ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 125-131

Article Title: **Yolo Algorithm Based Multimodal Sentiment Analysis on Image and Text Data**

Yolo Algorithm Based Multimodal Sentiment Analysis on Image and Text Data

[*1]J. Sindhuja A, [2]C. Brintha

[1,2]Department of computer science and Engineering, Udaya School of Engineering, Udaya Nagar, Kanyakumari, Vellamodi, Tamil Nadu, India.
sindhujanew2810@gmail.com

ABSTRACT

Sentiment analysis's objective is to examine public sentiment in a way that will support corporate growth. It emphasizes emotions as well as polarity (positive, negative, and neutral). It makes use of a variety of Natural Language Processing techniques, including Automatic, Hybrid, and Rule-based. Users are becoming accustomed to uploading text and photographs on social networks to express their feelings or ideas. As a result, multimodal sentiment analysis has drawn more attention as a research area in recent years. Usually, an image has emotional areas that trigger human emotion, which are typically expressed by corresponding words in comments. Similarly, while writing visual descriptions, people frequently depict the emotive areas of an image. As a result, for multimodal sentiment analysis, the association between picture affective areas and the accompanying text is extremely important. This paper exhibits one of the best CNN representatives You Only Look Once (YOLO), which breaks through the CNN family's tradition and innovates a completely new way of solving object detection with most simple and highly efficient way. Its name derives from the fact that, unlike earlier object detector algorithms like Faster R-CNN, it only requires that an image or video travel once through its network. Its outcomes surpassed the performance of Faster R-CNN greatly. The performance of YOLO is compared with faster R-CNN in terms of accuracy and F1 Measure.

Keywords: sentiment analysis, cross modal alignment module, cross modal gating module, softmax classifier.

1 Introduction

Sentiment analysis (also known as opinion mining or emotion AI) is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information [1]. Sentiment analysis is widely applied to the voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine. With the rise of deep language models, such as RoBERTa, more difficult data domains can also be analyzed, e.g., news texts where authors typically express their opinion/sentiment less explicitly.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 125-131

Article Title: Yolo Algorithm Based Multimodal Sentiment Analysis on Image and Text Data

Sentiment analysis is the process of classifying whether a block of text is positive, negative, or neutral [2]. Sentiment analysis is contextual mining of words which indicates the social sentiment of a brand and helps the business to determine whether the product which they are manufacturing is going to make a demand in the market or not. It uses various Natural Language Processing algorithms such as Rule-based, Automatic, and Hybrid.

2 Related Work

The pre-trained VGG16 network was used to extract visual features and fine-tune on the MVSA-Multiple and T4SA datasets for image sentiment classification. The Mask-RCNN model was then exploited to determine the objects in the images and were converted into text [3]. This paper was proposed Multimodal Translation for Sentiment Analysis (MTSA), a multimodal framework that helped to improve the quality of visual and audio features by translating them to text features extracted by Bidirectional Encoder Representations from Transformers (BERT). Experiments on two benchmark datasets CMU-MOSI and CMU-MOSEI were conducted for the better understanding of the suggested method than the state-of-the-art methods on both datasets across all the metrics [4].

3 Proposed Methodology

This work presents an image text interaction network for multimodal sentiment analysis, which focuses on the alignment between image regions and text words and integrates both visual and textual context information. The image-text interaction network comprises a cross-modal alignment module and a cross-modal gating module.

3.1 Object Detection

For computers, however, detecting objects is a task that needs a complex solution. For a computer to “detect objects” means to process an input image (or a single frame from a video) and respond with information about objects on the image and their position.

Other, slower algorithms for object detection (like Faster R-CNN) typically use a two-stage approach:

In the first stage, interesting image regions are selected. These are the parts of an image that might contain any objects.

- In the second stage, each of these regions is classified using a convolutional neural net.

Cross-modal Alignment Module

For the input image I, we detect image regions and their associated representations utilizing YOLO which explained later in this paper.

