TWO-STREAM HYBRID ATTENTION NETWORK FOR MULTIMODAL CLASSIFICATION

Qipin Chen[1] *Zhenyu Shi*[2] *Zhen Zuo*[2] *Jinmiao Fu*[2] *Yi Sun*[2]

[1] Pennsylvania State University [2] Amazon LLC

ABSTRACT

On modern e-commerce platforms like Amazon, the number of products is fast growing, precise and efficient product classification becomes a key lever to great customer shopping experience. To tackle the large-scale product classification problem, a major challenge is how to leverage multimodal product information (e.g., image, text). One of the most successful directions is the attention-based deep multimodal learning, where there are mainly two types of frameworks: 1) *keyless attention*, which learns the importance of features within each modal; and 2) *key-based attention*, which learns the importance of features using other modalities. In this paper, we propose a novel Two-stream Hybrid Attention Network (HANet), which leverages both key-based and keyless attention mechanisms to capture the key information across product image and title modalities. We experimentally show that our HANet achieves state-of-the-art performance on Amazon-scale product classification problem.

Index Terms— multimodal classification, hybrid attention mechanism

1. INTRODUCTION

Product classification is a key functionality for the e-commerce platform to categorize the products. Customers rely on such information to efficiently search for products based on key words (e.g., t-shirt, dress). The recommendation systems can also leverage the category information to provide customers with reliable shopping experiences (e.g., recommend similar products from the same category, return relevant products from the target categories that match customers' searching queries).

Product classification has been normally addressed as a text classification problem as most of the product information is represented as textual features such as product titles [1]. However, since images are also available for most products, it is natural to include both text and image information to complement each other, which forms the problem into a multimodal classification problem.

In the deep learning area, there has been extensive research in the multimodal learning domain. Popular applications include but not limit to spatial and temporal multimodal fusion in video classification [2], and image caption generation [3]. In these researches, attention mechanism [4, 5] has been applied as a key lever to distill salient components among input features. Based on the ways of feature selection, attention mechanism can be categorized into two major directions [6]: key-based attention which uses keys to search prominent features, and keyless attention where the attention scores are generated without keys.

Keyless attention mechanism can only capture the intra modal salient information, while key-based attention can only capture the inter modal complementary information. To incorporate both information, we propose a novel end-to-end two-stream Hybrid Attention Network (two-stream HANet), which leverages both keyless and key-based attention to fuse the features from product images and text. More specifically, keyless attention focuses on extracting information of each stream (i.e. image/text stream), while key-based attention focuses on the correlation between two streams.

2. RELATED WORKS

2.1. Deep multimodal learning

The focus of deep multimodal learning is to project the feature vectors from different modalities into the same feature space. There are three major types of frameworks [6]: (1) Joint representation, which aims to project unimodal representations into a shared semantic subspace. The projected representations can then be fused and leveraged for tasks such as classification. Typical applications include video classification [2, 7], speech recognition [25], and visual question answering [8], etc. (2) Coordinated representation, which learns separated but constrained representations for each modality in a coordinated subspace. Typical applications include cross-modal retrieval [9], cross-modal embedding [10], etc. (3) Encoder-decoder models, which learns an intermediate representation to map one modality into another. Typical applications include image captioning [3], text to image synthesis [11], etc. In this paper, we focus on joint representation, which matches with our product classification application.

Typical methods of fusing unimodal representations include direct concatenation, linear combination and outer production across unimodal representations [12]. [8] utilized Multimodal Compact Bilinear pooling to avoid the computation cost of outer product. [13] added regularization to encourage the statistic distribution of the activations in the intermediate layers across modalities to be consistent. Instead of fusing the unimodal representation and building neural nets on top of it for the classification task, [25] leveraged an additional policy network to choose between the unimodal classification outputs. Existing methods either do not train the end-to-end network by considering the unimodal representations as fixed, or do not consider the feature importance within unimodal representations. In this work, we apply attention mechanism to help the network focus on salient unimodal information and train the network in an end-to-end fashion.

2.2. Attention mechanism

There are multiple mechanisms which can be integrated into the deep multimodal learning framework to strengthen its performance, including generative adversarial network [14] and attention mechanism [15]. In this work, we focus on the attention mechanism. In recent years, many advancements on deep multimodal learning with attention mechanism have validated the feasibility and superiority of both key-based attention [16, 17] and keyless attention [18, 19]. For applications with spatial and temporal information, such as video classification and image captioning, a state-of-the-art structure – attention mechanism [4, 5] has been applied to distill salient components or regions among input features. Based on the ways of selecting features, attention mechanism can be categorized into two directions [6]: key-based attention which uses keys to search prominent features, and keyless attention where the attention scores are generated without keys. In two-stream HANet, we leverage the advantages of both key-based and keyless attention to exploit multimodal information.

