Image Captioning with Complementary Visual and Textual Cues

Bharathwaaj Venkatesan,
Ravinder Kaur Sond

***Abstract*—Describing an image with natural sentence without human involvement can be achieved using Deep Neural network, it requires knowledge of both image processing and Natural language processing. Most of the existing works are based on single modality model with Encoder-Decoder architecture where input images are encoded using Convolution Neural Network (CNN) and caption is generated by Recurrent Neural Network (RNN). In this paper, we propose image captioning model with complementary visual and textual cues. Our model performs early fusion by combining encoded image features from different CNNs, along with combined textual features from different word embedding techniques. The fused inputs are passed to our language model Long Short-Term Memory (LSTM) which generate captions. The result shows that our model with additional complementary information outperforms existing single modality models.**

***Index Terms*—Image captioning, deep learning, multimodal learning**

I. Introduction

Generating appropriate captions for a given image has become one of the most interdisciplinary research areas, where it predominantly combines computer vision and natural language processing. The application of image captioning is wide in range, such that some of the applications are used to help visually impaired people to understand the content of the image. In the best case, the image captioning should have the ability to express the sentiment of the visuals through natural linguistic understanding. In addition to this, Google uses the image captioning mechanism to classify photos into categories, like mountains, sea, business structure, etc. with general album tags. Another important application is visual content-based report generation, like damage estimation for insurance claim.

Image classification has improved rapidly because of the establishment of large amount of datasets, like MSCOCO [1], Flickr8k, Flickr30k and the advancement of deep neural network, for instance, CNN's [2]. Likewise, ImageNet [3] is a dataset of over 15 million images belonging to roughly 22,000 categories. Those images were collected from the web and labeled by humans using Amazon's Mechanical Turk crowdsourcing tool. The goal of image classification is to classify a picture into maximum possible categories and the most common way is to use a convolutional base which is used to perform the feature extraction from images. In the image classification task, the weights will be initialized using the artificial neural networks like VGG16, ResNet [4].

In general, visual caption generation model has two inputs, one is the features extracted from the images and the other is a vector representation of the captions. It is implemented as an Encoder-Decoder (EnDec) architecture, wherein a CNN subnetwork is used as image encoder and another subnetwork of Recurrent Neural Network (RNN) is employed as a decoder for the caption generation.

The existing works concentrate on single modality representation for images as well as their respective captions. This research work aims at the introduction of a multimodality learning approach for visual and textual representation using multiple feature extractors, viz. VGG16, ResNet50, Word2vec and GloVe.

II. Related work

The image captioning problem and its proposed solutions have existed since the emergence of the Internet and its widespread adoption to sharing of images. Several algorithms and interesting techniques have been suggested by researchers from different perspectives.

In that case, few years back in 2015, Karpathy and Fei-Fei [5] introduced a multimodal RNN for caption generation. First, they aligned sentence snippets to the target visual regions through a multimodal RNN embedding. The VGG16 network was used for image feature extraction. Even though the model outperformed and the results were encouraging, the model can only generate a description of an input at a fixed resolution because of the region level model which focuses only on a certain part of the image and does not consider other regions in the image which is certainly used for generating visual caption descriptions.

There are many limitations in traditional methods of image captioning like retrieval based and template-based methods. Convolution neural network and recurrent neural network are combined to solve such limitations. From then on, neural network-based image caption methods are used. To generate image caption related to the image content, the model extracts the information from the image and then fuses them to obtain much finer results. Unimodality models that have an attention-based approach, have the shortcoming of losing the features, and inorder to overcome this, we propose the multimodality approach.

Talking about fusion using the multimodality approach in use, there are three different methods to fuse textual and vi-

B. Venkatesan, R. Sond are with the Department of Computer Science, Lakehead University, Thunder Bay, ON, Canada. (e-mail: { bvenkat2, rsond}@lakeheadu.ca).

sual features. Those include simple operation-based, attention-based as well as tensor-based fusion methods. Vectorized features from different sources of knowledge can be combined in deep learning using a simple process, such as concatenation or weighted sum, which often has just a few or even number of parameters involved because, the joint training of the deep models will change the layers for high-level extractions of features to compensate for the process needed. Concatenation may be used to combine either low input [6], [7] characteristics or high feature derived from the pre-trained models [8], [9]. Proposed model uses the first technique that is simple operation-based fusion where the vectorized features from images are integrated using concatenation. Zhang's [10] multimodal approach uses similar type of fusion techniques, specifically attention based and bilinear pooling fusion. In the model's text extraction part, the textual features are fused using the addition operation which was the key factor for the model's improvement in its performance.

