Gated Multi-Task Network for Text Classification

Liqiang Xiao and **Honglun Zhang** and **Wenqing Chen**
State Key Lab of Advanced Optical Communication System and Network,
Shanghai Jiao Tong University
Artificial Intelligence Institute, Shanghai Jiao Tong University
800 Dongchuan Road, Shanghai, China
{xiaoliqiang,zhanghonglun,wenqingchen}@sjtu.edu.cn

Abstract

Multi-task learning with Convolutional Neural Network (CNN) has shown great success in many Natural Language Processing (NLP) tasks. This success can be largely attributed to the feature sharing by fusing some layers among tasks. However, most existing approaches just fully or proportionally share the features without distinguishing the helpfulness of them. By that the network would be confused by the helpless even harmful features, generating undesired interference between tasks. In this paper, we introduce gate mechanism into multi-task CNN and propose a new Gated Sharing Unit, which can filter the feature flows between tasks and greatly reduce the interference. Experiments on 9 text classification datasets shows that our approach can learn selection rules automatically and gain a great improvement over strong baselines.

1 Introduction

The combination of multi-task learning and neural networks has shown its advantages in many tasks, ranging from computer vision (Misra et al., 2016; Ruder et al., 2017) to natural language processing (Collobert and Weston, 2008). Multi-task learning (MTL) has the ability to share the knowledge among the joint tasks, which implicitly increases the training materials (Caruana, 1997). The shared knowledge help the network learn a more universal representation for the inputs. Inspired by this, more DNN-based approaches (Liu et al., 2015; Zhang et al., 2017) utilize multi-task learning to improve their performance.

The scheme for information sharing is the linchpin for designing an elaborate multi-task network. Most existing work attempts to find a appropriate proportion to sharing the layers between tasks, despite they entirely reuse the shallow layers (Liu et al., 2015; Caruana, 1993) or add the layers up at a ratio (Fang et al., 2017). And recently, the latter one shows its advantages for controlling relational intensity among tasks and become prevailing. More models adopt this thought to enhance the performance (Liu et al., 2015, 2016).

However, under the scheme of proportional addition (Ruder et al., 2017; Misra et al., 2016), all the features are shared with the same weight between every pair of tasks. Helpless or harmful features may be transported between tasks with the same importance as helpful ones, namely, the interference is generated. This would burden the network for distinguishing the helpful features and even mislead the predictions.

To solve above problem, we propose a new CNN-based architecture for multi-task learning, which can share features in a selective way. Our model allocates a private subnet to each task and transport the features between the subnets with a well-designed module—*Gated Sharing Unit*. It has the ability to filter features with gate mechanism (Chung et al., 2014; Srivastava et al., 2015) and select the helpful ones to benefit the tasks in hand, which expands the feature spaces and provides more evidence for right predictions. Our model is an end-to-end method and the proposed Gated Sharing Unit is easy to train.

We conduct extensive experiments on 9 benchmark datasets for text classification. The results show that our model greatly improves the performance and surpasses the single-task models and other competitors.

2 Gated Multi-Task Network

To make full use of multiple datasets and, meanwhile, avoid the interference, we introduce a new structure for multi-task learning in this section. The new structure is designed in a separative way—every task owns a private subnet. To share

Proceedings of NAACL-HLT 2018, pages 726–731
New Orleans, Louisiana, June 1 - 6, 2018.

