VISUAL-TEXTUAL SENTIMENT ANALYSIS IN PRODUCT REVIEWS

Jin Ye[⋆], *Xiaojiang Peng*[⋆], *Yu Qiao*[⋆], *Hao Xing*[†], *Junli Li*[†], *and Rongrong Ji*[‡].

[⋆]Shenzhen Key Lab of Computer Vision and Pattern Recognition, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[†]VIPShop Company, GuangZhou, China
[‡]School of Information Science and Engineering, Xiamen University, Xiamen, China

ABSTRACT

Sentiment analysis has attracted increasing attention recently due to its potential wide applications in opinion analysis, recommendation system, etc. Visual-textual sentiment analysis aims to improve the performance of sentiment analysis by leveraging both visual and textual signals. In this paper, we address the visual-textual sentiment analysis in product reviews. Our main contributions are two-fold. First, instead of crawling data from Flickr or Twitter with positive and negative labels in existing works, we introduce a new dataset for visual-textual sentiment analysis, termed as Product Reviews-150K (PR-150K), which is collected from the product reviews of online shopping websites. Second, we propose a deep Tucker fusion method for visual-textual sentiment analysis, which efficiently combines visual and textual deep representations based on the Tucker decomposition and a bilinear pooling operation. Extensive experiments on our PR-150K, MVSO, and VSO datasets show that our method outperforms several state-of-the-art methods.

Index Terms— sentiment analysis, product reviews, tucker decomposition, DTF

1. INTRODUCTION

With the rapid development of the mobile Internet and mobile devices, increasing people are willing to share their opinions by uploading visual and textual information on social platforms and online shops, such as Twitter, Flickr, Ebay, Taobao, etc. Sentiment analysis aims to categorize textual or other information into several sentiment classes, such as positive, negative, and neutral. Due to its potential wide applications in opinion analysis and recommendation system, sentiment analysis with textual and visual information has become a highly active research area recently [1, 2, 3, 4, 5, 6].

Visual-textual sentiment analysis aims to leverage both visual and textual information for accurate sentiment analysis. Texts and images are two main clues for sentiment analysis, thus visual representations, textual representations, and fusion strategies are vital for sentiment analysis performance. Most of early methods are based on text information only, and consider the task as a special case of text classification [2, 3, 4, 5, 7, 8]. Hu et al. [4] propose an unsupervised method which leverages correlated emotional signals from social media for sentiment analysis. Maas et al. [2] propose to use a mix of unsupervised and supervised techniques to learn word vectors which capture both semantic information and rich sentiment content. Le et al. [7] propose to learn distributed representations for documents and utilize it to textual sentiment analysis. Several other sentiment analysis researches are based on images only. Wang et al. [9] propose a Deep Coupled Adjective and Noun (DCAN) neural network for visual sentiment classification. You et al. [10] present a Progressively Trained Convolutional Neural Network (PCNN) for image sentiment analysis, and transfer the model from Flickr images to Twitter images. Recent works address the sentiment analysis by leveraging multimodal data such as texts, images, and acoustic information, which aim to improve traditional textual sentiment analysis. Inspired by the attention work in [11], You et al. [6] present a tree-structured LSTM (T-LSTM) model for visual-textual sentiment analysis, which treats visual and textual information jointly in a structural fashion. You et al. [12] propose a cross-modality consistent regression (CCR) model, which combines both the state-of-the-art visual and textual sentiment analysis techniques. Chen et al. [13] also propose to combine visual and textual deep neural networks with feature concatenation for sentiment analysis.

Though large progress has been made in visual-textual sentiment analysis recently, most of the works [14, 15, 6] are limited on social media data such as Twitter and Flickr, where images are usually in good quality. In this paper, as the first contribution, we introduce a new practical dataset, termed as Product Reviews-150K (PR-150K), which is collected from the product reviews of online shopping websites. PR-150K consists of three categories (i.e. positive, neutral, and negative) with more than 150k manually-labeled image-text pairs. Experts are asked to annotate PR-150K by carefully reading texts and watching images. PR-150K is a very challenging dataset due to that i) the 'neutral' images are very similar

Xiaojiang Peng and Jin Ye are equally-contributed authors. Xiaojiang Peng is the corresponding author. Email: xj.peng@siat.ac.cn

Fig. 1: Image-text pairs from PR-150K. Left: positive samples. Middle: neutral samples. Right: negative samples.

with 'positive' and ii) images are usually in low quality since they are captured casually by customers. Several previous works [8, 16] also conduct sentiment analysis in product reviews, but they are limited in textual information.