Fusion is a basic research problem in multimodal studies, which integrates information extracted from different unimodal data into one compact multimodal representation. Distinctive researches from the past, categorizes the fusion as, early fusion that are feature-level fusion which directly combines the feature extracted from unimodal data, whereas late fusion are strong intra-modality interactions [11], [12]. Mechanism of attention is commonly used for fusion, which often refers to a weighted total of a collection of vectors using dynamically generated scalar weights by a small attention model at each time phase [13], [14]. By using a two-dimensional weight matrix, the bilinear representation is converted linearly into an output vector, which is similar to three-dimensional tensor operator to combine two input function vectors [15].

Unlike current methods [16], where there is no early fusion using visual features and textual features, the proposed model has used a word embedding technique called GloVe [17], inaddition to Word2vec. It is a Global vector (GloVe) for word representation in recent times. This is one of the most recent methodologies for learning vector space representations of words used in caption generation.

The proposed approach has different stages of the process such as visual feature extraction, text interpretation and fusion training. To accomplish the task of visual captioning through multimodality approach, we propose the above mentioned method of fusion of feature vectors from the word embedding techniques as well as merging of image features from multiple artificial neural networks which overcomes the problem of losing features. Hence to achieve this goal and outperform existing approaches, the proposed model is experimented on Flickr8k dataset and competitive results are achieved.

III. Methodology

A. General Image Captioning Approach

A general image captioning approach consist of two core modules: i. visual pre-processor and feature extractor, and ii. sequence processor and interpreter as shown in Fig. 1.

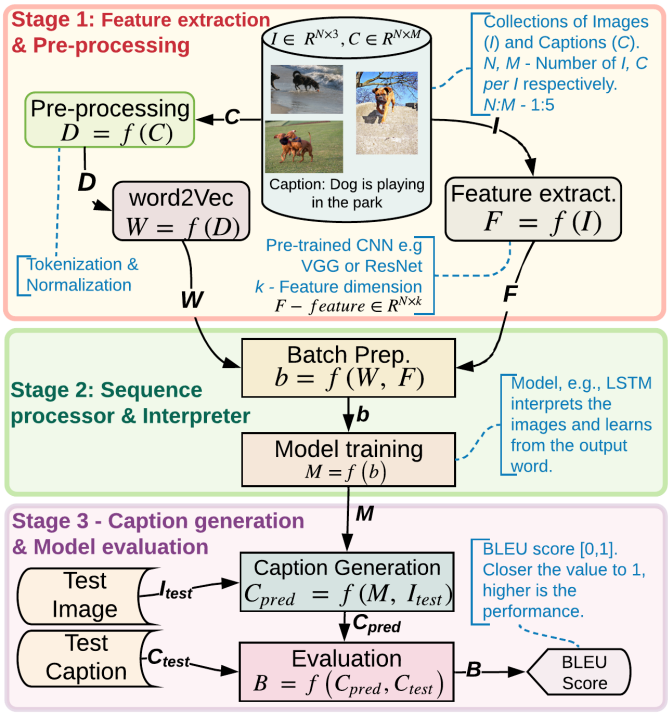

Fig. 1: Operational flow of a general visual caption generation model.

*1) **Feature extraction and Pre-processing**:* This is the first stage in the image captioning model and involves image and their captions for pre-processing. Textual pre-processing starts with tokenization and is followed by different normalization techniques, like removal of unary words, conversion to lower case and removal of punctuation as defined in the equation (1).

Feature extraction from images is performed through transfer learning whereby a ImageNet pretrained CNN, like VGG16 is used to extract the features. The loaded images are resized to the preferred size of the CNN model (3 channels of 224×224 pixel image). The features from the Fully Connected (FC) layer are extracted from the model in 1-dimensional array with the shape of 1×4096 as shown in equation (2).