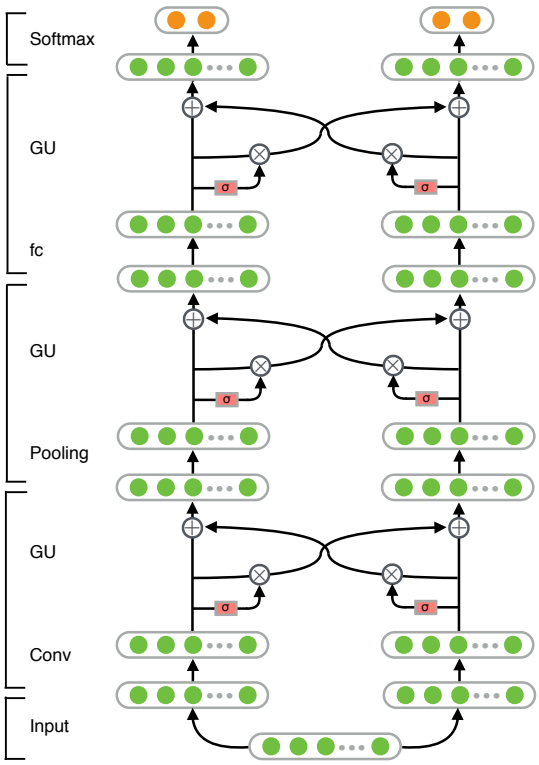

Figure 1: Illustration of the architecture for Gated Multi-Task CNN .

features across the subnets, gate mechanism is designed to selectively allow the features been exchanged. Our new model can be trained end-to-end, needing no extra supervision or handcraft hyperparameters. And it can be easily transferred to other networks such as DNN, RNN, LSTM, etc. Figure 1 illustrates the design of model structure and other details.

2.1 Model Architecture

Multi-task model with deeper layers shared can augment deeper knowledge and greatly increase the feature space (Zhang et al., 2017). But undesirable interference inevitably and simultaneously comes with the benefits, especially between less-related tasks. This would burden the models with the overhead on distinguishing helpful features. To overcome this problem, we assign each task a private subnet as illustrated in Figure 1. Tasks are relatively separated and can borrow the useful information from others through a bridge, Gated Sharing Unit (GSU). The weight of each feature in this unit is automatically learned from previous layers, needing no extra supervision, so there is more selectivity across the tasks. By filtering out useless features, tasks receive less interference

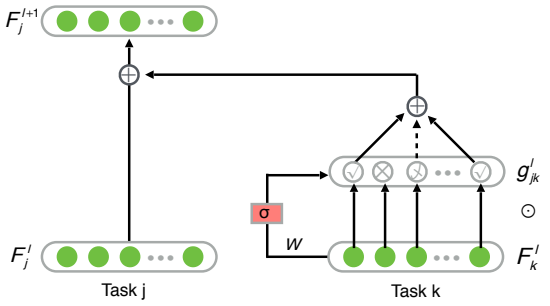

Figure 2: Illustration of Gated Sharing Unit

from each other.

2.2 Gated Sharing Unit

For reducing interference, it important to filter the information flows among the tasks. Hence, in this section, we introduce the mechanism of gate, which originates from the cells of recurrent neural networks like Long-Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), Gated Recurrent Unit (GRU) (Chung et al., 2014). Gated mechanism in existing studies not only shows its convenience for training (Srivastava et al., 2015), but also behaves as tool to route the information (He et al., 2016).

Inspirited by gate mechanism, we propose a new module GSU to control the information flows and selectively share the features among tasks. The details of this module is shown in Figure 2. For notation, we refer to C as the collection of N tasks and $C = \{1, 2, \cdots, N\}$. For a sample from arbitrary task j, a series of feature maps are generated in subnets. When task j borrows the features from task k, a gate $\mathbf{g}$ is inserted to select the helpful ones, which is calculated from the prior layer by

$$\mathbf{g}_{jk}^l = \sigma(\mathbf{W}_{jk}^l \cdot \mathbf{F}_k^l + \mathbf{b}_{jk}^l) \quad (1)$$

where l means the level of the layers and σ denotes the nonlinear activation of sigmoid, which guarantees the values of $\mathbf{g}$ in the $[0, 1]$. Note that the gate $\mathbf{g}_{jk}^l$ is vector. Each component in it controls the pass of a corresponding feature. Their states move between pass and interception, or choose a middle ground if needed.