To deal with the visual-textual sentiment analysis, we propose a deep Tucker fusion method as another contribution in this paper, which efficiently combines visual and textual deep representations based on the Tucker decomposition [17] and a bilinear pooling operation. Bilinear pooling on visual and textual features helps to learn high level associations between textual and visual information [18, 19], however, it suffers from huge dimensionality issues. To this end, our deep Tucker fusion method factorises the pooling/fusion tensor using a Tucker decomposition module, which consists of several fully-connected layers and a smaller fusion tensor, to efficiently parametrize bilinear interactions.

We provide several baselines for the proposed PR-150K dataset including several state-of-the-art image-based methods, text-based methods, and fusion methods. We extensively evaluate the deep Tucker fusion method on both PR-150K and the public MVSO (collected from Flickr) dataset. We achieve state-of-the-art performance on MVSO and VSO, and also show that the proposed PR-150K is more challenging than these datasets collected from Flickr.

2. OUR DATASET AND APPROACH

In this section, we first introduce our PR-150K dataset in details, and then present our deep Tucker fusion method for visual-textual sentiment analysis.

2.1. The PR-150K Dataset

Considering the practical application of product recommendation, we collect the Product Reviews-150K (PR-150K) dataset by crawling images along with the texts in product reviews from several online shopping websites including VIPShop[1], Taobao[2], etc. The product reviews are mainly selected according to the product categories of these websites.In total, PR-150K contains 90 product categories with 151,158 image-text pairs. All the image-text pairs of PR-150K are annotated by several experts with three sentiment classes, i.e. positive, neutral, and negative, within two months. Some image-text examples are illustrated in Figure 1, where the Chinese review texts are translated by Youdao API[3]. From the visual information, we observe that only negative samples have some characteristics (such as broken, dirty, etc) to distinguish from the other two classes. PR-150K is imbalanced with ratio 7:1.5:1.5 in positive, neutral, and negative. Most of the samples are positive which may be explained by that i) people are more likely to write reviews for satisfied products, and ii) sellers may encourage and reward customers who upload positive reviews. In addition, most of the reviews have number of words ranging from 15 to 35.

Protocols. We randomly split the data in each sentiment class to 8:2 for training and testing. Since PR-150K is imbalanced, we compute *precision*, *recall* for each category, and report the average *F1 score* for our evaluation metrics.

2.2. Our Visual-Textual Sentiment Analysis System and Deep Tucker Fusion

Our framework of visual-textual sentiment analysis is shown in Figure 2. It mainly consists of four modules, i.e. deep visual representation module, deep textual representation module, Tucker fusion module, and the final classifier.

Deep Tucker Fusion (DTF). We propose the Deep Tucker Fusion (DTF) method to efficiently combine visual and textual information for sentiment analysis. DTF is composed of a Tucker decomposition operation and a bilinear pooling operation. Bilinear pooling [20] or second-order pooling [21] is first introduced in image representations, and is further extent to Visual Question Answering (VQA) for jointly visual-textual representation learning [18, 19]. The main issue with these bilinear pooling methods is related to the number of parameters, which quickly becomes intractable with respect to the input and output dimensions.

Let $x_v \in R^{d_v}$ and $x_t \in R^{d_t}$ denote the visual and textual features, respectively. With a learned bilinear model T, we obtain the final fusion representation $x_o \in R^{d_o}$ as follows,

$$\boldsymbol{x_o} = (\boldsymbol{T} \times_1 \boldsymbol{x_v}) \times_2 \boldsymbol{x_t} \quad (1)$$

[1] www.vipshop.com
[2] www.taobao.com
[3] https://github.com/chenjiandongx/youdao-wd