For an input sentence T with n words, BERT-Base method applied to embed each word into a 768-dimensional embedding vector x_i, $i \in [1, n]$. Then, a bidirectional GRU was employed to summarize context information in the sentence.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 125-131

Article Title: Yolo Algorithm Based Multimodal Sentiment Analysis on Image and Text Data

$$\vec{h}_i = \vec{GRU}\left(x_i, \vec{h}_{i-1}\right), i \in [1, n] \quad (1)$$

$$\overleftarrow{h}_i = \overleftarrow{GRU}\left(x_i, \overleftarrow{h}_{i+1}\right), i \in [1, n] \quad (2)$$

In the above equations $\vec{h}_i \in R^d$ denotes the forward hidden state and $\overleftarrow{h}_i \in R^d$ denotes the backward hidden state. The final word feature is the average of bidirectional hidden states and is given below,

$$w_i = \frac{\vec{h}_i + \overleftarrow{h}_i}{2} i \in [1, n] \quad (3)$$

The region-word affinity matrix is first computed as,

$$A = \left(\hat{W}_r R\right)\left(\hat{W}_t W\right)^T \quad (4)$$

Where W_r *and* W_t represent the projection matrices to obtain k dimensional region and word features. For the region word affinity matrix $A \in R^{m \times n}$, A_{ij} denotes the affinity between the i^{th} region and j^{th} word.

The affinity matrix A as follows.

$$\bar{A} = soft\max\left(\frac{A}{\sqrt{k}}\right) \quad (5)$$

Then, we aggregate all word features about each region on the basis of normalized matrix $\bar{A}$:

$$U = \bar{A}.W \quad (6)$$

Where the i^{th} row of U denotes the interactive textual features corresponding to the i^{th} region. Therefore, U can be used to explore the interaction of information flowing between images and words.

3.2 Cross-modal Gating Module

The cross-modal alignment module generates the most corresponding word-level information for each region, and fragment messages passing across the two modalities allow fine-grained cross-modal interactions. However, in practice, not all the learned region word pairs are perfectly aligned. Therefore, we further propose a cross-modal gating module utilizing a soft gate to control feature fusion intensity adaptively, with the aim to eliminate the influence of negative region-word pairs and further enhance the interactions of cross-modality information.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 125-131

Article Title: Yolo Algorithm Based Multimodal Sentiment Analysis on Image and Text Data

3.3 Multimodal Sentiment Classification

The objective of multimodal sentiment classification is to predict the sentiment label $y \in \{Positive, Neutral, Negative\}$ of an input image-text pair (I,T) . Therefore, we feed the feature vector F into a softmax layer for predicting the final sentiment:

$$y = soft\ \max\left(W_f F + b_f\right) \tag{7}$$

Where W_f *and* b_f are learnable parameters.

$$L = -\sum_i \hat{y}_i \log y_i \tag{8}$$

Where $\hat{y}_i$ denotes the ground truth sentiment label, and y_i is the output of the softmax layer.