For task j, the output $\mathbf{F}_j^{l+1}$ of gates is calculated by fusing the lower layers $\mathbf{F}^l$ from all the tasks by

$$\mathbf{F}_j^{l+1} = \sum_{k \in C, k \neq j} \mathbf{g}_{jk}^l \odot \mathbf{F}_k^l + \mathbf{F}_j^l \quad (2)$$

where $\odot$ denotes element-wise multiplication. To

represent the output for all the tasks C, we can stack Eq. (2) in matrix form

$$\begin{bmatrix} \mathbf{F}_1^{l+1} \\ \mathbf{F}_2^{l+1} \\ \vdots \\ \mathbf{F}_N^{l+1} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{g}_{12}^l & \cdots & \mathbf{g}_{1N}^l \\ \mathbf{g}_{21}^l & 1 & \cdots & \mathbf{g}_{2N}^l \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{g}_{N1}^l & \mathbf{g}_{N2}^l & \cdots & 1 \end{bmatrix} \begin{bmatrix} \mathbf{F}_1^l \\ \mathbf{F}_2^l \\ \vdots \\ \mathbf{F}_N^l \end{bmatrix} \tag{3}$$

From Eq. (2) and (3), we know that, in the GSU, the feature map for current task directly passes into the next layer. But the features from other tasks are merged into current task after the selection by gates. In this way, the shared features tend to be pure and helpful for current task, which avoids the harmful interference existing in conventional models.

For comparison, here we briefly introduce the methods that share the features by proportional addition (Misra et al., 2016; Ruder et al., 2017; Fang et al., 2017). They can be constructed by inserting a scaler weight α_{jk}^l between every two tasks i, j. α_{jk}^l is updated by back-propagation and reflects the degree of association between tasks, but do not select the features. In this paper, this kind of models is alluded to as PA-CNN.

2.3 Output Layer and Loss

In the last layer of task j, vector representations $\hat{\mathbf{F}}_j$ of input sequences are ultimately fed into corresponding softmax layers to fit the number of classes, which emits the prediction of probability distribution for the task j

$$\hat{y}_j = softmax(\mathbf{W}_j \hat{\mathbf{F}}_j + b_j) \tag{4}$$

where $\hat{y}_j$ is predictive result; $\mathbf{W}_j$ is the weight of the full-connected layer; and b_j is the bias term.

Given the prediction of all tasks, a global loss function forces the model to minimize the cross-entropy of prediction and true distribution for all the tasks:

$$\Phi = \sum_{j=1}^{N} \lambda_j L(\hat{y}_j, y_j) \tag{5}$$

where λ_j is the weight for the task j. In this paper, we set λ_j to $1/N$ for all N tasks to make a balance.

3 Experiments

In this section, we demonstrate the empirical performance of our model on 9 related benchmark tasks for text classification. And the results are compared with the state-of-the-art models.

Dataset	Train	Dev.	Test	V	L
Books	1398	200	400	22K	159
Electronics	1398	200	400	11K	111
DVDs	1400	200	400	22K	189
Kitchen	1400	200	400	10K	93
Apparel	1400	200	400	8K	64
Baby	1300	200	400	9K	173
RN	7860	1122	2246	29K	147
SUBJ	8000	1000	1000	21K	23
TREC	4907	545	500	10K	10

Table 1: Statistics of the text classification datasets. Train, Dev. and Test denote the size of train, development and test set respectively; C: Vocabulary size; L: Average sentence length.

3.1 Datasets

As Table 1 shows, we select 9 related benchmark datasets for text classification.

The first 6 datasets are all about product reviews, which are comprised of Amazon product reviews in 6 domains, including books, DVDs, cameras, etc. These corpora are classified according to the sentiment of positiveness or negativeness. They are collected from the raw data published by (Blitzer et al., 2007).

The rest 3 datasets are RN, SUBJ and TREC. RN is a dataset about news topic classification, which is collected from Reuters Newswire and published by (Velasco et al., 1994); SUBJ is a subjectivity dataset, whose task is to classify a sentence level text as being subjective or objective (Pang et al., 2004); TREC dataset has the task of classifying a question into 6 types (the questions are about location, person, numeric information, etc.)(Li and Roth, 2002).