3. METHODS

3.1. Overview

Fig.1 shows the structure of the two-stream HANet, which consists of three parts: 1) image model (e.g. EfficientNet [20]); 2) text model (e.g. BERT [21]); 3) two-stream fusion network with hybrid attention mechanism that combines both keyless and key-based attention.

As shown in Fig.1, we first leverage the state-of-the art BERT [21] for the text embedding generation, and EfficientNet [20] for the image embedding generation to get high-quality inputs. Second, we build the two-stream fusion network with hybrid attention, which integrates the keyless attention to capture intra modal salient information, and key-based attention to capture the inter modal complementary information. Finally, we connect the two stream HANet to the categorization task to optimize for product category prediction.

3.2. Hybrid attention

3.2.1. Keyless Attention

Fig. 2 (a) shows the structure of keyless attention mechanism of HANet. Keyless attention [18] directly generates attention scores without any key involved. To extract intra-information of each stream, we apply keyless attention on text and image embeddings separately:

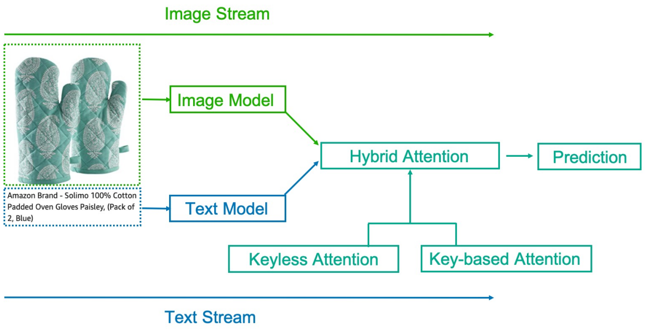

Fig. 1: The structure of two-stream HANet.

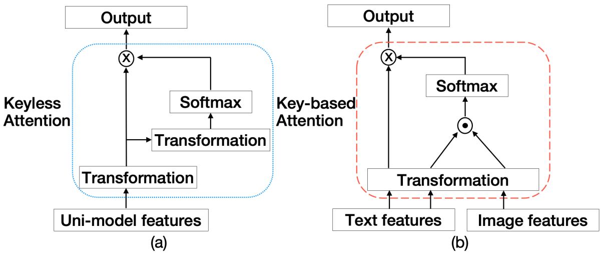

Fig. 2: (a) The structure of keyless attention mechanism of HANet for unimodal feature. (b) The structure of key-based attention mechanism of HANet for text stream.

$$\begin{aligned} wf_t &= s_{t;l} \cdot v_t \\ wf_i &= s_{i;l} \cdot v_i, \end{aligned} \tag{1}$$

where wf_t, wf_i are weighted text feature and weighted image feature. v_t and v_i are values of text and image, which are computed as:

$$\begin{aligned} v_t &= W_t f_t + b_t \\ v_i &= W_i f_i + b_i, \end{aligned} \tag{2}$$

where f_t and f_i are the features of text and image respectively. W_t, W_i, b_t and b_i are learnable parameters. $s_{t;l}$ and $s_{i;l}$ are the keyless attention scores which are computed by:

$$[s_{t;l}, s_{i;l}] = softmax([rs_{t;l}, rs_{i;l}]). \tag{3}$$

where $rs_{t;l}$ and $rs_{i;l}$ are raw scores which are also linear transformations of v_t and v_i respectively. By applying the keyless attention, two-stream HANet is able to decide which stream is more prominent for final prediction based on the keyless attention scores. However, when generating attention scores, keyless attention will only focus on the feature within each single modality. Namely, keyless attention can only extract the intra-information of each stream and ignore the inter-information between text and image streams.

3.2.2. Key-based Attention

Key-based attention, which uses keys to generate attention scores and extract inter-information between two streams. Inspired by the work of [15], we employed the Dot-Product attention mechanism for our two-stream HANet as the inter-information extractor. Fig. 2 (b) shows the structure of key-based attention mechanism of two-stream HANet. Similarly, attention scores are directly applied on the values of each stream:

$$\begin{aligned} wf_t &= s_{t;b} \cdot v_t \\ wf_i &= s_{i;b} \cdot v_i, \end{aligned} \tag{4}$$

where wf_t, wf_i are weighted text feature and weighted image feature. v_t and v_i are values of text and image in equation (2). $s_{t;b}$ and $s_{i;b}$ are key-based attention scores of text and image stream, can be computed by the following:

$$[s_{t;b}, s_{i;b}] = softmax([k_t^T q_i, k_i^T q_t]), \tag{5}$$

where k_t and q_t are key and query of text features, while k_i and q_i are key and query of image features. All keys and queries are linear transformations of their corresponding features. Different from keyless attention, the scores of key-based attention are generated by evaluating the correlation between text and image features.