In the Fig.1. the different processes involved in the existing model are demonstrated using formulas.

$$P_c = f(t_i), \tag{1}$$

where $f(.)$, i and t are the pre-processing operation, index of token ranging from 1 to N with N being the total number of tokens and token generated from the captions respectively.

$$F_{V\rightarrow VGG16} = f\left(\sum_{i=1}^{N} w_i^V x_i + b^V\right), \tag{2}$$

where $f(.)$, N, i, w_i^V, x_i and b^V are the activation function, total number of neurons in the FC layer, neuron index, weight of ith neuron of FC layer, input to the FC layer and bias of

VGG16's FC layer respectively.

*2) **Sequence processor and Interpreter**:* Before the processed descriptions and images are loaded into interpreter, the descriptions are converted to vector representations using Word2vec and batches of input-output sequences are formed. For each pair of image-caption, the input-output sequences will the input being image as well as caption and the output being the predicted word from the caption as defined in equation (3).

$$\hat{b} = f\left(P_c, F_{V \to VGG16}\right), \tag{3}$$

where $f(.), P_c$ and $F_{V \to VGG16}$ are the function for creating input-output pairs from equations (1) and (2), the preprocessed captions and the image features respectively. The ratio of images and their captions for the Flickr8k dataset is $1:5$, viz. $N:M$.

These batches of input-output pairs are then fed into the interpreter, i.e. LSTM for learning the captions for that respective image, hence the entire process is termed as Model training as shown in the equation (4).

$$\tilde{M} = f\left(\hat{b}\right), \tag{4}$$

where $f(.)$ and $\hat{b}$ are the model training of the batches of input-output pairs and the batch of the input-output pairs respectively.

*3) **Caption generation and Model evaluation**:* This is the final stage in the model where the trained model is used to predict the captions for the test image as defined in equation (5). The predicted captions are then evaluated with the actual test captions, and the similarity in the generated captions and the actual captions are calculated using the BLEU scores as shown in equation (6).

$$\hat{C} = f\left(\tilde{M}, I\right), \tag{5}$$

where $f(.)$, $\tilde{M}$ and I are the caption generation function, the trained model and the test images respectively.

$$B = f\left(C, \hat{C}\right), \tag{6}$$

where $f(.)$, C and $\hat{C}$ are the evaluation function, actual captions and predicted captions for testing respectively.

B. Proposed Approach

The proposed solution introduces two enhancements done to the existing model which are (i) fusion of word embedding techniques and (ii) feature fusion from two different feature extractors. Changes to the existing model takes place on the inputs to the LSTM. The baseline visual caption generator model's architecture definition is given in Fig.3. The proposed model has two inputs one is set of images and on the other hand the descriptions/captions related to the image. The first part of the model has the dropout layer connected to the Dense layer of 256 units with 'relu' activation function. The ReLu activation function is used to perform a threshold operation to every input element that is given where values less than zero are set to zero.

In the second part of the model where the descriptions are taken as input and the input is sent to the embedding layer where the tokenization is done and that layer is linked to one more dropout layer with 0.5 units and is connected to the LSTM with 256 units/layers.

The values from the Dense layer from image and the LSTM are sent to the decoder and are combined and captions are decoded. The activation function used here is the softmax and categorical crossentropy is used for the loss compilation. Using Adam's optimizer the model achieved its higher performance and generated commendable accuracy.

