Cross-Active Connection for Image-Text Multimodal Feature Fusion

JungHyuk Im[KAIST][0000−0002−0967−6512], Wooyeong Cho[KAIST][0000−0002−9131−8086], and Dae-Shik Kim[KAIST][0000−0002−6001−1911]

[1] KAIST, South Korea
[2] http://www.kaist.ac.kr/

Abstract. Recent research fields tackle high-level machine learning tasks which often deal with multiplex datasets. Image-text multimodal learning is one of the comparatively challenging domains in Natural Language Processing. In this paper, we suggest a novel method for fusing and training the image-text multimodal feature. The proposed architecture follows a multi-step training scheme to train a neural network for image-text multimodal classification. In the training process, different groups of weights in the network are updated hierarchically in order to reflect the importance of each single modality as well as their mutual relationship. The effectiveness of Cross-Active Connection in image-text multimodal NLP tasks was verified through extensive experiments on the task of multimodal hashtag prediction and image-text feature fusion.

Keywords: Multi-modal Learning · Feature Fusion · Natural Language Processing.

1 Introduction

The development of high-performance language models has brought remarkable advance in machine learning based language tasks. As recently emerged methods [5] [14] are able to represent the complex semantic properties of words regardless of tasks, comprehension abilities of word-wise encoders have been strengthened enough to tackle challenging NLP tasks. Nevertheless, high-level tasks involving both natural language understanding and text generation such as dialogue and question answering face another drawback. In real life communication between human beings, semantic representation of text is also dependent of visual information as they affect the context of words used in an utterance. Current trends of research reflect efforts to contemplate this multimodal dependency of humanlike communication tasks. A variety of models were developed to solve the challenge of Visual Question Answering [1], along with attempts to fuse image and text features for multimodal classification tasks[15] [6]. However, many of the high-performance models are implemented with the ensemble of multiple networks dealing with different modalities. This shows that multimodal feature fusion is a field of research that still needs advancement.

Our research focuses on building a training scheme that can effectively fuse image and text features. Researchers of this field are already informed that baseline methods for image-text feature fusion involve concatenating the image and text representations separately extracted from two neural networks. But concatenation is not enough to achieve rich representation that reflects the mutual relationship between text and image information. We designed a two-phase training scheme to subdue the limitations of end-to-end models that use concatenated multimodal feature as the input. The complete architecture of the proposed model in this paper integrates two single feature extracting models with a multi-label classifier. The training scheme of the neural network multi-label classifier breaks itself down to be equivalent to training the ensemble of four individual networks; two individual single-modal networks and a set of two complementary multimodal networks.

The evaluation of our model was conducted with the task of image-text hashtag prediction. Researchers are now aware that single-modal hashtag prediction is a limited area of research as the majority of online SNS platforms deal with both image and text. It is challenging to achieve solid performance in multimodal hashtag prediction, as hashtags do not directly represent the objects in the image or written captions. To be considered as a successful approach, a multimodal predictor must not only perform better than single-modal predictors, but also be capable of handling cases when one of the two modalities does not relate to the ground truth. The results of our experiments show that the implementation of cross-active connection within the neural network is effective for building a multi-label classification model with multimodal inputs.

The main contributions of our research are summarized below:

- We present Cross-Active ConNet (CACNet), a novel network design for image-text multimodal classifier.
- Training process of CACNet involves fusing the features of two modalities within the hidden layer. As a result, the weights of the hidden layers become effective image-text feature extractor.
- The multi-level training scheme that we propose is effective for multimodal feature fusion, but too complex to implement without Batch Gradient Descent. We simplified the implementation by grouping the weight matrices into sub-sections and utilizing a virtual sigmoid output. As a result, the multi-step training scheme is reduced into the problem of training four sub-networks that add up to build CACNet, and we can apply Mini-Batch Stochastic Gradient Descent.
- Cross-Active weights are updated when the two modalities share similar latent features. This selective updating algorithm helps the network to build a complementary relationship between two modalities, making the classifer less vulnerable to cases in which one of the two inputs do not relate to the label.