3.4 System Architecture

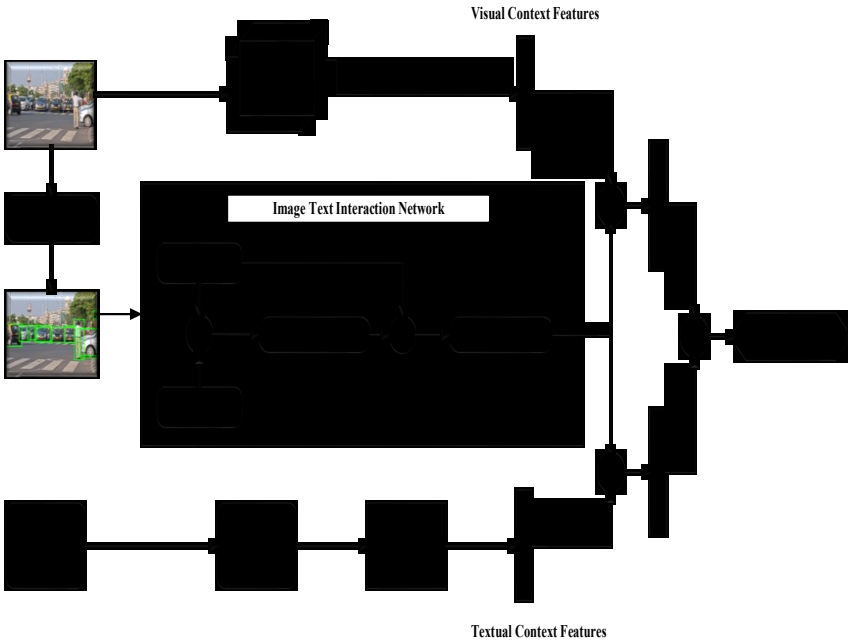

Figure 1: *Architecture of proposed system*

Despite the impressive advances in multimodal sentiment tasks, little attention has been paid in exploring cross-modal interactions for image-text sentiment analysis. Most of the existing approaches simply concatenate features extracted from different modalities or learn the relation between image and text at a coarse level, which leads to suboptimal predictions. Considering the mutual influences and intricate relationship between the two modalities, image-text interaction network (ITIN) for multimodal sentiment analysis is utilized in this paper.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 125-131

Article Title: Yolo Algorithm Based Multimodal Sentiment Analysis on Image and Text Data

3.5 Yolo Algorithm Description

You Only Look Once (YOLO) is a viral and widely used algorithm. YOLO is famous for its object detection characteristic. The core of the YOLO target detection algorithm lies in the model's small size and fast calculation speed. The structure of YOLO is straightforward. It can directly output the position and category of the bounding box through the neural network. The speed of YOLO is fast because YOLO only needs to put the picture into the network to get the final detection result, so YOLO can also realize the time detection of video. YOLO directly uses the global image for detection, which can encode the global information and reduce the error of detecting the background as the object. YOLO has a strong generalization ability because YOLO can learn highly generalized features to be transferred to other fields. It converts the problem of target detection into a regression problem, but detection accuracy needs to be improved. The original YOLO architecture consists of 24 convolution layers, followed by two fully connected layers.

4 Result and Discussion

We use the accuracy and F1-score as the evaluation metrics.

Table 1: *Comparison of different methods on MVSA datasets*

Method	MVSA Single		MVSA Multiple	
	Accuracy	**FI**	**Accuracy**	**F1**
Senti Bank & Strength [11]	0.5205	0.5008	0.6562	0.5536
CNN-Multi [12]	0.6120	0.5837	0.6639	0.6419
DNN-LR [13]	0.6142	0.6103	0.6786	0.6633
HSAN [14]	-	0.6690	-	0.6776
MultiSentiNet [15]	0.6984	0.6963	0.6886	0.6811
CoMN [16]	0.7051	0.7001	0.6992	0.6983
MVAN [17]	0.7298	0.7298	0.7236	0.7230
ITIN [18]	0.7519	0.7497	0.7352	0.7349
YOLO Based [19]	0.829	0.8	0.793	0.781

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 125-131

Article Title: Yolo Algorithm Based Multimodal Sentiment Analysis on Image and Text Data