3.2 Hyperparameters and Training

For all the experiments, we employ Word2Vec (Mikolov et al., 2013) to initialized the word vectors, which is trained on Google News with 100 billion words. The vectors have dimensionality of 300 and are trained by continuous bag-of-words architecture. All the other parameters are initialized with random values from uniform distribution in [-0.1, 0.1]. For every subnet we use: rectified linear units, filter windows of 3,4,5 with 100 feature maps each, mini-batch size of 50, dropout rate of 0.5, l_2 constrain of 3, learning rate of 10^{-3}. All the hyper-parameters are chosen via a small grid search on dev set. For the dataset without a stan-

Dataset	Single-Task (%)			Multi-Task (%)				
	DCNN	LSTM	BiLSTM	MT-DNN	MT-RNN	MT-CNN	PA-CNN	GMT-CNN
Books	80.7	79.5	81.0	82.3	83.3	84.0	82.2	**84.4**
Electronics	78.3	80.5	78.5	81.6	84.6	83.1	84.8	**86.9**
DVDs	80.6	81.7	80.5	83.8	84.2	84.0	83.7	**85.4**
Kitchen	79.8	78.0	81.2	80.8	**86.0**	83.4	85.1	85.9
Apparel	84.2	83.2	86.0	85.1	86.3	83.6	**87.2**	87.0
Baby	84.1	84.7	84.5	88.0	87.6	87.8	86.5	**88.3**
RN	83.6	83.5	83.7	83.9	84.2	84.3	83.6	**85.0**
SUBJ	93.0	93.1	93.2	92.7	**94.1**	92.9	93.1	94.0
TREC	93.0	92.7	93.0	93.2	93.5	93.7	93.3	**94.2**
Avg.	84.1	84.1	84.6	$85.7_{(+1.1)}$	$87.0_{(+2.4)}$	$86.3_{(+1.7)}$	$86.6_{(+2.0)}$	**87.9**$_{(+3.3)}$

Table 2: Accuracies of our model against other state-of-the-art methods. Single-Task column shows the results of plain DCNN(Kalchbrenner et al., 2014), LSTM(Jozefowicz et al., 2015) and BiLSTM. First 3 models in the Multi-Task column shows the results of multi-task models: MT-DCNN (Liu et al., 2015), MT-RNN (Zhang et al., 2017), MT-CNN (Collobert and Weston, 2008). The remaining columns of PA-CNN and GMT-CNN shows the performance of proportional addition or gate mechanism. Number in round bracket denotes the average improvement over BiLSTM.

dard dev set we randomly select 10% as dev set. The whole network is trained through stochastic gradient decent using Adadelta update rule (Zeiler, 2012).

3.3 Performance of Multi-task CNN

Table 2 shows the comparison of the accuracies. All the results for multi-task learning models are achieved by training simultaneously on 9 datasets. From the table, we can see that the models employing multi-task learning improve the performance on most tasks beyond the single-task models, in which our model achieves the highest accuracies. Specifically, our model boosts the performance by 3.3% over the best single-task model BiLSTM, outstripping other multi-task models by at least 0.9%. Additionally, we also compare our model with the PA-CNN, a variant keeps the structure of GMT-CNN but shares the features by proportional additions. For PA-CNN, performance on several datasets is decreased than single-task due to the interference. In contract, our model shows steady improvement in all the datasets and surpasses PA-CNN by 1.3%, which indicates the effectiveness of gate mechanism.

3.4 Visualization

To intuitively show the selection process, we design an experiment to show the values of gates and how they block the useless features. For the first convolutional layer and GSU, we visualize the activations $\mathbf{F}^1_j$ of the filters with normalized values and show their corresponding weights $\mathbf{g}^1_{jk}$ in the gate units. By that we can easily find what kind of features are discarded as interference.