– Experiments conducted in our paper show that CACNet is an effective approach for image-text multimodal classification and image-text feature fusion.

2 Related Work

Multimodal Feature Fusion Several recently published works deal with multimodal feature fusion. Multi-modal gender prediction model[15] was implemented with Gated Multimodal Units[2]. Another recent approach [6] exploits the well-known CNN sentence classification model [9] to fuse image and text features. They have shown that the fused feature performs better than baseline models of image and text single-modal classification.

Image based Hashtag Prediction HARRISON[12] is a benchmark dataset for image based hashtag prediction, which is provided along with prediction experiment results using a baseline method. The authors suggest three models for evaluation, which use features extracted from VGG-Object, VGG-Scene and both of them respectively. The evaluation results of these baseline models are included in the result section of this paper for comparison of our model against single-modal classifiers.

Multimodal Hashtag Prediction Not many published works tackle multimodal hashtag prediction. However, several online authors propose models that can handle the task. Previous work on public online repository introduces a hierarchical ensemble model of CNN [8] and word feature extractor [11] [13] for image-text hashtag prediction. They have also conducted an ablation study on the importance of hashtag segmentation in terms of text pre-processing. We constructed our own dataset to conduct the experiments of our research, adapting parts of the pre-processing methods described in their works. A multimodal hashtag predictor implemented by concatenating text feature extracted from [10] and visual feature from [16] won second place on OpenResource Hackathon 2019.

3 Methods

The Overall Architecture
The complete architecture for multi-label hashtag prediction is shown in Figure 1. The model integrates two feature extractors and a multi-label classifier. The extracted features of two modalities serve as the inputs of the multi-label classifier CACNet.

Feature Extraction
We use VGG16 model pre-trained on the 1.2 million ImageNet dataset [7] as our image feature extractor. The 1x4096 vector output is reduced into the dimension

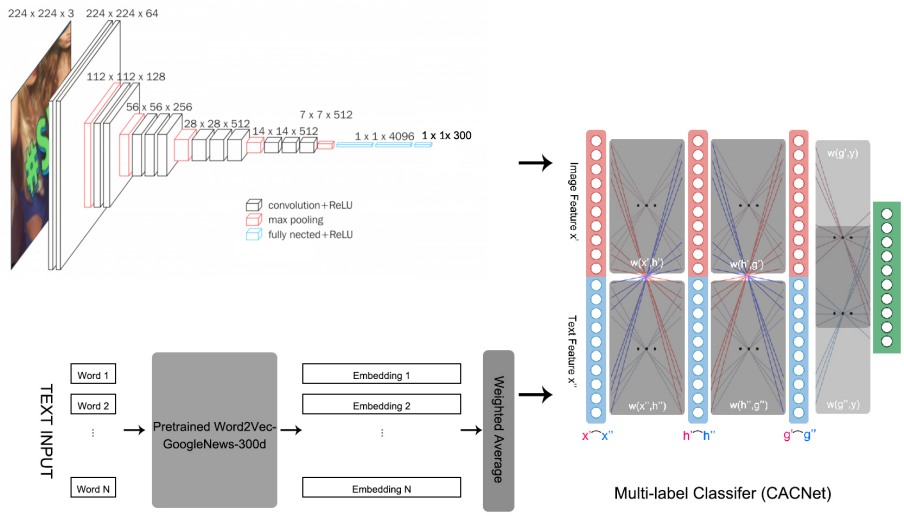

Fig. 1. The overall architecture of the multi-label classifier proposed in this paper.

of 300, which is directly used as the image feature input of the multi-label classifier. Word2Vec model pretrained on Google News corpus of over 3 million words was used as the text feature extractor. Although there exist various methods to form representations of sentence-level texts using Recurrent Neural Network based encoding methods, previous work[3] have proven that the weighted average of word embedding can strongly represent sentences. Particularly in the task of hashtag prediction, the importance of word sequence in the captions are reduced compared to other types of tasks such as reading comprehension. Thus our model takes the weighted average of the word vectors to represent the caption of an Instagram post as the text feature input instead of taking RNN based approaches.