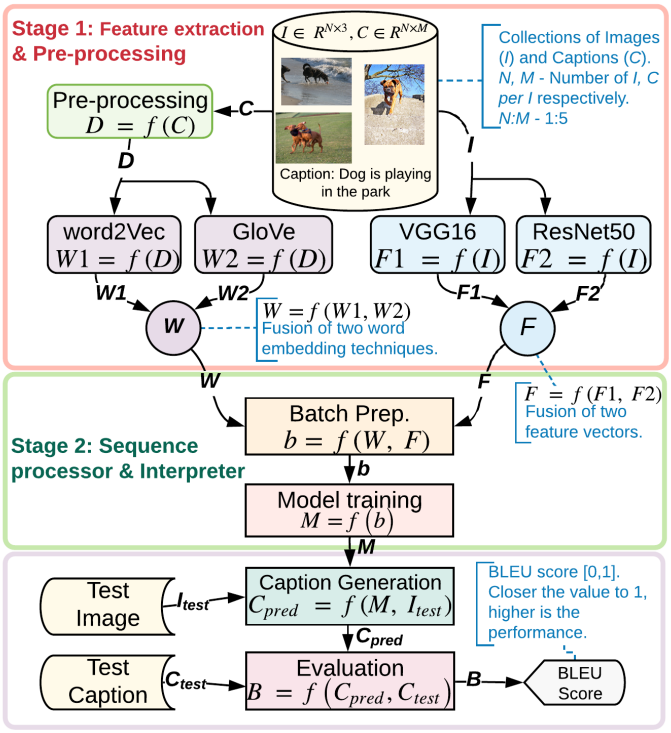

Fig. 2: Proposed multimodality feature learning for visual caption generation.

*1) **Textual feature extraction**:* There are two word embedding techniques used in the proposed model, viz. word2vec and GloVe. GloVe is a count based model which generates word vectors from their co-occurrence information. Unlike word2vec, GloVe does not rely on local contextual information of words, but incorporates global statistics to obtain word vectors. The vector representations from word2vec and GloVe are fused to get generalized representation of the captions as defined in equation (7).

$$\tilde{W} = f\left(W^{w_2v}, W^{GloVe}\right), \tag{7}$$

where $f(.)$, W^{w_2v} and W^{GloVe} are the arithmetic addition operation on the word vectors, word2vec based vector representation of the descriptions and GloVe based vector

representation of the descriptions respectively.

*2) **Visual Feature fusion**:* Similarly, there are two feature extraction models used, viz. VGG16 and ResNet50. VGG16 has the image features from the FC layer in a 1-dimensional array which is different from ResNet50. The latter has the image features from the FC layer in the shape of 1×2048. Hence, before the fusion of the two feature vectors, the output from ResNet50 has to be resized to the output shape of VGG16. The two feature vectors which are having same shapes are fused together either by addition or by multiplication as defined in equation (8).

$$\tilde{F} = f\left(F_{V \to VGG16}, F_{R \to ResNet50}\right), \tag{8}$$

where $f(.)$, $F_{V \to VGG16}$ and $F_{R \to ResNet50}$ are the operations like arithmetic addition or multiplication, features extracted from the FC layer of VGG16 and features extracted from the FC layer of ResNet50 respectively.

The two fused models from equations (7) and (8) are then fed into the interpreter for training the model and from there, the rest of the stages in the Caption generation model remains the same.

IV. Experimental Setup

A. Dataset

Flickr8k dataset is used for many computer vision tasks and due to its smaller size, training the model using this dataset becomes easier on low-end laptops or desktops. This dataset consists of 8092 images in JPEG format with varying shapes and sizes. Of which, 6000 images are used for training, 1000 for validation and 1000 for testing. There are 5 captions for each image, thereby resulting in a total of 40460 captions.

B. Training configurations

The hyper-parameters for the experimental tasks were fixed on the different models to avoid ambiguities in the comparative study. Every model is trained for a number of epochs by using early stopping with the optimizer set as Adam. Adam is an efficient stochastic optimization algorithm and it computes individual adaptive learning rates, thereby eliminating the need to set the learning rate explicitly. A generator function is used to send batches of images for model training instead of setting a batch size parameter and 2 images are sent per batch. Input sequences with a pre-defined length are sent to sequence processor model through an embedding layer and dropout rate is 0.5. This is followed by an LSTM layer with 256 memory units.

