SATNet: Symmetric Adversarial Transfer Network Based on Two-Level Alignment Strategy towards Cross-Domain Sentiment Classification (Student Abstract)

Yu Cao,[1,2] **Hua Xu**[1,2]
[1]State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, China
[2] Beijing National Research Center for Information Science and Technology(BNRist), Beijing 100084, China
caoy18@mails.tsinghua.edu.cn, xuhua@tsinghua.edu.cn

Abstract

In recent years, domain adaptation tasks have attracted much attention, especially, the task of cross-domain sentiment classification (CDSC). In this paper, we propose a novel domain adaptation method called Symmetric Adversarial Transfer Network (SATNet). Experiments on the Amazon reviews dataset demonstrate the effectiveness of SATNet.

Introduction

Sentiment classification is an important task in natural language processing (NLP), and it aims to assign the sentiment polarity towards a given text. However, the performances of neural network-based sentiment analysis methods are highly dependent on large manually labeled training data. Thus, cross-domain sentiment classification, which aims to borrow knowledge learned on labeled data from related domains (called source domain) to a new domain (called target domain), becomes a promising direction.

Over the last decades, researchers have proposed various methods for cross-domain sentiment classification . For example, traditional pivot-based work (Blitzer, Dredze, and Pereira 2007) attempts to infer the correlation between pivot words, i.e., the domain-shared sentiment words, and non-pivot words, i.e., the domain-specific sentiment words by using multiple pivot prediction tasks. However, this method needs to manually select the pivots before adaptation. Recently, many researchers apply the unlabeled data for training to learn representations shared across domains, i.e., Neural Network with Auxiliary Task(AuxNN) (Yu and Jiang 2016).Recently, some existing adversarial learning methods (Ganin et al. 2016; Li et al. 2017; Zhang, Miao, and Wang 2019) reduce feature difference by fooling a domain discriminator. However, these adversarial methods simply align the marginal distribution of the two domains ,ignoring the category-specific decision boundaries. Some recent works with category and domain level alignment have been explored in computer vision applications (Zhang et al. 2019).

This paper proposes a novel design of Symmetric Adversarial Transfer Network (SATNet) to facilitate, via adversarial training,the alignment of the joint distributions of feature and category across data domains. Our SATNet is based on the symmetric design of a classifier that shares its layer neurons with the source and target sentiment classifiers. In addition, we propose a novel adversarial training method to learn the SATNet, which includes category-level and domain-level alignment losses and can thus enhance domain-invariant feature learning towards the category level.

Symmetric Adversarial Transfer Network

In this section, we first present an overview of our proposed SATNet model. Then we detail the model with three related works. Finally, we explain the entire training procedure of our method.

Our method is composed of feature extractor G and fully-connected (FC) layer as the classifiers which include source domain classifier F^t and target domain classifier F^s. The domain discriminator D, which shares its layer neurons with F^s and F^t, is used to distinguish features of samples from the two domains. In order to classify the target samples correctly, we aim to align joint distributions of feature and category across domains by utilizing three related tasks .

Related Task A: In the task A, two classifiers (F^s, F^t)are used to classify source labeled samples correctly. We also train a domain discriminator D on feature representations of different domains extracted by G.

Related Task B: In the task B, we aim to learn G to achieve category-level alignment. For source domain data, we force the consistency of the output of the category-corresponding neurons between the source domain classifierF^s and the target domain classifier F^t to achieve the category alignment.

Related Task C: In the task C, we consider learning G to maximally “confuse” the two domains to achieve domain-level alignment. To maximize the role of domain confusion, we decided to use target domain data to balance the output of source domain classifierF^s and target domain classifierF^t .

The whole training procedure can be divided into two steps. In the step one, we consider both minimizing the classification error and domain prediction error. In the step two,