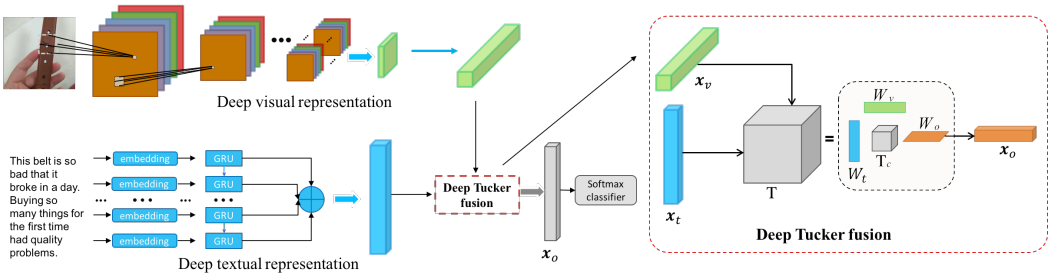

Fig. 2: Our visual-textual sentiment analysis system.

where $\boldsymbol{T}^{d_v \times d_t \times d_o}$ is the fusion tensor, and $\times_i$ represent the $i-mode$ product between a tensor and a matrix.

With ResNet (pool5) for image and GRU for text, we usually get $d_v = d_t = 2048$. Consequently, the full tensor $\boldsymbol{T}$ becomes about 10^{10} when the output dimension d_o is 1024, which is not practical due to limited memory and computation resources. To address the problem, we propose to apply Tucker decomposition for the full tensor $\boldsymbol{T}$. The Tucker decomposition[17] decomposes the full tensor $\boldsymbol{T}^{d_v \times d_t \times d_o}$ into a set of matrices $\boldsymbol{W_v}, \boldsymbol{W_t}, \boldsymbol{W_o}$ and a small core tensor $\boldsymbol{T}_c \in R^{d_1 \times d_2 \times d_3}$:

$$\boldsymbol{T} = ((\boldsymbol{T_c} \times_1 \boldsymbol{W_v}) \times_2 \boldsymbol{W_t}) \times_3 \boldsymbol{W_o}, \tag{2}$$

where $\boldsymbol{W_v} \in R^{d_v \times d_1}$, $\boldsymbol{W_t} \in R^{d_t \times d_2}$, and $\boldsymbol{W_o} \in R^{d_o \times d_3}$. Then we can rewrite Eq. (1) as follows,

$$\boldsymbol{x_o} = ((\boldsymbol{T_c} \times_1 (\boldsymbol{x_v W_v})) \times_2 (\boldsymbol{x_t W_t})) \times_3 \boldsymbol{W_o}. \tag{3}$$

In practice, $\boldsymbol{W_v}, \boldsymbol{W_t}$, and $\boldsymbol{W_o}$ are identically fully-connected layers. As for $\boldsymbol{T}_c$, we factorise the last dimension, and simply use d_3 fully-connected layers with a summation operation.

3. EXPERIMENTS

In this section, we first review two visual-textual public datasets: MVSO [15] and VSO [14], and then make extensive evaluations with our deep Tucker fusion method on our PR-150K and MVSO, and finally we validate our method on MVSO and VSO with a comparison to the state of the arts.

3.1. Public Dataset

MVSO. The MVSO dataset consists of 15,600 concepts in 12 different languages. These concept are defined as adjective and noun pairs (ANPs), which are crawled from Flickr. Each ANP has hundreds of images and has a score to assess the sentiment tendency. In this paper, we follow the protocol in [13], which only use the English dataset for evaluation.

VSO. The VSO dataset contains millions of images collected by querying Flickr with thousands of ANPs. The sentiment label of each image is decided by sentiment polarity of the corresponding ANP. We also use the same protocol with [13] for evaluation.

3.2. Implementation Details

We implement our method based on Pytorch. For visual models, we initialize the learning rate ($\boldsymbol{lr}$) to 10^{-3}, and divide it by 10 after 5, 8 and 10 epochs. We stop training after 14 epochs. We use the SGD method for optimization with a momentum of 0.9 and a weight decay of 10^{-3}. In training phase, images are resized to 224×224 with random flipping. We use ImageNet-pretrained models and finetune them on target datasets. We test the visual model on resized 224×224 images. For textual models, words are first embedded with the 'nn.embedding' layer in Pytorch, and then processed by GRU [22] or CNN [23]. For simplicity, we fix input text length for training as 26 by default. We set the $\boldsymbol{lr}$ to 2×10^{-4}, and use Adam [24] to optimize our model. For visual-textural fusion models, we first train both visual and textual models separately, and then train the fusion module with the same training settings as textual model training. Following the common setting in VQA task of [19], the dimensions (d_1, d_2, d_3) for Tucker decomposition in Eq. (2) are set to $(2 \times d, d, d)$ with $d = 310$ by default.