V. Results and Evaluation

A. BLEU Score calculation

BLEU Score is basically the averaged percentage of n-gram matches. In other words, for each i-gram where $i = 1, 2, ...N$, you compute the percentage of the $i - gram$ tuples in the hypothesis that also occur in the references (this is also called

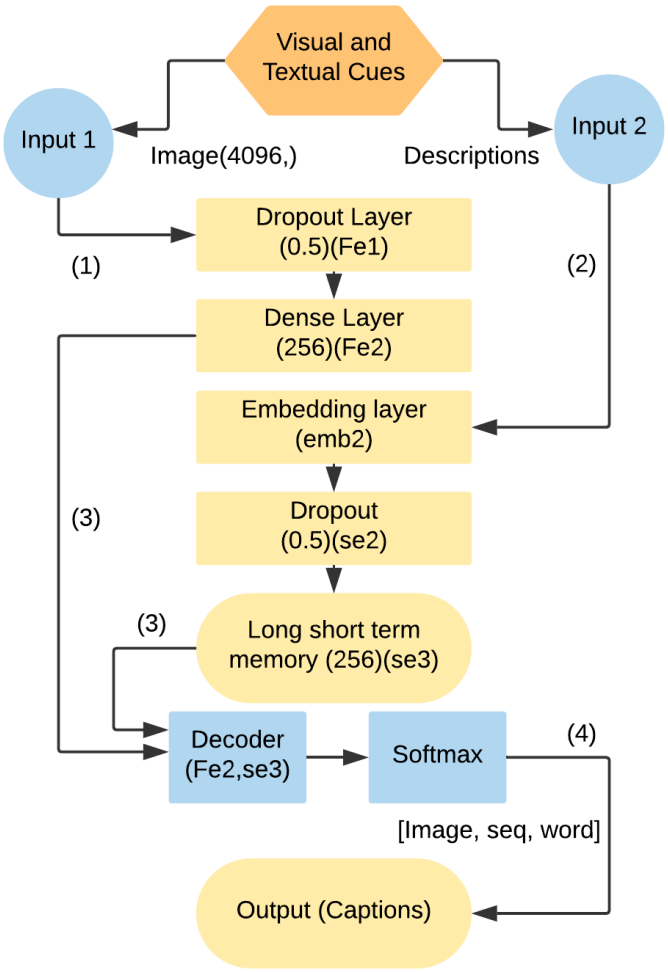

Fig. 3: Caption Generator from proposed model.

the precision):

$$P(i) = \frac{Matched(i)}{H(i)} \tag{9}$$

where $H(i)$ is the number of i-gram tuples in the hypothesis. For a hypothesis of length n words, $H(1) = n$, $H(2) = n$-1 and $H(3) = n$-2.

$$Matched(i) = \sum_{t_i} min\{C_h(t_i), \max_j C_{hj}(t_i)\}, \tag{10}$$

where t_i is an i-gram tuple in hypothesis h, $C_h(t_i)$ is the number of times t_i occurs in the hypothesis, $C_{hj}(t_i)$ is the number of times t_i occurs in reference j of this hypothesis (Remember a hypothesis may have multiple references). Randomly generating very long translation output, though will increase $Matched(i)$, the precision will stay low.

B. Quantitative analysis

The Flickr8k is run through two CNN's, viz. VGG16 and ResNet50 and also on their fusion. The performance of

unimodality based learning for caption generation is compared with proposed multimodality fusions using the corpus level BLEU score.

TABLE I: Performance Evaluation of Various Models: PSTS - Per Sample Processing Time for Caption Generation

Models	F-e(s)	We-tech	BLEU (%)	PSPT (s)
M1	VGG16	W2v	52	0.23
M2	ResNet50	W2v	51	0.30
M3	VGG16	GloVe	51	0.20
M4	ResNet50	GloVe	53	0.24
M5	VGG16 + ResNet50	W2v	51	0.43
M6	VGG16 + ResNet50	GloVe	54	0.21
M7	VGG16	W2v + GloVe	55	0.27
M8	ResNet50	W2v + GloVe	54	0.27
M9	VGG16 + ResNet50	W2v + GloVe	56	0.33

In the above Table I, F-e(s) are the feature extractors, We-tech is the word embedding technique used for textual representation of captions and BLEU is the test data accuracy in percentage on model's performance. It depicts the nine different variants of the model in which M1, M2, M3 and M4 are unimodal whereas M5, M6, M7, M8 and M9 are multimodal.

TABLE II: Comparison of best models

Model type	Models	BLEU (%)	PSPT (s)	BLEU	PSPT
Unimodal	M4	53	0.24	-	-
Visual multimodal	M6	54	0.21	1%↑	3s↓
Textual multimodal	M7	55	0.27	2%↑	3s↑
Fusion of all	M9	56	0.33	3%↑	9s↑

Fig. 4 compares the different model's accuracy based on the same parameters along with its average caption generation time and it is evident that multimodal fusion performs better compared to unimodal with 56% average BLEU score accuracy.