Cross-Active Connection Network

The proposed network design of our research, CACNet serves as the multi-label classifier. It consists of two fully connected hidden layers of 600 dimension each and a sigmoid output for 300 categories of hashtags that our training dataset contains. The architecture of CACNet seems similar to a general Multi Layer Perceptron model with two hidden layers and the concatenated vector input of image and text features. The training algorithm of CACNet to be described later differentiates our classifier from general MLP for single-modal classification tasks. Using sigmoid function as the activation allows our classifer to perform multi-label classification.

In a Mini-Batch Stochastic Gradient Descent training scenario, CACNet updates the weights with a two-phase hierarchy. The idea of this training algorithm is

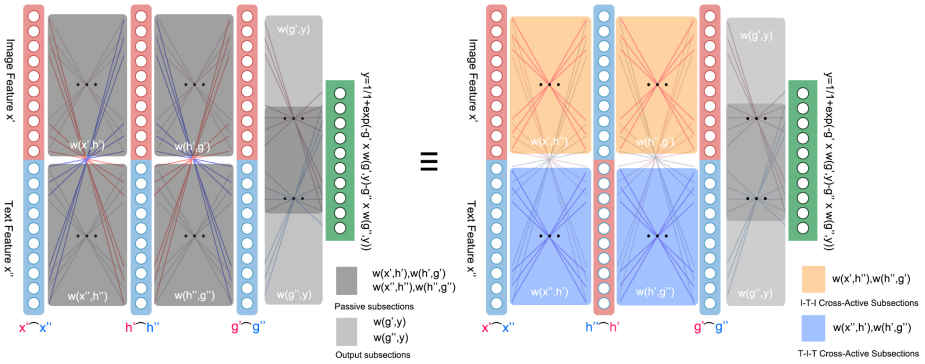

Fig. 2. Two different ways to visualize one equivalent CACNet multi-label classifier. Change of neuron alignments in the hidden layer help visualize how the weight parameters of CACNet are grouped into different sub-sections.

to maintain the relationship between the output and each single-modality while also reflecting the complementary relationship each modality shares per single iteration. The concept is similar to adaptive dropout [4] in the sense that selective parts of the neurons are deactivated in each training phase according to a control variable. Figure 2 illustrates the structure of CACNet. Each layer is notated as a concatenation of two subsections of the neural network for convenience in mathematical formulation, and the weights that connect each subsection are grouped by the notation. The network on the right side of Figure 2 is equivalent to the one on the left, where a slight change of arrangements of weights has been made. All layers are fully connected and the weights are grouped into 10 subsections as labeled in the figures.

The first phase of training involves minimizing the cross entropy cost function for the passive subsections when the cross-active connections are deactivated. We cannot directly derive the loss function from the sigmoid output y when Cross-Active subsections are deactivated, as it is connected to both of the output subsections. We introduce the concept of creating a virtual sigmoid output which is only of temporary use for deriving the cross entropy independent of the other output subsection. Solving to minimize the error between the ground truth output and the virtual sigmoid output lets each output subsection lose dependency to the other, thus we can derive the following chain rule of partial derivatives to update passive subsections related to the image feature, where t is ground truth output.

$$y' = \sigma(g' \cdot w(g', y)) \tag{1}$$

$$E' = -y' log t - (1 - y') log(1 - t) \tag{2}$$

$$\frac{\partial E'}{\partial w(g', y)} = \frac{\partial E'}{\partial y'} \cdot \frac{\partial y'}{\partial g' \cdot w(g', y)} \cdot \frac{\partial g' \cdot w(g', y)}{\partial w(g', y)} \tag{3}$$

$$\frac{\partial E'}{\partial w(h',g')} = \frac{\partial E'}{\partial g'} \cdot \frac{\partial g'}{\partial h' \cdot w(h',g')} \cdot \frac{\partial h' \cdot w(h',g')}{\partial w(h',g')} \tag{4}$$

$$\frac{\partial E'}{\partial w(x',h')} = \frac{\partial E'}{\partial h'} \cdot \frac{\partial h'}{\partial x' \cdot w(x',h')} \cdot \frac{\partial x' \cdot w(x',h')}{\partial w(x',h')} \tag{5}$$

