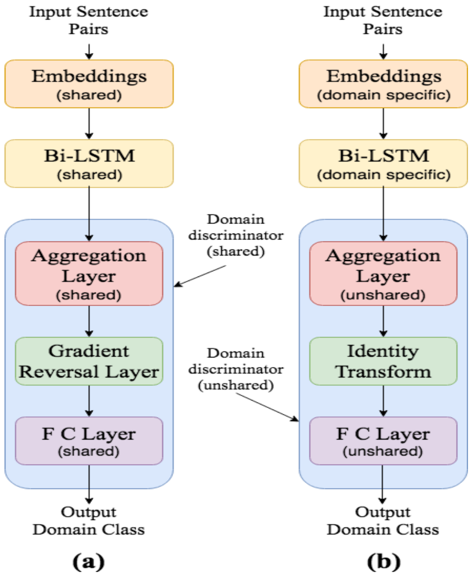

Figure 1: (a) Architecture for data flow of pass 1, (b) Architecture for data flow of passes 2 and 3

2.2 Adversarial Domain Adaptation Methods

The overall architecture used for prediction makes use of both shared and domain specific features. The shared features are learned in an adversarial fashion wherein the desired feature layer that needs to be shared sends its output to a domain discriminator. For our experiments, we plug in this domain discriminator at the base of the model, right after the Bi-LSTM layer. This is to ensure that the layers following Bi-LSTM are trained only for the duplicate classification task, and use domain invariant features generated by the Bi-LSTM. Our work uses two domain discriminators - shared domain discriminator with gradient rever-

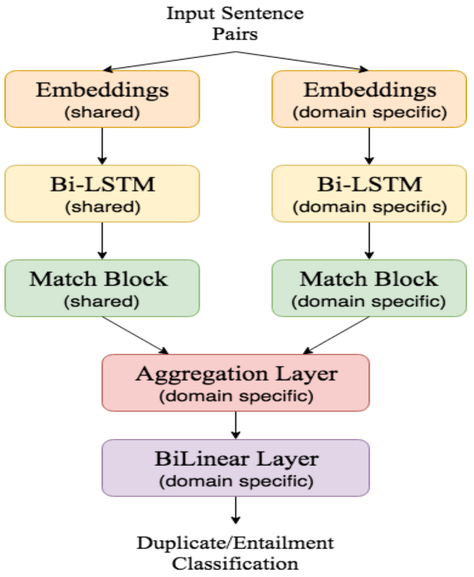
Input Sentence Pairs
Embeddings (shared)
Embeddings (domain specific)
Bi-LSTM (shared)
Bi-LSTM (domain specific)
Match Block (shared)
Match Block (domain specific)
Aggregation Layer (domain specific)
BiLinear Layer (domain specific)
Duplicate/Entailment Classification

Figure 2: Architecture for data flow of passes 4 and 5

sal layer (explained below), that is used to train shared Embedding and Bi-LSTM layers to generate domain invariant features, and unshared domain discriminator that is used to train all the domain specific Embedding and Bi-LSTM layers to generate highly domain specific features. These discriminators consist of an aggregation layer (attention mechanism), followed by a fully connected layer for domain classification (see Figures 1(a) and 1(b)).

The shared domain discriminator uses a **Gradient Reversal Layer (GRL)** (see Figure 1(a)) that acts as an identity transform in the forward pass through the network. During the backward pass however, this layer multiplies the incoming gradient by a negative factor $-\lambda$ which reverses the gradient direction. The use of this layer allows the domain discriminator to be trained in a minimax game fashion, where the domain classification layer tries to minimize the domain classification loss, thus trying to be better at this task, while feature extraction layers (layers before GRL) act as adversaries by trying to make the task harder for domain classification layer. This ensures that feature extraction layers are as ineffective as possible for domain classification, thus bringing the feature maps of both domains closer. As a result, the desired feature layers should generate shared feature representations that are almost indistinguishable by the domain classification layer. The shared features obtained from shared Bi-LSTM should also be more effective to transfer than the ones obtained by simply maximizing the domain classification loss throughout the domain discriminator and base model layers.

The domain specific features are learned using an unshared domain discriminator that is identical to the domain discriminator used for shared features, except that the GRL is replaced by identity transform layer (see Figure 1(b)). This layer however, multiplies the incoming gradient by a positive factor $+\lambda$ to maintain uniformity in gradient magnitudes with shared domain discriminator. This domain discriminator tries to minimize the domain classification loss, as do the preceding layers and thus the desired feature layer learns to generate highly domain specific feature representations.

A block diagram of the proposed adversarial learning framework for domain adaptation has been shown in Figure 3.