Tasks	SVM	mSDA	AuxNN	DANN	AMN	DAS	HAGAN-C	ACAN	SATNet-D	SATNet-M	SATNet
D→B	75.20	78.50	80.80	81.70	81.52	82.05	81.69	82.35	**83.75**	82.25	83.50
E→B	68.85	76.15	78.00	78.55	77.80	80.00	79.23	79.75	77.50	77.50	**80.50**
K→B	70.00	75.65	77.85	79.25	79.37	80.05	78.99	80.80	76.50	79.50	**82.25**
B→D	77.15	80.60	81.75	82.30	81.32	82.75	82.38	83.45	80.75	80.50	**83.75**
E→D	69.50	76.30	80.65	79.70	77.51	80.15	80.65	**81.75**	78.00	80.00	81.00
K→D	71.40	76.05	78.90	80.45	80.03	81.40	80.91	82.10	81.00	**82.50**	82.00
B→E	72.15	75.55	76.40	77.60	80.07	81.15	80.12	**81.20**	77.75	78.00	79.25
D→E	71.65	76.00	77.55	79.70	80.00	81.55	80.99	**82.80**	79.75	81.00	82.75
K→E	79.75	84.20	84.05	**86.65**	81.97	85.80	85.23	86.60	85.25	86.25	86.00
B→K	73.50	75.95	78.10	76.10	81.00	82.25	82.00	**83.05**	77.75	80.25	80.50
D→K	72.00	76.30	80.05	77.35	**83.88**	81.50	81.50	78.60	76.50	77.50	78.25
E→K	82.80	84.45	84.15	83.95	87.10	84.85	84.99	83.35	86.50	87.00	**88.00**
Avg	73.66	77.98	79.85	80.29	80.96	81.96	81.56	82.15	80.08	81.02	**82.31**

Table 1: Accuracies on the Amazon dataset.

category and domain level alignment losses for updating the feature extractor G, we also adopt entropy minimization principle (Grandvalet and Bengio 2005) to increase the prediction confidence on the target domain.

Experiments

In this section, we conduct the experiments on the Amazon reviews dataset (Blitzer, Dredze, and Pereira 2007) , which has been widely used for cross-domain sentiment classification. This dataset contains reviews from four different domains: Books (B), DVD (D), Electronics (E), Kitchen (K).

In our implementation, the feature encoder G consists of three parts including a 300-dimensional word embedding layer, a one-layer CNN with ReLU activation function adopted in (He et al. 2018) and a max-over-time pooling through which final sentence representation is obtained. Similarly, we add batch normalization layer before activation function.

Experiment Results: SATNet-D is a variant of our model which removes domain-level alignment loss. SATNet-M is a variant of our model which removes entropy minimization loss. Table 1 reports the classification accuracies of different methods on the Amazon reviews. It is obvious to see that the proposed **SATNet** outperforms all other methods generally, such as SVM, DANN (Ganin et al. 2016), mSDA (Chen et al. 2012), DAS (He et al. 2018), AuxNN (Yu and Jiang 2016), AMN (Li et al. 2017), HAGAN-C (Zhang, Miao, and Wang 2019), ACAN (Qu et al. 2019).

Conclusion

In this paper, we propose a novel adversarial approach termed symmetric adversarial transfer network (SATNet) for cross-domain sentiment classification, which achieves category-level and domain-level alignment across domains via two-level alignment strategy. Experiments on Amazon reviews dataset verify the effectiveness of our proposed SATNet.

Acknowledgments:Project (Grant No: 61673235) supported by National Natural Science Foundation of China.

References

Blitzer, J.; Dredze, M.; and Pereira, F. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, 440–447.

Chen, M.; Xu, Z.; Weinberger, K.; and Sha, F. 2012. Marginalized denoising autoencoders for domain adaptation. *arXiv preprint arXiv:1206.4683*.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17(1):2096–2030.

Grandvalet, Y., and Bengio, Y. 2005. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, 529–536.

He, R.; Lee, W. S.; Ng, H. T.; and Dahlmeier, D. 2018. Adaptive semi-supervised learning for cross-domain sentiment classification. *arXiv preprint arXiv:1809.00530*.

Li, Z.; Zhang, Y.; Wei, Y.; Wu, Y.; and Yang, Q. 2017. End-to-end adversarial memory network for cross-domain sentiment classification. In *IJCAI*, 2237–2243.

Qu, X.; Zou, Z.; Cheng, Y.; Yang, Y.; and Zhou, P. 2019. Adversarial category alignment network for cross-domain sentiment classification. 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)*, 2496–2508.

Yu, J., and Jiang, J. 2016. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 236–246.

Zhang, Y.; Tang, H.; Jia, K.; and Tan, M. 2019. Domain-symmetric networks for adversarial domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5031–5040.

Zhang, Y.; Miao, D.; and Wang, J. 2019. Hierarchical attention generative adversarial networks for cross-domain sentiment classification. *arXiv preprint arXiv:1903.11334*.