Notations were written as matrix multiplication for convenience. The gradients in the chain rule can all be calculated since *E'* is independent of the weights of the text feature related subsections and the cross-active weights. The same procedure can be processed through the text feature subsection of CACNet vice versa, by creating another virtual output *y"*. Notice that we maintain the notation of the weight matrix *w(g',y)* to emphasize that the connection between the output subsection and *y'* is temporary. The procedure of the first phase of training is then equivalent to updating the weights of two Passive sub-networks that work as independent single-modal classifiers, illustrated in Figure 3.

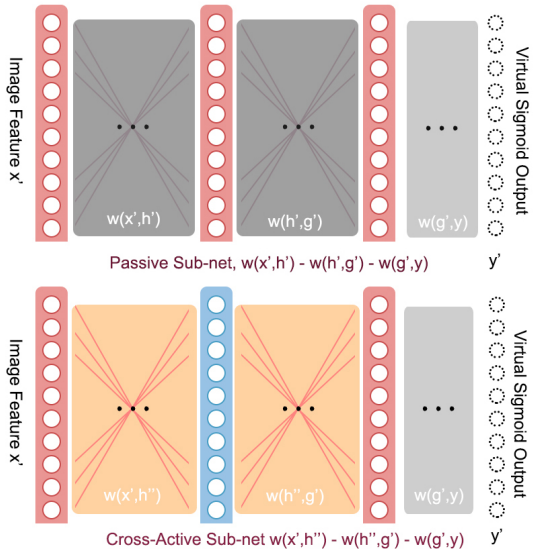

Fig. 3. The first phase of training is equivalent to training a pair of passive sub-networks to minimize the error between target output and virtual sigmoid output, when given each single-modal feature. Case 2 of the second phase is equivalent to training the Cross-Active Sub-net shown in the figure and the counterpart of it.

The second phase of training is divided into two cases controlled by the activation control variable γ,

$$\alpha = \frac{1}{N} \sum_{i=1}^{N} \frac{E'_i + E''_i}{2(-y_i log t_i - (1-y_i) log(1-t_i))} \tag{6}$$

$$\beta = \alpha/1 + \alpha \tag{7}$$

The activation parameter β shows how effective the whole network performs compared to the single-modal subsections. Low value of β also implies cases in which one of the two modalities do not reflect the training batch well. We define a control variable γ that ranges between 0 and 1 as a threshold that divides the high level training into two cases:

- **Case 1** If $\beta < \gamma$, update the whole network in an end-to-end manner without grouping the weights layer to minimize the cross-entropy loss of y. As described earlier, low value of β implies that the two modalities do not relate well, and it is better not to isolate the Passive subsections.
- **Case 2** If $\beta \geq \gamma$, activate the Cross-Active connection weights and deactivate passive subsections. Minimize the cross entropy cost function of a virtual sigmoid output by updating the cross-active weights.

In Case 2, where Cross-Active subsections are activated, the partial derivatives for calculating the gradients differ from the first phase training as follows.

$$y' = \sigma(g' \cdot w(g', y)) \tag{8}$$

$$\frac{\partial E'}{\partial w(g', y)} = \frac{\partial E'}{\partial y'} \cdot \frac{\partial y'}{\partial g' \cdot w(g', y)} \cdot \frac{\partial g' \cdot w(g', y)}{\partial w(g', y)} \tag{9}$$

$$\frac{\partial E'}{\partial w(h'', g')} = \frac{\partial E'}{\partial g'} \cdot \frac{\partial g'}{\partial g' \cdot w(h'', g')} \cdot \frac{\partial g' \cdot w(h'', g')}{\partial w(h'', g')} \tag{10}$$

$$\frac{\partial E'}{\partial w(x', h'')} = \frac{\partial E'}{\partial h''} \cdot \frac{\partial h''}{\partial h'' \cdot w(x', h'')} \cdot \frac{\partial h'' \cdot w(x', h'')}{\partial w(x', h'')} \tag{11}$$