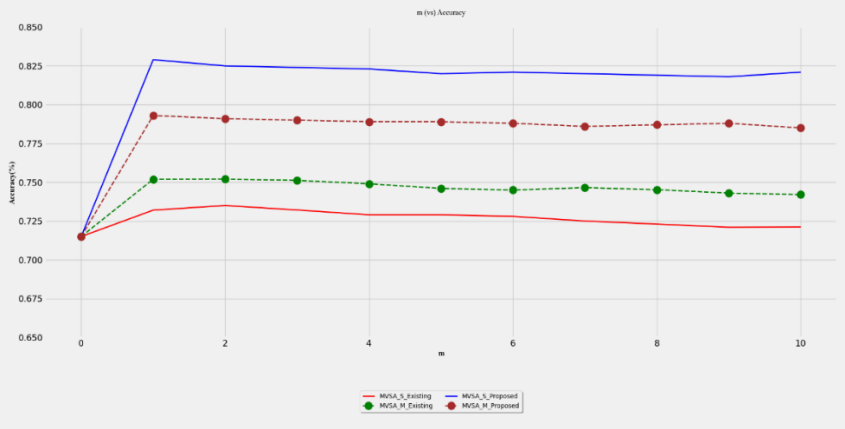

Figure 2: *Accuracy with different image region number (m)*

The performance improvements benefit from the superiority of the proposed method. First, with the cross-modal alignment module and cross-modal gating module, the interactions between image and text can be captured thoroughly at a finer level.

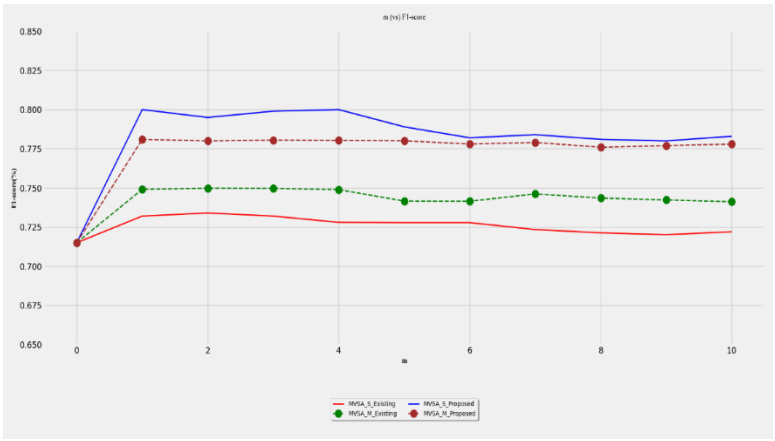

Figure 3: *F1 score with different image region number (m)*

On the MVSA-Single dataset, our model outperforms the existing best model ITIN in terms of accuracy and F1-score respectively. Overall, these results demonstrate the advantage of the proposed YOLO based method for multimodal sentiment analysis.

5 Conclusion and Future Enhancement

5.1 Conclusion

This paper exhibits one of the best CNN representatives You Only Look Once (YOLO), which breaks through the CNN family's tradition and innovates a completely new way of solving object detection with most simple and highly efficient way. Its outcomes surpassed the performance of Faster R-CNN greatly. The performance of YOLO is compared with faster R-CNN in terms of accuracy and F1 Measure. The performance of the proposed work is evaluated in the python environment, and it shows better enhancement in multimodal sentiment analysis.

5.2 Future Enhancement

- Even though YOLO outperforms faster R-CNN methodology it holds some restrictions.
- In an image it struggles to recognize small objects in a group.
- It also creates bounding errors while implementation.
- Therefore, the future work will diminish the error with new methodology to improve the accuracy or considering that most of the current multimodal methods focus on sentiment classification, the investigation of multimodal continuous emotion intensity is emerged as future work.

References

1. Medhat Walaa; Ahmed Hassan; Hoda Korashy, Year: 2014, "Sentiment analysis algorithms and applications: A survey", Ain Shams engineering journal, Vol: 5, no: 4, pp. 1093-113.

2. Fang Xing; Justin Zhan, Year: 2015, "Sentiment analysis using product review data", Journal of Big Data, Vol: 2, no: 1, pp. 1-4.

3. Ghorbanali Alireza; Mohammad Karim Sohrabi; Farzin Yaghmaee, Year: 2022, "Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks", Information Processing & Management, Vol: 59, no: 3, pp. 102929.

4. Yang Bo; Bo Shao; Lijun Wu; Xiaola Lin, Year: 2022, "Multimodal sentiment analysis with unidirectional modality translation", Neurocomputing, Vol: 467, pp. 130-7.