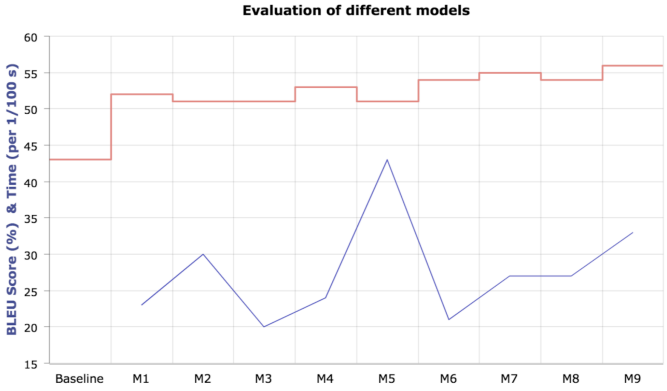

Fig. 4: Performance analysis of different models.

C. Qualitative analysis

From the output images in Fig. 5 the captions generated are classified into good and bad captions. From the first image Fig. 5a, the captions generated are

Actual: boy in his blue swim shorts at the beach
Predicted: guy stands in the ocean lifting up his hand
BLEU score: 0.49

The BLEU score for good caption indicates that the predicted captions are not exactly similar to the reference caption but few words match resulting in that score.

From the second image Fig. 5b, the captions generated are as follows.

Actual: closeup of white dog that is laying its head on its paw
Predicted: two big dogs wade in the ocean
BLEU score: 0

The BLEU score above is an indication of a perfect mismatch between reference and the predicted caption resulting in that score.

D. Timing analysis

The timing analysis is carried out on a laptop with an Intel Core i7 processor that uses Google Colaboratory cloud based programming environment with the specifications as follows, Tesla k80 GPU having 2496 CUDA cores, 12GB GDDR5 RAM and a hard disk space of 33 GB. Table II shows the average caption generation time in seconds, which helps to perform comparative study on processing time of different models for the same training procedures. It is evident from the results that though the accuracy of models M5 and M9 are same, M9 takes comparatively less processing time than M5, thereby making it a overall good fusion model.

VI. Conclusion

This research proposes multimodal visual feature and textual feature fusion strategies to improve the performance of automatic visual caption generation model. The proposed models show better results than the unimodal based image captioning system. The future work is to investigate the model's performance on different data sets which has more number of images than Flickr8k and also to incorporate visual attention.

References

[1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[4] F. Chollet, "Xception: Deep learning with depthwise separable convolutions, 2016," *arXiv preprint arXiv:1610.02357*, 2016.

[5] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

(a) Good caption: **Actual** - boy in his blue swim shorts at the beach. **Predicted** - guy stands in the ocean lifting up his hand.

(b) Bad caption: **Actual** - closeup of white dog that is laying its head on its paw. **Predicted** - two big dogs wade in the ocean.

Fig. 5: Examples for Predicted Good and Bad Captions.

[6] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 284–288, 2016.

[7] H. Wang, A. Meghawat, L.-P. Morency, and E. P. Xing, "Select-additive learning: Improving generalization in multimodal sentiment analysis," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 949–954, IEEE, 2017.

[8] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "Centralnet: a multilayer approach for multimodal fusion," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 0–0, 2018.

[9] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," *arXiv preprint arXiv:1512.02167*, 2015.

[10] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *arXiv preprint arXiv:1911.03977*, 2019.

[11] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, and P. Natarajan, "Multimodal feature fusion for robust event detection in web videos," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1298–1305, IEEE, 2012.

[12] G. Ye, D. Liu, I.-H. Jhuo, and S.-F. Chang, "Robust late fusion with rank minimization," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3021–3028, IEEE, 2012.

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[14] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *arXiv preprint arXiv:1410.5401*, 2014.

[15] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[16] D. Zhao, Z. Chang, and S. Guo, "A multimodal fusion approach for image captioning," *Neurocomputing*, vol. 329, pp. 476–485, 2019.

[17] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.