3.3. Deep Tucker Fusion

We evaluate our deep Tucker fusion method in Figure 3 with comparison to several popular fusion strategies, namely feature concatenation (concat), summation (sum), elementwise multiplication (mul), score summation (late). The visual and textual backbone networks are ResNet-101 and GRU-based RNN model, respectively. ResNet-101 is pretrained on Imagenet [25], and the textual model is trained from scratch.

From Figure 3, our deep Tucker fusion method outperforms all the other fusion strategies in average F1 score on both PR-150K and MVSO datasets. On PR-150K, we evaluate our method with both Chinese and English. On both languages, 'concat' method performs slightly worse than Tucker fusion. Though translation from Chinese to English is unsatisfied in sentence level, the performance are very similar

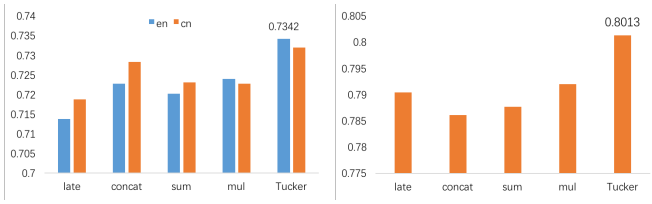

Fig. 3: Comparison between deep Tucker fusion and other popular fusion strategies. Left: PR-150K. Right: MVSO.

Table 1: Evaluation of visual and textual models on PR-150K.

	Text		Image			Tucker fusion
	CNN	GRU	ResNet101	DRN105	SENet154	SENet154+CNN(GRU)
pos.	0.863	0.845	0.891	0.889	0.891	0.888 (0.898)
neu.	0.234	0.245	0.439	0.451	0.464	0.476 (0.465)
neg.	0.760	0.760	0.754	0.756	0.776	0.816 (0.836)
avg.	0.620	0.617	0.694	0.699	0.710	0.726 (0.733)

by using either Chinese or English. Our deep Tucker fusion with English achieves the best F1 score, i.e. 0.7342. We will use English for our PR-150K dataset in the remainder of this paper. On MVSO, all the other fusion methods perform similarly and are inferior to our deep Tucker fusion method.

Visual and textual models. To select better models for both visual and textual information, we evaluate three state-of-the-art visual CNN architectures including ResNet101 [26] , SeNet154 [27], and DRN105 [28] and two popular textual models, i.e. CNN [23] and a GRU recurrent model [22].

Table 1 shows the F1 scores of each category with different visual and textural models on PR-150K. For the textual models, the CNN model performs slightly better than the GRU recurrent model. For the visual models, SENet154 outperforms ResNet101 by 1.6% and DRN105 performs on par with ResNet101. Both DRN105 and SENet154 are built upon ResNet architectures with elaborate design, while the DRN architecture emphasizes large receptive fields without pooling and the SENet architecture enhances feature maps with channel-wise attention. Having stronger models on visual modality, we select SENet154 for visual model and conduct deep Tucker fusion with both CNN and GRU textual models. The last column of Table 1 shows that these fusion results are slightly inferior to the default Tucker fusion setting where ResNet101 and GRU are used. This may be explained by that an overemphasis on visual modality (a stronger image model) degrades the complementary of both modalities. We will use ResNet101 for visual model and GRU for textual model in the remainder of this paper.