Vice-versa can be done for the text-feature input involving counterpart subsection *w(x",h') — w(h',g") — w(g",y)* The loss function is independent to the deactivated weights when we minimize error between virtual output and the target output, so gradients involved in the partial derivatives are all easy to calculate. Notice that in the second phase, we are training the weights of the hidden layers connecting to the other modality, which we named Cross-Active subsections. This process is equivalent to training another complementary pair of sub-networks of structure labeled as Cross-Active Sub-net in Figure 3.

When $\beta \geq \gamma$, the complexity of weight updates involving activating and deactivating parts of the network makes the procedure difficult to implement, especially for Mini-Batch Stochastic Gradient Descent scenarios. By utilizing virtual sigmoid outputs and grouping the weight matrix into subsections, we simplified the two-level training process into an equivalent problem of updating weights for 4 sub-networks given the same input and target output. After an iteration of Case 2 in second phase training, the 4 sub-networks jointly form CACNet.

4 Experiments

Our model was implemented with PyTorch 1.6.0, under a multi-GPU environment with 4 NVIDIA Titan Xp GPUs installed and CUDA Toolkit 10.2. The training procedure was conducted by Mini-Batch Stochastic Gradient Descent of batch size 20. Our complete dataset consists of 30k pairs of image-text multimodal inputs and text output. The CACNet classifier was trained over 500 epochs on the training dataset. The activation control variable γ described in the Methods section was set to 0.4 at the start of the training, and linearly increased up to 0.8 in the last 100 epochs. **Dataset**
There are some benchmark datasets for image-based hashtag prediction[12] [17], but there are no public dataset available for use in image-text multimodal hashtag prediction. For training and evaluation of our classifier, we constructed our own datasets. The process was conducted by scraping Instagram posts using Selenium over top 300 popular hashtags, as last updated on 2020-08-20. Non-english segments of the post including emoticons and special characters were removed, and the characters were converted into lower case.There has been a study about hashtag segmentation using the Viterbi algorithm to overcome the complexity caused by hashtags in Instagram posts combining multiple words into a single tag. Our multi-label classifier does not involve the ensemble of word embeddings in the prediction stage, so the pre-processing method was unnecessary. The ground truth outputs of our dataset consists of up to 10 hashtags used in a post. The details of the training and evaluation datasets are explicitly shown in Table 1 and Figure 4.

	CACTrain	**CACEval**
# of posts	25,017	5,000
average # of words per caption	12.82	13.1
average # of Hashtags per post	8.71	9.11
Hashtag categories	300	300
average # of <unk> per caption	2.51	2.27

Table 1. Details of the dataset used in our research. Train and Evaluation sets were both collected with Selenium web crawler.

We conducted experiments to evaluate our model against baseline methods in two tasks, Image-Text Feature Fusion and Image-Text Multimodal HashTag Prediction.
Feature Fusion The weight parameters of the trained CACNet can be extracted to serve the task of image-text feature fusion. We evaluated the feature fusion performance of CACNet against baseline methods published with UPMC Food-101, a large multimodal dataset that contains over 100k food recipes classified

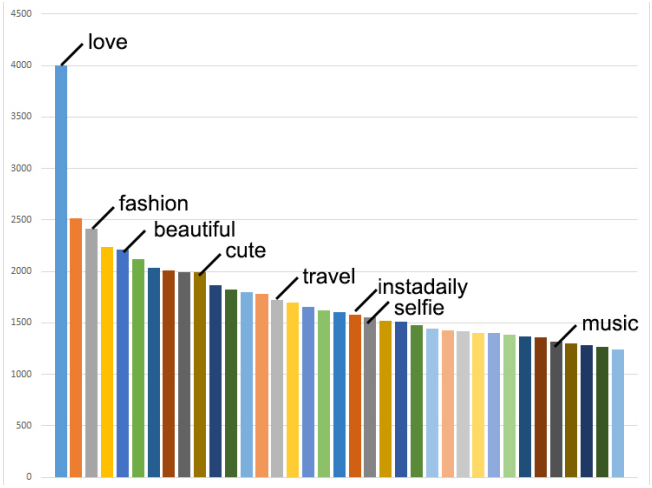

Fig. 4. The number of posts containing 40 mostly appearing hashtags in our training dataset. Single post is labeled with up to 10 multiple hashtags

.

in 101 categories.

