An Image-Text Sentiment Analysis Method for Small Samples Based on Image Captioning

Shuailin Chen *

Department of Software Engineering, East China Normal University, Shanghai, China

* Corresponding author: 10215101413@stu.ecnu.edu.cn

Abstract. With the wide popularity of personal terminals, people prefer social media to share their lives, which provides a rich source for sentiment analysis methods. However, challenges still exist in small-sample sentiment analysis methods. A sentiment analysis method for Small Samples based on Image Caotion and BERT is proposed. Specifically, the model takes a pre-trained language model as the image description decoder and uses a cross-modal attention mechanism to eliminate the effects of misaligned regions. This can further increase the interaction from image to text. Then, the generated descriptions are coupled with the original text in the dataset. The BERT model is used to extract word vectors and output sentiment analysis results. The COCO dataset is used to train the model for image Captioning, and the MVSA dataset is used for training and evaluation of sentiment analysis. The experiment creates Less Sample Segmentation by randomly selecting samples from the dataset. Accuracy and F1 value are used to compare with baseline models to evaluate the model performance. The results show that the Image Captioning-BERT model has a certain performance improvement in sentiment analysis of image-text pairs with small samples.

Keywords: Image Captioning, BERT, pre-trained language model, Sentiment Analysis.

1. Introduction

In the era of rapid Internet development, social media platforms have been the main avenue for users to express their emotions and views. The multimodal data published by users provides a rich source of information for sentiment analysis. Image-text sentiment analysis has been receiving increasing research attention in recent years as an important branch of multimodal sentiment analysis. A number of sentiment analysis methods have been proposed, and some models mention using a concatenation of feature vectors from different modalities to fuse multimodal features. Xu and Mao proposed MultiSentiNet, which extracts semantic visual information, scenes and objects from images. The features of text and images are merged by using visual features as the LSTM output of attention-guided text [1]. The Hierarchical Attention Network (HSAN) was proposed in subsequent research by Xu, which simultaneously processes text features by including semantic annotations from images, getting a better fusion of image and text features [2]. The Multi-View Attention Network (MVAN) was proposed by Yang et al. to obtain deeper multimodal feature vectors, which use multi-view image-text interactions to strengthen the extraction and fusion ability of emotional features [3]. Yang et al. later proposed the MGNNS model that by combining graph neural networks and multi-head attention mechanisms, so as to effectively capture and merge the features of image and text [4]. On the other hand, Tan and Bansal proposed the LXMERT model that uses a cross-modal encoder based on the Transformer architecture. LXMERT is effective in merging image and text information through pretraining for a lot of downstream tasks [5]. In order to improve the performance of small samples, Gao et al. proposed a method by improving the pre-trained language model, which provides new ways and ideas for multimodal sentiment analysis [6]. Zhu et al. proposed a new Image-Text Interaction Network (ITIN) which introduces a new module to find region-word correspondences [7]. ITIN also uses adaptive gating units to fuse features across modalities and these changes helped to improve the performance.

However, current multimodal sentiment analysis methods still have some challenges, such as limited data leading to insufficient model training to adequately learn the complex relationships between images and texts. In addition, fusion and alignment of different modal data is more difficult

in small sample contexts, where traditional methods perform ineffectively. An improved method based on image captioning and BERT model is proposed for enhancing small-sample image-text sentiment analysis in this paper. The model uses a pre-trained language model for image captioning decoder and a cross-modal alignment and gating mechanism in order to improve the model's adaptability on less sample segmentation and the accuracy of sentiment analysis.

2. Methodology

Image Captioning-BERT model is proposed to improve the problem of small sample image-text sentiment analysis. The model involves two parts, one part of Transformer for image captioning and the other part of BERT model for text sentiment analysis.