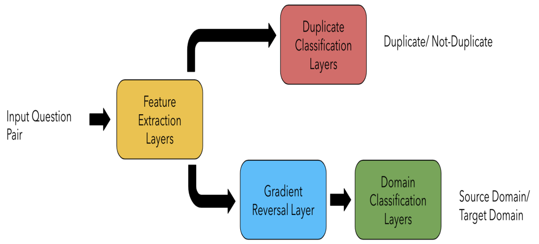
Input Question Pair
Feature Extraction Layers
Duplicate Classification Layers
Duplicate/ Not-Duplicate
Gradient Reversal Layer
Domain Classification Layers
Source Domain/ Target Domain

Figure 3: Adversarial Learning Framework for Domain Adaptation

2.3 Model Architecture

The training data has sentence pairs (Q_S) from source domain S, and sentence pairs (Q_T) from target domain T. Figures 1 and 2 show the overall architecture of the model. The initial layers of the network - Embedding, Bi-LSTM and multi-perspective match block - are of two kinds: shared and domain specific. Shared layers are used in the network for sentences of all domain types, whereas the domain specific layers work on sentences of only corresponding domains. The Embedding layers can be appropriately initialized and trained end-to-end along with the rest of the network. Each domain has domain specific aggregation and classification (fully connected) layers as well. The aggregation layer takes in the domain specific and shared features as inputs (Figure 2), aggregates them and concatenates these aggregated vectors to form a combined representation.

This combined feature vector is passed to the classification layers for task classification.

2.4 Model Training

The forward propagation through the model involves 5 passes, which are listed below:

- **Pass 1 (Figure 1(a)) -** Q_S and Q_T through shared layers and shared domain discriminator (Loss = L_1).
- **Pass 2 (Figure 1(b)) -** Q_S through domain specific layers and unshared domain discriminator (Loss = L_2).
- **Pass 3 (Figure 1(b)) -** Q_T through domain specific layers and unshared domain discriminator (Loss = L_3).
- **Pass 4 (Figure 2) -** Q_S through domain specific and shared layers for task classification (Loss = L_4).
- **Pass 5 (Figure 2) -** Q_T through domain specific and shared layers for task classification (Loss = L_5).

The source domain layers are trained by minimizing L_S (Equation 1). The target domain layers are trained by minimizing L_T (Equation 2). The shared embedding, Bi-LSTM and aggregation layers are learned by minimizing L_{Sh} (Equation 3), while fully connected layer of shared domain discriminator minimizes L_1.

$$L_S = L_2 + L_4 \quad (1) \qquad L_T = L_3 + L_5 \quad (2)$$

$$L_{Sh} = L_4 + L_5 - \lambda L_1 \quad (3)$$

Note that not all domain specific layers contribute to losses L_2 and L_3, and thus the gradient due to these losses affects only the Embedding and Bi-LSTM layers for all domains. We trained all the models and tuned all the hyperparameters to optimize the validation set performance on target domain data.

3 Experiments

3.1 Datasets

For question duplicate detection, we use the Quora question pairs dataset(Quora, 2017) as the source domain dataset and 5 datasets that are from different and diverse set of domains as our target domains. The Android, Mathematica, Programmers and Unix question datasets were used from the Stack Exchange dataset (StackExchchange, 2018). We obtained the Tax Domain Qs from a popular forum for tax related question answers, which we plan to make public shortly. For RTE, the Stanford Natural Language Inference (SNLI) (SNLI, 2015) has been used as source domain, and for target domains we used The Guardian Headlines RTE (RTE, 2012) and SICK (SICK, 2014) datasets. The size for all these datasets has been mentioned in Table 1 in the (train/ validation/ test) format.

3.2 Results

In Table 1 we compared the base model BiMPM **(base)** trained only on the target domains to three variants of the same model, each obtained after a different approach for adversarial domain adaptation. **Model T1** was trained by using both the shared and domain specific features, but maximizing the domain classification loss to learn shared features. **Model T2** used only the shared features learned using gradient reversal strategy, along with fine-tuned features obtained from later layers of the network. **Model T3** used both the domain specific features as well as the shared features learned using the gradient reversal method. The accuracy of these models for five different question deduplication and two RTE target domains is reported in Table 1. Comparisons of accuracy numbers between different rows are fairly consistent across all domains[1], enabling us to draw the following empirical claims:

T1, T2 and T3 outperform baseline, hence enforcing the effectiveness of adversarial domain adaptation in all tasks in Table 1.

T3 outperforms T2, thus indicating that learning a combination of domain specific and shared representations is quite beneficial for all domain transfer experiments in Table 1. This observation was also noted by Qiu *et al* (Qiu et al., 2018), even if without the use of gradient reversal.

Both T2 and T3 outperform T1, hence providing strong evidence that GRL significantly improves overall feature learning if compared to maximizing the domain classification loss. In particular, the comparison between T3 and T1, shows that learning exactly the same feature set using GRL for adversarial domain adaptation is more effective than maximizing the loss.