3.2.3. Two-Stream Hybrid Attention

Since keyless attention and key-based attention focus on intra- and inter-information of different modalities respectively, to fully exploit the advantages of multimodalities, we propose hybrid attention mechanism which consists of both keyless and key-based attention. The hybrid attention was implemented by embedding the keyless attention into the Dot-Product key-based attention. Fig. 3 shows the structure of hybrid attention mechanism of two-stream HANet. The weighted features were also generated by applying attention scores on the values of each stream:

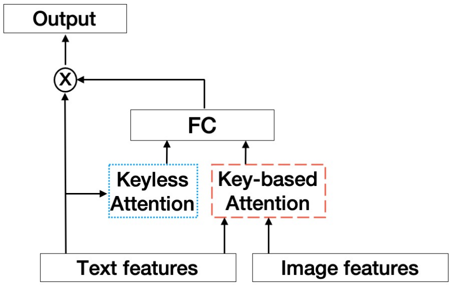

Fig. 3: The structure of hybrid attention mechanism of HANet for text stream.

$$\begin{aligned} wf_t &= s_t \cdot v_t \\ wf_i &= s_i \cdot v_i, \end{aligned} \tag{6}$$

where s_t and s_i are the hybrid attention scores for text and image stream. For generating these scores, we used a one-hidden-layer network to fuse both keyless and key-based attention scores. Namely,

$$\begin{aligned} s_t &= \mathcal{D}([s_{t;l}, s_{t;b}]) \\ s_i &= \mathcal{D}([s_{i;i}, s_{i;b}]), \end{aligned} \tag{7}$$

where $\mathcal{D}$ is a N-hidden-layer fully connected network, (N=1 in our experiments)

4. EXPERIMENTS

4.1. Dataset

The dataset we used contains 1.5 million products with 5.6K classes selected from Amazon catalog system. The 5.6K classes describe a product's classification in a very granular level, such as men's casual t-shirt. In the previous multimodal classification studies, [1, 22, 23, 24, 25] used the data set with size ranges from 10K-1M and number of labels from 17-3K. In contrast, our problem is more challenging considering its larger data size and number of classes. We divide the data into training/validation datasets with 1.2 million products and 0.3 million products, respectively.

4.2. Experiment details

To train a baseline text model before the multimodal joint learning, we used BERT-base model which contains 12 layers (transformer blocks), 12 self-attention heads with the hidden size of 768 to extract text features. The total number of parameters is 110M. The BERT-base model was pre-trained on the BooksCorpus [26] and English Wikipedia which contains 800M and 2,500M words respectively. During the training process of transfer learning, the learning rate and batch size were set to be 2×10^{-5} and 32 respectively. We employed AdamW [27, 28] optimizer with weight decay of 0.01. After training for 10 epochs, we got a validation accuracy of 83.53%.

For pre-training image model, we used EfficientNet-b0 as the image feature extractor. During the training process, the learning rate was set to be 0.1 at the beginning and was decreased by 0.1 for every 30 epochs. We employed Stochastic Gradient Descent (SGD) [29] optimizer with weight decay of 1×10^{-4}. After training for 90 epochs with batch size of 512, the validation accuracy reached 66.73%.

For our two-stream HANet training, the learning rate was set to be 2×10^{-5} and a linear warm-up schedule with warm-up proportion of 0.1 was applied. The optimizer and batch size were also AdamW and 32 as pre-training text model. All the results are reported on validation dataset after training 10 epochs on training dataset.

4.3. Error Analysis

In Table 1, we did error analysis on the validation data with baseline text and text models. We observed that the text

On Validation Data For Baseline Model	Percentage
Text model is correct image model is wrong	20.49%
Text model is wrong image model is correct	3.68%
Both model are correct	63.05%
Both model are wrong	12.78%

Table 1: Top-1 prediction error analysis for unimodal models.

Model Name	Validation Acc.
Attention-free network	83.80%
Two-stream keyless attention network	84.55%
Two-stream HANet	**84.82**%

Table 2: Performance comparison among different attention mechanisms.

Model Name	Validation Acc.
EfficientNet-b0 [20]	66.73%
BERT [21]	83.53%
Decision-level fusion network	83.60%
Policy network [25]	83.74%
Feature level fusion network	83.80%
Keyless attention only	84.55%
Two-stream HANet	**84.82**%

Table 3: Performance comparison among different multimodal classification approaches.

(BERT) model performs much better than image (EfficientNet) model. We want to point out this is not because we did not train the image model well. Instead, this is natural in ecommerce domain, where most of the time, the product title is able to accurately describe the item itself. Unlike text, it is hard for image models to distinguish classes with subtle difference such as sport shirt vs casual shirt. However, image model could still make 3.68% potential contribution to improve the baseline text model, which could be considered as the improvement upper bound for all the fusion models.