2.1. Transformer for Image Captioning

The Transformer has become the standard paradigm for handling image description generation, with its core being the self-attention mechanism [8]. It assumes that the input information is X, and the query vector q as well as the key-value pairs (K, V) are generated from X, as shown below:

$$\mathrm{Q} = \mathrm{XW_Q} \tag{1}$$

$$\mathrm{K} = \mathrm{XW_K} \tag{2}$$

$$\mathrm{V} = \mathrm{XW_V} \tag{3}$$

Where, $\mathrm{W_Q}$, $\mathrm{W_K}$, $\mathrm{W_V}$are trainable parameters, and D is the scaling factor, $\mathrm{X} = [\mathrm{x_1}, \mathrm{x_2}, \ldots, \mathrm{x_n}]$. The scaled dot-product model is used as the aattention-scoringformula. When self-attention is computed between all input vectors x_i, the formula is as shown below:

$$\mathrm{Self-Attention((K,V),q)} = \mathrm{Softmax}(\frac{\mathrm{Q \cdot K^T}}{\sqrt{\mathrm{d_k}}}) \cdot \mathrm{V} \tag{4}$$

The multi-head attention mechanism consists of multiple self-attention mechanisms. In the task of image description generation, let the output of the visual encoder be I, and the current state of the encoder be H. The cross-attention mechanism is represented as:

$$\mathrm{EncDecAttention(H,I)} = \mathrm{Self-Attention(IW_K, IW_V, HW_Q)} = \mathrm{Softmax}(\frac{\mathrm{HW_Q \cdot IW_K{}^T}}{\sqrt{\mathrm{d_k}}}) \cdot \mathrm{IW_V} \tag{5}$$

The subsequent Add & Norm layer contains residual linking and layer normalisation, which are the expressions:

$$\mathrm{LayerNorm(X + MutltiHeadAttention(X))} \tag{6}$$

$$\mathrm{LayerNorm(X + FeedForward(X))} \tag{7}$$

X represents the input to either the attention mechanism or the feedforward neural network. The decoder uses two gates to determine when to utilize visual information and when to use linguistic information, balancing the influence of both [9]. The formulas are as follows:

$$\mathrm{Output} = \sigma(\mathrm{H})\mathbf{1}_{\sigma(\mathrm{H})>\tau} \otimes \mathrm{EncDecAten(H,I)} + (1-\sigma(\mathrm{H}))\mathbf{1}_{1-\sigma(\mathrm{H})>\tau} \otimes \mathrm{H} \tag{8}$$

Where, the function $\sigma()$ converts each element's value into a probability between 0 and 1. The indicator function $\mathbf{1}()$ is used to introduce sparsity by setting the output of the function to zero when it is less than the threshold τ.

By coupling the decoder's current state with the visual encoder's output through a probability transformation operation, when one part of the output is zero, its gradient is also zero. This prevents the accidental overwriting of linguistic knowledge, while the gradient of the other part remains usable. Effective protection is provided for the knowledge in the pre-trained model. It was able to perform well with a small amount of data.

2.2. BERT for Text Sentiment Analysis

BERT is a pre-trained representation model [9], consisting of an input layer, stacked Transformer encoder units, and an output layer. The followed Fig 1 is the BERT's basic structure.