T3 outperforms all other models, showing that our proposed approach consistently beats all other settings for domain adaptation in both ques-

[1] All row differences are statistically significant on paired t-test(p-value< 0.05)

Model (BiMPM)	Adversarial Approach	Features	Question Duplicate Detection					Textual Entailment	
			Tax Domain	**Android**	**Mathematica**	**Programmers**	**Unix**	**Guardian**	**SICK**
			(3k/ 1k/ 1k)	(7k/ 1.5k/ 1.5k)	(5.4k/ 1.2k/ 1.2k)	(6.5k/ 1.5k/ 1.5k)	(7k/ 1.5k/ 1.5k)	(23k/ 5k/ 5k)	(6.8k/ 1.5k/ 1.5k)
base	–	DSF	84.7	90.7	80.0	90.7	88.7	92.3	69.5
T1	maxLoss	SF + DSF	87.6	91.3	82.1	91.6	89.6	94.3	72.7
T2	GRL	SF	88.1	92.0	82.6	91.9	90.8	96.4	73.8
T3	GRL	SF + DSF	**89.3**	**92.6**	**83.0**	**92.4**	**91.1**	**97.4**	**76.4**

Table 1: Comparison of Accuracy for different domain adaptation methods; Source domain for question duplicate detection: Quora (240k/ 80k/ 80k), Source domain for RTE: SNLI (550k/ 10k/ 10k); **SF:** shared features, **DSF:** domain specific features, **maxLoss:** maximizing domain discriminator loss, **GRL:** gradient reversal layer

tion duplicate classification and RTE.

4 Discussion and Conclusion

We systematically evaluated different adversarial domain adaptation techniques for duplicate question detection and RTE tasks. Our experiments showed that adversarial domain adaptation using gradient reversal yields the best knowledge transfer between all textual domains in Table 1. This method outperformed existing domain adaptation techniques, including recently proposed adversarial domain adaptation method of maximizing the domain classification loss by a discriminator. Furthermore, we show that the models that use both domain specific features and shared features outperform the models that use only either of these features.

References

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *30th Conference on Neural Information Processing Systems (NIPS 2016)*, pages 343–351.

Yue Cao, Mingsheng Long, and Jianmin Wang. 2018. Unsupervised domain adaptation with distribution matching machines. In *AAAI*.

Daniel Cohen, Bhaskar Mitra, Katja Hofmann, and W. Bruce Croft. 2018. Cross domain regularization for neural ranking models using adversarial learning. In *SIGIR'18: 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1025–1028. ACM.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21. ACL.

Lisheng Fu, Thien Huu Nguyen, Bonan Min, and Ralph Grishman. 2017. Domain adaptation for relation extraction with domain adversarial neural network. In *Proceedings of the The 8th International Joint Conference on Natural Language Processing*, pages 425–429. ACL.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *ICML'15: Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1180–1189. ACM.

Doris Hoogeveen, Karin M. Verspoor, and Timothy Baldwin. 2015. Cqadupstack: A benchmark data set for community question-answering research. In *ADCS*.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. 2017. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2208–2217, International Convention Centre, Sydney, Australia. PMLR.

Minlong Peng, Qi Zhang, Yu-gang Jiang, and Xuanjing Huang. 2018. Cross-domain sentiment classification with target domain specific information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2505–2513. ACL.

Minghui Qiu, Liu Yang, Feng Ji, Wei Zhou, Jun Huang, Haiqing Chen, Bruce Croft, and Wei Lin. 2018. Transfer learning for context-aware question matching in information-seeking conversations in e-commerce. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–213. Association for Computational Linguistics.

Quora. 2017. Quora Duplicte Questions Dataset. `https://www.kaggle.com/c/quora-question-pairs/data`.

Guardian Headlines RTE. 2012. The Guardian Headlines Entailment Training Dataset. `https://github.com/daoudclarke/rte-experiment`.

Swami Sankaranarayanan, Yogesh Balaji, Carlos D. Castillo, and Rama Chellappa. 2018. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512. IEEE.

Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. 2016. Learning transferrable representations for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, pages 2110–2118.

Darsh Shah, Tao Lei, Alessandro Moschitti, Salvatore Romeo, and Preslav Nakov. 2018. Adversarial domain adaptation for duplicate question detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1056–1063. Association for Computational Linguistics.

SICK. 2014. Sentences Involving Compositional Knowledge (SICK). http://clic.cimec.unitn.it/composes/sick.html.

SNLI. 2015. The Stanford Natural Language Inference Corpus. https://nlp.stanford.edu/projects/snli/.

Kihyuk Sohn, Wenling Shang, Xiang Yu, and Manmohan Chandraker. 2019. Unsupervised domain adaptation for distance metric learning. In *International Conference on Learning Representations*.

StackExchchange. 2018. Stack Exchange Data Dump. https://archive.org/download/stackexchange.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176. IEEE.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4144–4150. AAAI.

Kai Zhao, Liang Huang, and Mingbo Ma. 2017. Textual entailment with structured attentions and composition. *arXiv preprint arXiv:1701.01126*.