Hashtag Prediction Despite the efforts of researchers on image-text multimodal tasks, there are no available published work that we can evaluate performance of CACNet on multimodal HashTag prediction against. To prove the validity of our multimodal classifier, we evaluated our model under the metrics of [12], as they provide a benchmark dataset for image-based Hashtag prediction and a baseline model. For generic evaluation, we also trained and evaluated their baseline model with our independent dataset.

5 Results

Feature Fusion Performance in image-text feature fusion task was evaluated using the UPMC Food-101 dataset. As the authors describe, higher scores with text-only baseline method result from the bias introduced by their data crawling protocol. Evaluation was performed by comparing our results against their baseline models[18]. Our classifier CACNet achieved higher performance in classification than the baseline models. The results are shown in Table 2.

Prediction examples shown in Figure 5 show successful prediction examples in challenging cases, all of which single-modal baseline classifiers fail to predict accurately. CACNet successfully predicts hashtags in cases even when the input image does not relate to the ground truth hashtags, or when the words in the input caption are useless. **HashTag Prediction** We used *Precision@K*, *Recall@K*, *Accuracy@K* as the evaluation measures for quantitative comparison against the

Methods	Avg.Precision
Very Deep (Vision only)	40.21%
TF-IDF (Text only)	82.06%
TF-IDF + Very Deep (Fusion)	85.10%
VGG16-Word2Vec300-CACNet(Fusion)	**87.63%**

Table 2. Evaluation on the UPMC Food-101 dataset.

Methods@Dataset	Precision@1	Recall@5	Accuracy@5
VGG-Object@HARRISON	28.30%	20.83%	50.70%
VGG-Scene@HARRISON	25.34%	18.66%	46.30%
VGG-Object + VGG-Scene@HARRISON	30.16%	21.38%	52.52%
VGG16-Word2Vec300-CACNet@CACEval	**59.7%**	**42.72%**	**71.13%**

Table 3. Comparison of performance for hashtag prediction against image baseline models trained with HARRISON benchmark dataset.

Methods	Accuracy@1	Accuracy@3	Accuracy@5
VGG-Object	8.41%	37.56%	48.12%
VGG-Scene	7.8%	33.74%	47.71%
VGG-Object + VGG-Scene	9.64%	38.44%	54.81%
VGG16-Word2Vec300-CACNet(γ=1)	9.11%	41.47%	55.19%
VGG16-Word2Vec300-CACNet	**12.82%**	**48.91%**	**71.13%**

Table 4. Generic evaluations of baseline models and our model measured with *Accuracy@K*

baseline models introduced in the HARRISON benchmark dataset.[12]
Precision@K is the portion of top K ranked hashtags that match ground truth output. *Recall@K* is the portion of ground truth hashtags that match top K ranked hashtags. *Accuracy@K* is defined as 1 if there exists at least one match between top K ranked hashtags and the ground truth hashtags. The evaluation results for Hashtag prediction are shown in Table 3. The HARRISON benchmark contains 1,000 categories of hashtags while our dataset contains 300 categories. Thus the quantitative comparison of best results might not be reliable. For generic evaluation, we conducted further research by evaluating the baseline methods provided by HARRISON benchmark on our dataset, CACEval. Instead of using precision and recall as the metric, we evaluated the models on *Accuracy@K* only. To show the validity of our training scheme, we also trained a version of CACNet with the control variable γ set to 1. Table 4 shows the generic evaluation results.