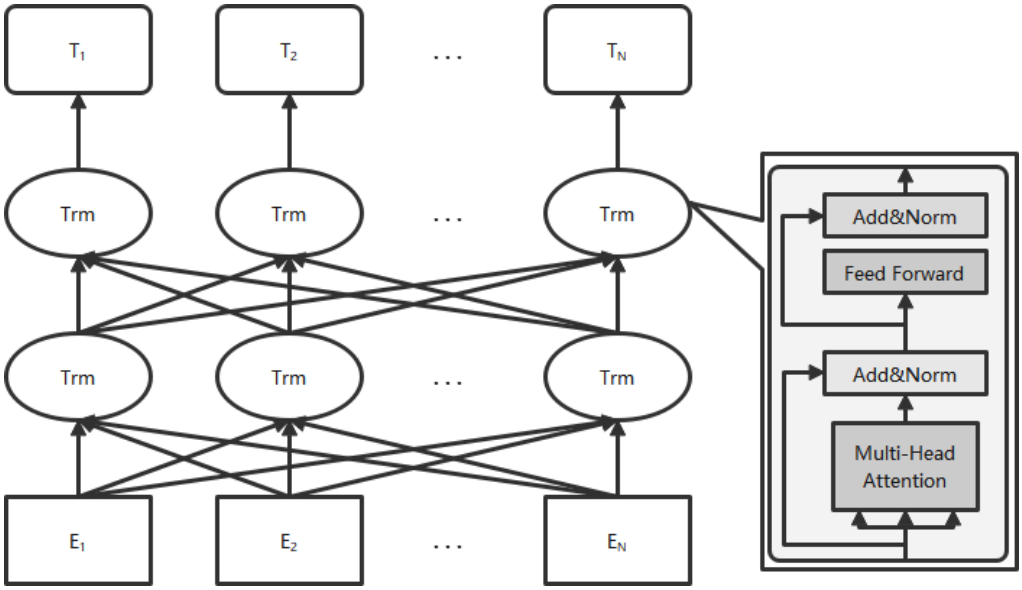

Figure 1. Model structure of BERT

Downstream task-based model fine-tuning is the most common learning method for pre-trained models such as BERT, where a specific network structure is usually added after the Transformer encoder to match different task outputs. Given the pre-trained language model M, the input X is first converted into a token sequence $X' = [CLS]X[SEP]$, which is then mapped to a vector sequence $\{h_i \in \mathbb{R}^d\}$ through the input layer of M. After feeding the vector sequence into the model backbone, M and the added Softmax classifiers are trained by maximizing the number of pairs of correct labels, i.e:

$$\hat{y} = \arg\max_{y \in Y} p(y|X; \theta, \theta') = \arg\max_{y \in Y} \text{Sofrmax}(Wh_{[CLS]} + b) \tag{9}$$

Where θ denotes the parameters in M that have been trained with and θ' denotes the parameters in the added classifier. $W, b \in \theta'$ is a randomly initialized parameter in the classifier. $h_{[CLS]}$ is the output vector of token "[CLS]" after passing through multiple layers of the Transformer encoder, and it represents the entire input sentence for classification, resulting in the predicted label $\hat{y} \in Y$. This process is illustrated in Fig 2.

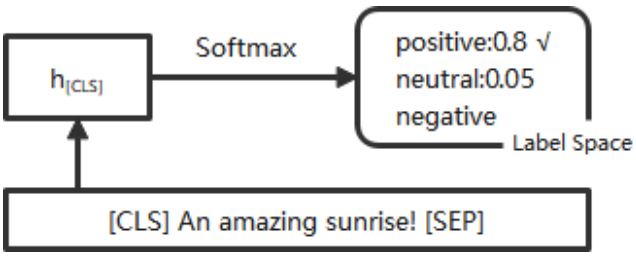

Figure 2. A fine-tuned approach to BERT in an emotion categorization task

2.3. Image Captioning-BERT Model

The model first receives the input image to the image encoder and extracts the objects in the image and their spatial locations. A pre-trained GPT-2 is used as an object detection network to extract features. Visual features and location information about the object are represented in the form of embedding vectors extracted from it. Encoder consists of N layers of Transformer. The left of Fig 3 illustrates the decoder utilized in image captioning. The visual features derived from the encoder's

output and the decoder's current state are conveyed through the cross-attention mechanism and subjected to processing in a backward iteration.

In the sentiment analysis phase, the algorithm loads the model weights trained during the image description generation phase. As shown on the right side of Fig 3, for each image from the MVSA dataset, the algorithm uses the trained encoder to extract visual features and generates a textual description of the image through the decoder. The BERT model tokenizes and processes the generated description, which is then combined with the original sentiment text to obtain the sentiment information of the input text. The combined text is processed by the BERT model through its weights, outputs the sentiment information of the text, and computes the sentiment label.