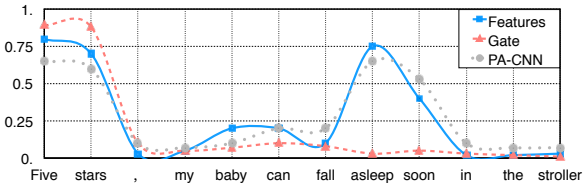

Figure 3: Features line illustrates feature weights in $\mathbf{F}^1_{dvds}$ in DVDs subnet. And Gate line shows the value of $\mathbf{g}^1_{baby \leftarrow dvds}$ that filters the features from DVDs subnet to Baby subnet. PA-CNN line visualizes the feature weights in the first layer of PA-CNN.

Figure 3 illustrates the behavior of GSU on a random selected sentence from Baby task. We visualize the results of the first feature map for DVDs subnet and the gate unit that filters the features from DVDs to Baby task. For the positive sentence "Five stars, my baby can fall asleep soon in the stroller", we can see that subnet for DVDs task focuses on two critical positions "Five stars" and "asleep". The word "asleep" is negative for DVDs task, but actually neutral for Baby task. Successfully, our gated unit lowers the intensity of the interference "asleep", making a correct prediction. However, PA-CNN wrongly makes a negative prediction for lacking resistance to interference. This indicates the effectiveness of our gate mechanism for the feature selection in MTL.

4 Conclusion and Future Work

In this paper, we introduce gate mechanism in multi-task CNN to reduce the interference. The proposed model has an ability to select the potentially useful features, which can reduce the interference among tests. The effectiveness of our method is fully validated on 9 datasets for text classification and further illustrated by visualization experiment.

In future work, we would like to investigate the effect of memory mechanism for multi-task learning, which is similar to gate mechanism but more complex. It originates from recurrent neural network and have been proven effective for feature selection.

5 Acknowledge

We would like to thank Yaohui Jin and Yongkun Wang for their careful guidance. Besides, we appreciate the constructive advices from Naoki Yoshinaga and the valuable comments from anonymous reviewers. This research was funded by National Natural Science Foundation of China under Grant No. 61371048.

References

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. http://aclweb.org/anthology/P07-1056.

Rich Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Machine Learning, Proceedings of the Tenth International Conference, University of Massachusetts, Amherst, MA, USA, June 27-29, 1993*. pages 41–48.

Rich Caruana. 1997. Multitask learning. *Machine Learning* 28(1):41–75.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR* abs/1412.3555. http://arxiv.org/abs/1412.3555.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*. pages 160–167. https://doi.org/10.1145/1390156.1390177.

Yuchun Fang, Zhengyan Ma, Zhaoxiang Zhang, Xu-Yao Zhang, and Xiang Bai. 2017. Dynamic multi-task learning with convolutional neural network. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. pages 1668–1674. https://doi.org/10.24963/ijcai.2017/231.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. pages 770–778. https://doi.org/10.1109/CVPR.2016.90.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. 2015. An empirical exploration of recurrent network architectures. In *International Conference on International Conference on Machine Learning*. pages 2342–2350.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*. pages 655–665. http://aclweb.org/anthology/P/P14/P14-1062.pdf.

Xin Li and Dan Roth. 2002. Learning question classifiers. *Coling* 12(24):556–562.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning .

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 912–921.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 26:3111–3119.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. 2016. Cross-stitch networks for multi-task learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pang, Bo, Lee, and Lillian. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of Acl* pages 271–278.

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Sgaard. 2017. Sluice networks: Learning what to share between loosely related tasks .

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks. *CoRR* abs/1505.00387. http://arxiv.org/abs/1505.00387.

E Velasco, L. C. Thuler, C. A. Martins, L. M. Dias, and V. M. Gonalves. 1994. Automated learning of decision rules for text categorization. *Acm Transactions on Information Systems* 12(3):233–251.

Matthew D. Zeiler. 2012. Adadelta: An adaptive learning rate method. *Computer Science* .

Honglun Zhang, Liqiang Xiao, Yongkun Wang, and Yaohui Jin. 2017. A generalized recurrent neural architecture for text classification with multi-task learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. pages 3385–3391. https://doi.org/10.24963/ijcai.2017/473.