4.4. Ablation study

To validate the superiority of two-stream hybrid attention mechanism, in Table 2, we compared the performances of two-stream HANet, two-stream keyless attention network and attention-free network which naively concatenates text and image features.

As we can observe from Table 2 2^{nd} row, the attention-free network has lower accuracy than networks with attention mechanism. Moreover, as show in Table 2 3^{rd} row and 4^{th} row, the performance of two-stream HANet is better than two-stream keyless attention network, which means that two-stream hybrid attention mechanism leverages the advantages of both keyless and key-based attention. Intra- and inter-information are complementarily combined for a better prediction.

4.5. Comparison with other multimodal classification approaches

In Table 3, we compare two-stream HANet with different structures of multimodal classification networks including policy network decision-level fusion network and feature level fusion network. We also list keyless attention only result here for reference. For policy network [25], it is a fully-connected structure which is designed to learn the strategy of choosing the right model between text and image models given the concatenations of top-k class probabilities of each model. Decision-level fusion network is also a fully-connected network which takes the concatenation of all class probabilities of each model as input and directly gives the classification prediction. Feature level fusion network simply concatenates text and image features, add classification layers on top of it and trains the entire network end to end.

In Table 3, we observed that all multimodal networks perform better than unimodal models. For multimodal network comparison, we choose the decision level fusion result (83.60%) as the baseline. Policy network shows a comparable performance with decision level network, with (+0.14%) improvement. Feature level fusion network is able to achieve (+0.20%) improvement as the network has more generalization capabilities with end to end model training. Keyless attention model is able to achieve (+0.95%) improvement, as it is able to capture additional intra modal salient information compared to feature level fusion network. Two-stream HANet beat all the benchmark models with (+1.22%) improvement due to its well designed structure.

5. CONCLUSION

In this paper, we propose a two-stream hybrid attention network for multimodal large-scale product classification. By leveraging advantages of both keyless and key-based attention mechanisms, multimodal information and learning ability of two-stream HANet are fully utilized which lead to an improvement in classification accuracies. Experiment results on Amazon-scale dataset have validated the superiority of two-stream HANet in multimodal classification.

6. REFERENCES

[1] A. Kannan, P. P Talukdar, N. Rasiwasia, and Q. Ke, “Improving product classification using images,” in *2011 ICDM*. IEEE, 2011, pp. 310–319.

[2] Y. Liu, X. Feng, and Z. Zhou, “Multimodal video classification with stacked contractive autoencoders,” *Signal Process*, vol. 120, pp. 761–766, 2016.

[3] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, at-

tend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[4] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[5] I. Sutskever, O. Vinyals, and Q. V Le, "Sequence to sequence learning with neural networks," in *NeurIPS*, 2014, pp. 3104–3112.

[6] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.

[7] Y. Jiang, Z. Wu, J. Wang, X. Xue, and S. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 40, 2018.

[8] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *EMNLP*, 2016, pp. 457–468.

[9] Y. Peng, J. Qi, and Y. Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," in *IEEE Transactions on Image Processing*, 2018, vol. 27, pp. 5585–5599.

[10] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *CVPR*, 2016, pp. 4594–4602.

[11] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," *ICCV*, pp. 5907–5915.

[12] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. Morency, "Tensor fusion network for multimodal sentiment analysis," in *EMNLP*, 2017, pp. 1103–1114.

[13] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, and A. Torralba, "Cross-modal scene networks," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 40, no. 10, pp. 2303–2314, 2018.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Cambridge, MA, USA, 2014, vol. 2, pp. 2672–2680, MIT Press.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[16] H. Nam, J. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *CVPR*, 2017, pp. 299–307.

[17] C. Hori, T. Hori, T.Y Lee, Z. Zhang, B. Harsham, J. R Hershey, T. K Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *ICCV*, 2017, pp. 4193–4202.

[18] X. Long, C. Gan, G. De Melo, X. Liu, Y. Li, F. Li, and S. Wen, "Multimodal keyless attention fusion for video classification," in *AAAI*, 2018.

[19] A. Zadeh, P. P Liang, N. Mazumder, S. Poria, E. Cambria, and L. Morency, "Memory fusion network for multi-view sequential learning," *arXiv preprint arXiv:1802.00927*, 2018.

[20] M. Tan and Q. V Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019.

[21] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[22] S. Poria, E. Cambria, N. Howard, G. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, 2016.

[23] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *ICML*, 2011.

[24] A. Frome, G. S Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *NeurIPS*, 2013, pp. 2121–2129.

[25] T. Zahavy, A. Magnani, A. Krishnan, and S. Mannor, "Is a picture worth a thousand words? a deep multimodal fusion architecture for product classification in e-commerce," *arXiv preprint arXiv:1611.09534*, 2016.

[26] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *ICCV*, 2015, pp. 19–27.

[27] D. P Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[29] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.