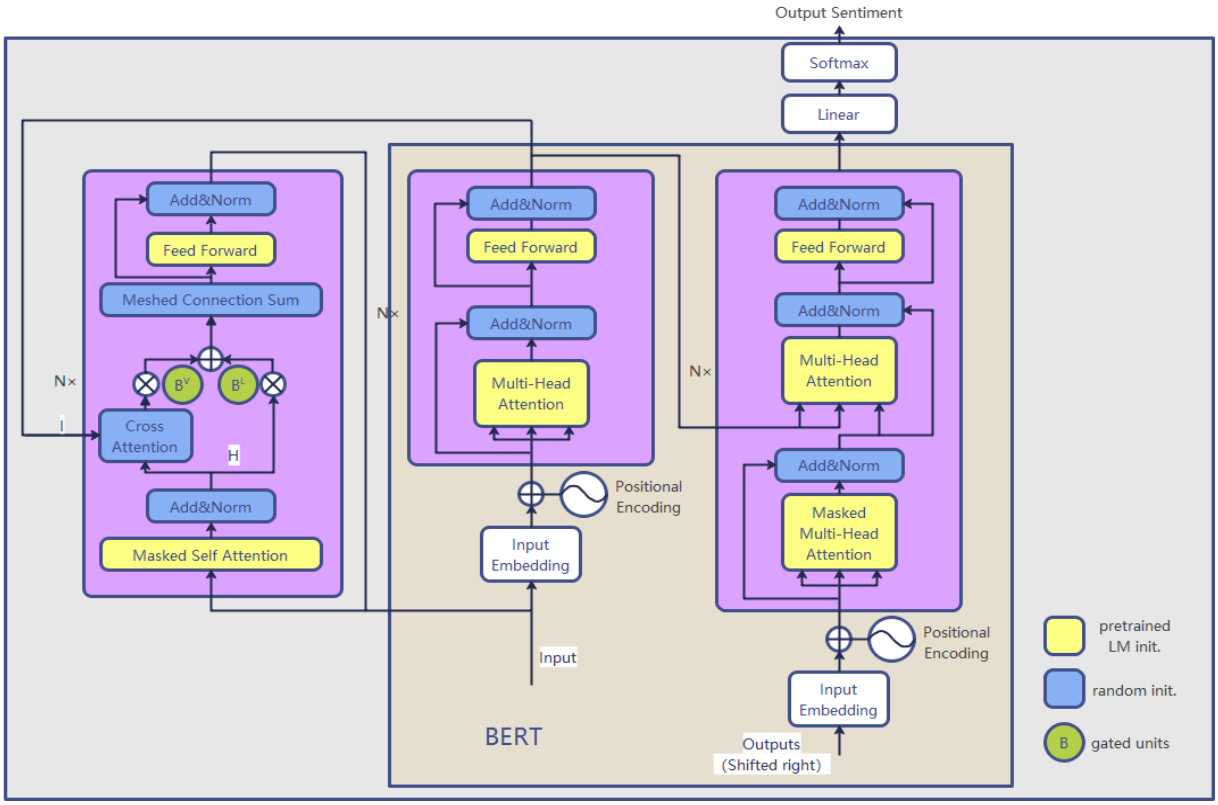

Figure 3. Image Captioning-BERT model structure

3. Evaluation Results and Analyses

3.1. Datasets

In the stage of generating image descriptions, the experiments first train and evaluate the model for image captioning with the MS COCO dataset [10]. Each image in the MS COCO dataset has 5 different caption annotations, with a total of 123,287 images. The experiment used the Karpathy split method for validation and test set division. To create a small-scale training dataset, the study randomly selected 0.1% of the image-description pairs from the MS COCO dataset for training, consisting of 567 pairs, and the training loop was executed 100 times.

The MVSA dataset was used for the training and validation of the sentiment analysis phase, which is divided into MVSA-Single and MVSA-Multiple, both predicted to be from the Twitter platform [11]. These two datasets were divided into training, validation, and test sets. The Less Sample Segmentation consisted of 1% of the full training and experimental sets and was chosen randomly from the divided training set to observe the performance under small-sample training. Table 1 displays the statistics of the dataset.