3.4. Comparison on VSO and MVSO

We further validate our deep Tucker fusion method on two popular visual-textual sentiment analysis datasets, namely VSO and MVSO. To our knowledge, there are only a handful

Table 2: Comparison on VSO

Methods	Prec.	Rec.	F1 score
visual-ResNet101	0.664	0.655	0.657
textual-GRU	0.849	0.832	0.838
Deep Fusion [13]	0.830	**0.857**	0.844
PCNN [10]	0.759	0.826	0.791
T-LSTM [6]	0.821	0.833	0.833
Ours DTF (ResNet-101+GRU)	**0.861**	0.853	**0.856**

Table 3: Comparison on MVSO

Methods	Prec.	Rec.	F1 score
visual-ResNet101	0.603	0.602	0.602
textual-GRU	0.777	0.769	0.771
Deep Fusion [13]	0.740	0.730	0.735
Ours DTF (ResNet-101+GRU)	**0.801**	**0.803**	**0.801**

of studies on these two datasets. Table 2 and Table 3 show the comparison between our method and several recent deep learning based methods on VSO and MVSO, respectively. On VSO, You et al. [6] use CNN-based image region features and a T-LSTM [29] textual attention model for robust visual-textual sentiment analysis. Deep Fusion [13] applies CNN models for both textual and visual modalities, and combines both features by concatenation. On both dataset, compared to word2vec embedding and CNN model in [13], our textual model, i.e. Pytorch's 'nn.embedding' for word embedding and GRU-based RNN, gets more promising results. Overall, our method achieves the state-of-the-art results on both datasets, which outperforms the best existing results on VSO and MVSO by 1.2% and 6.6%, respectively.

4. CONCLUSION

In this paper, we address the visual-textual sentiment analysis task with a novel deep Tucker fusion method . We introduce a new dataset, termed as Product Reviews-150K (PR-150K), for visual-textual sentiment analysis. We conduct extensive experiments on our PR-150K, MVSO and VSO. Experimental results show that our deep Tucker fusion method outperforms most of popular fusion methods and achieves state-of-the-art performance on both VSO and MVSO.

5. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (U1813218, U1613211, U1713208), Shenzhen Research Program (JCYJ20170818164704758, JSGG20180507182100698).

6. REFERENCES

[1] Bo Pang, Lillian Lee, et al., "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[2] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts, "Learning word vectors for sentiment analysis," in *ACL*, 2011.

[3] Johan Bollen, Huina Mao, and Alberto Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena.," in *ICWSM*, 2011.

[4] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu, "Unsupervised sentiment analysis with emotional signals," in *WWW*, 2013.

[5] Yafeng Ren, Yue Zhang, Meishan Zhang, and Donghong Ji, "Context-sensitive twitter sentiment classification using neural network.," in *AAAI*, 2016.

[6] Quanzeng You, Liangliang Cao, Hailin Jin, and Jiebo Luo, "Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks," in *ACM MM*, 2016.

[7] Quoc Le and Tomas Mikolov, "Distributed representations of sentences and documents," in *ICML*, 2014.

[8] Xing Fang and Justin Zhan, "Sentiment analysis using product review data," *Journal of Big Data*, 2015.

[9] Jingwen Wang, Jianlong Fu, Yong Xu, and Tao Mei, "Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks," in *IJCAI*, 2016.

[10] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks.," in *AAAI*, 2015, pp. 381–388.

[11] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.

[12] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang, "Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia," in *International conference on Web search and data mining*. ACM, 2016.

[13] Xingyue Chen, Yunhong Wang, and Qingjie Liu, "Visual and textual sentiment analysis using deep fusion convolutional neural networks," in *ICIP*, 2017.

[14] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *ACM MM*, 2013.

[15] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang, "Visual affect around the world: A large-scale multilingual visual sentiment ontology," in *ACM MM*, 2015.

[16] Mary McGlohon, Natalie Glance, and Zach Reiter, "Star quality: Aggregating reviews to rank products and merchants," in *ICWSM*, 2010.

[17] Ledyard R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, 1966.

[18] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[19] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome, "Mutan: Multimodal tucker fusion for visual question answering," in *CVPR*, 2018.

[20] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji, "Bilinear cnn models for fine-grained visual recognition," in *ICCV*, 2015.

[21] Joao Carreira, Rui Caseiro, Jorge Batista, and Cristian Sminchisescu, "Semantic segmentation with second-order pooling," in *ECCV*, 2012.

[22] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler, "Skip-thought vectors," in *NIPS*, 2015.

[23] Yoon Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[24] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[27] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[28] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser, "Dilated residual networks," in *CVPR*, 2017.

[29] Kai Sheng Tai, Richard Socher, and Christopher D Manning, "Improved semantic representations from tree-structured long short-term memory networks," *arXiv preprint arXiv:1503.00075*, 2015.