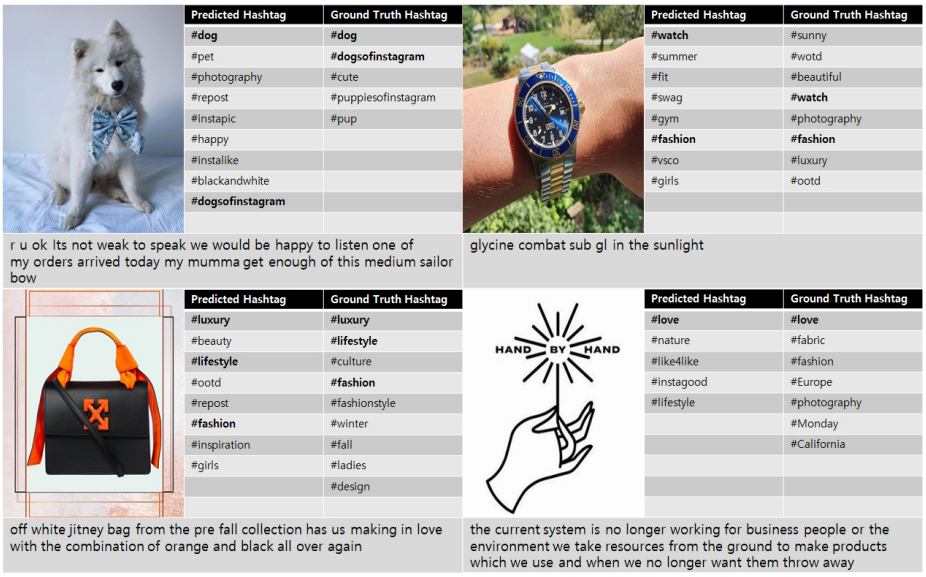

Fig. 5. Examples of successful predictions are shown above. Matching hashtags are in bold letters. Examples show cases in which our classifier was able to predict multiple matches with the ground truth when either one of caption or image are hard to relate to the hashtags.

6 Discussion

In this paper, we introduced a novel method for training a network with multimodal inputs. As far as we know, the multi-label classifier trained with our implementation, CACNet holds the state-of-the-art performance in hashtag prediction tasks. Our model has advantages over ensemble-based approaches and end-to-end approaches. The multi-phase training scheme lets the network maintain single-modal dependency as well as fusing the complimentary characteristics of two modalities. Another contribution of our research comes from introducing the concept of virtual outputs when observe the gradients from small sections of weight parameters in a whole network. This approach makes it possible to divide a network into sub-sections of weights and simplify complex training schemes. We expect our works to inspire fields of research involving image-text multimodal classification and feature fusion.

7 Acknowledgements

This work was supported by Institute for Information communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.2016-0-00563, Research on Adaptive Machine Learning Technology Development for Intelligent Autonomous Digital Companion)

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
2. Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992 (2017)
3. Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence embeddings (2016)
4. Ba, J., Frey, B.: Adaptive dropout for training deep neural networks. In: Advances in neural information processing systems. pp. 3084–3092 (2013)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Gallo, I., Calefati, A., Nawaz, S., Janjua, M.K.: Image and encoded text fusion for multi-modal classification. In: 2018 Digital Image Computing: Techniques and Applications (DICTA). pp. 1–7. IEEE (2018)
7. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning, vol. 1. MIT press Cambridge (2016)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
10. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019)
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
12. Park, M., Li, H., Kim, J.: Harrison: A benchmark on hashtag recommendation for real-world images in social networks. arXiv preprint arXiv:1605.05054 (2016)
13. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
14. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)
15. Sierra, S., González, F.A.: Combining textual and visual representations for multimodal author profiling. Working Notes Papers of the CLEF **2125**, 219–228 (2018)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
17. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
18. Wang, X., Kumar, D., Thome, N., Cord, M., Precioso, F.: Recipe recognition with large multimodal food dataset. In: 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). pp. 1–6. IEEE (2015)