Table 1. Statistical information on datasets

DataSet	Emotional labelling	Full Sample Segmentation			Less Sample Segmentation		
		Train	Dev	Test	Train	Dev	Test
MVSA-Single	Positive	2147	268	268	20	20	268
	Neutral	376	47	47	4	4	47
	Negative	1088	135	135	12	12	135
MVSA-Multiple	Positive	9056	1131	1131	88	88	1131
	Neutral	3528	440	440	32	32	440
	Negative	1040	129	129	10	10	129

3.2. Quantitative Results

The experiments use Accuracy and F1 values to assess method performance. Accuracy provides overall classification accuracy and is suitable for measuring model performance on all data. Whereas, the F1 value can help assess the model performance on a few classes. In order to test the validity of the models, experiments were selected to compare some multimodal sentiment analysis models with the models in this paper on the Full Sample Segmentation and Less Sample Segmentation. Table 2 and Table 3 show the results compared with the baseline model.

Table 2. Performance of different methods on the Full Sample Segmentation [12]

Method	MVSA-Single		MVSA-Multi	
	ACC	F1	ACC	F1
MultiSentiNet	69.84	69.63	68.86	68.11
MGNNS	73.77	72.7	72.49	69.34
Co-MN-Hop6	70.51	70.01	68.92	68.83
HSAN	69.88	66.90	67.96	67.76
ICSA Model	70.32	68.26	68.84	68.65

Table 3. Performance of different methods on the Less Sample Segmentation

Method	MVSA-Single		MVSA-Multi	
	ACC	F1	ACC	F1
MVAN	42.77	36.75	46.16	34.74
MGNNS	34.40	32.05	40.03	32.58
LXMERT	47.27	37.54	49.08	36.69
LM-BFF	57.35	51.39	54.60	45.22
ICSA Model	57.37	52.21	56.76	44.07

The results demonstrate that different modeling approaches perform differently for different amounts of data. This paper's proposed model for full sample segmentation is better in terms of accuracy and F1 value than MultiSentiNet and HSAN [1][2]. Although all these models obtain semantic descriptions from images because HSAN uses CNN to extract visual feature vectors of images and the visual query vectors are randomly initialised. MultiSentiNet may be fused in a simpler way and lacks in-depth modeling of complex inter-modal relationships. Image Captioning-BERT model extracts features using the pre-trained GPT-2, which is integrated with the text generation model through a cross-modal attention mechanism, so it has better performance.

However, in the case of full sample segmentation, the structure of the graph neural network allows MGNNS to better deal with noise and learn richer feature representations [4]. Therefore, it has a better performance than Image Captioning-BERT model. But in case of less sample segmentation, the Image Captioning-BERT model instead has better performance than the other baseline models.

Both MVAN and MGNNS, rely on graphical and textual interactions to enhance sentiment analysis [3] [4]. Therefore, in case of less sample segmentation, there is insufficient correlation information between modalities, making it difficult for the models to learn effective cross-modal feature interactions. The model is unable to learn sufficiently and instead performs worst in the baseline model. LXMERT uses a vision-language pre-trained model; however, its performance is mediocre in small-sample scenarios [5]. This may be due to the significant gap between the multimodal pretraining tasks and the downstream sentiment classification task, which affects its performance. Image Captioning-BERT model retains the pre-trained knowledge using a special gating mechanism to efficiently generate meaningful image descriptions. This dissolves the gap between image and text and hence gives better performance with small number of samples. LM-BFF performs well in small-sample cases, but due to its use of the exemplar-based fine-tuning strategy, it requires handling a large number of input combinations and conducting multiple training iterations on these combinations to optimize performance, which consumes a significant number of computational resources [6]. The model in this paper introduces sparsity during computation, which means that only a part of the neurons is activated, reducing the total amount of computation needed in the propagation process. Therefore, the model can still achieve excellent performance in the case of limited computational resources.

3.3. Discussion

This study tests the graphical sentiment analysis model by testing it on COCO dataset and MVSA dataset. The experiments show that using image captioning technique can significantly enhance the model's performance when handling a small amount of data. The Image Captioning-BERT model has significantly better accuracy and F1 value compared to the traditional model, as indicated by the results. In particular, it shows higher efficiency and better adaptability in the case of data scarcity.

This finding has important implications for practical applications in the field of graphical sentiment analysis. In brand management, graphic sentiment analysis can effectively monitor brand image and customer satisfaction. By analyzing pictures and comments posted by customers on social media, companies can keep abreast of consumers' real feedback and emotional tendencies. This not only identifies the brand's position in the market, but also potential crises and opportunities, helping enterprises to quickly adjust their market strategies and thus enhance market competitiveness.

In addition, depression and anxiety issues among adolescents are becoming increasingly severe. Traditional mental health monitoring methods rely on periodic questionnaires and face-to-face consultations, which consume more resources and may lead to data delays and subjective bias. By utilizing trained image-text sentiment analysis models, interactions on social media can be automatically analyzed to detect emotional fluctuations and mental health conditions in real time, providing support for early intervention and treatment. Compared to some classic image-text interaction neural network models, it demonstrates significant advantages, especially in small-sample scenarios.

4. Conclusion

The widespread use of social media has provided a rich corpus for image-text sentiment analysis, and research in this area has become an important method for understanding users' emotional fluctuations. An image-text sentiment analysis method based on image Captioning is proposed in this paper. The Image Captioning-BERT model's adaptability and accuracy on public databases are demonstrated by the experimental results. This is primarily due to the utilisation of a pre-trained language model, in conjunction with a bespoke gating mechanism, which serves to bridge the

semantic gap between disparate modalities. Additionally, the image captioning technology provides richer and more accurate semantics for sentiment analysis. This method can be used to monitor data from social media platforms with less computational expense, providing assistance in areas such as early prevention of youth mental health and corporate brand management. However, the study still has limitations, such as difficulty handling noise caused by insufficient useful information in the images or low relevance between images and text, which reduces the model's stability in small-sample scenarios. In the future, the model will be optimised on more domain-specific datasets to improve its migration capabilities and explore cross-domain multimodal sentiment classification methods in sample less scenarios.

References

[1] Xu N, Mao W. Multisentinet: A deep semantic network for multimodal sentiment analys. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017: 2399 - 2402.

[2] Xu N. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. 2017 IEEE international conference on intelligence and security informatics (ISI). IEEE, 2017: 152 - 154.

[3] Yang X, Feng S, Wang D, et al. Image-text multimodal emotion classification via multi-view attentional network. IEEE Transactions on Multimedia, 2020, 23: 4014 - 4026.

[4] Yang X, Feng S, Zhang Y, et al. Multimodal sentiment detection based on multi-channel graph neural networks. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 328 - 339.

[5] Tan H, Bansal M. Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv, 2019.

[6] Gao T, Fisch A, Chen D. Making pre-trained language models better few-shot learners. arXiv preprint, 2020.

[7] Zhu T, Li L, Yang J, et al. Multimodal sentiment analysis with image-text interaction network. IEEE transactions on multimedia, 2022, 25: 3375 - 3385.

[8] Vaswani A. Attention is all you need. Advances in Neural Information Processing Systems, 2017.

[9] Chen, Li, et al. DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks. arXiv preprint arXiv:1410.8586, 2014.

[10] Lin T, Maire M, Belongie S J, et al. Microsoft COCO: common objects in context. Computer Vision-ECCV 2014-13th European Conference (ECCV). 2014: 740 - 755.

[11] Niu T, Zhu S, Pang L, et al. Sentiment analysis on multi-view social data. MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22. Springer International Publishing, 2016: 15 - 27.

[12] Xu N, Mao W, Chen G. A co-memory network for multimodal sentiment analysis. The 41st international ACM SIGIR conference on research & development in information retrieval. 2018: 929 - 932.