ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9415007

HIERARCHICAL SIMILARITY LEARNING FOR LANGUAGE-BASED PRODUCT IMAGE RETRIEVAL

Zhe Ma[1], *Fenghao Liu*[1], *Jianfeng Dong*[2,3]*, *Xiaoye Qu*[4], *Yuan He*[5], *Shouling Ji*[1,3]

[1]Zhejiang University, [2]Zhejiang Gongshang University,
[3]Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies,
[4]Huazhong University of Science and Technology, [5]Alibaba Group

ABSTRACT

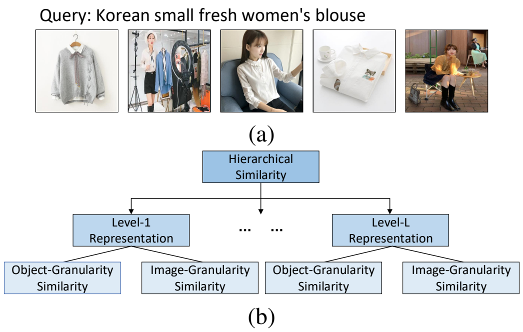
This paper aims for the language-based product image retrieval task. The majority of previous works have made significant progress by designing network structure, similarity measurement, and loss function. However, they typically perform vision-text matching at certain granularity regardless of the intrinsic multiple granularities of images. In this paper, we focus on the cross-modal similarity measurement, and propose a novel **H**ierarchical **S**imilarity **L**earning (HSL) network. HSL first learns multi-level representations of input data by stacked encoders, and object-granularity similarity and image-granularity similarity are computed at each level. All the similarities are combined as the final hierarchical cross-modal similarity. Experiments on a large-scale product retrieval dataset demonstrate the effectiveness of our proposed method. Code and data are available at https://github.com/liufh1/hsl.

Index Terms— Product retrieval, Hierarchical similarity, Multi-level representation, Cross-modal retrieval

1. INTRODUCTION

Cross-modal retrieval is a classical task at the intersection between computer vision and natural language processing, and has been widely explored [1, 2, 3, 4]. Recently, with the increasing popularity of e-commerce platforms [5, 6, 7, 8], language-based product image retrieval attracts increasing attention [9, 10, 11]. As exemplified in Fig. 1(a), given a textual query, the task is asked to retrieve images containing products that are specified by the given query. In contrast to general cross-modal retrieval, product images seem to be more diverse. As shown in Fig. 1(a), a clothing product can be shown in isolation or on the mannequin; can be folded or not; can be displayed with a clear or complex background, showing the challenging characteristics of the language-based product image retrieval task.

Recent works for language-based product image retrieval tend to exploit strong image and textual query representations to tackle the problem [9, 10, 11]. For instance, Huang *et al.* [9] borrow the ideas of Modular Co-Attention Networks (MCAN) [12] and VisualBERT [13], taking advantage of the power of multi-head self-attention/transformer [14] to encode images and textual queries. Deriving from LXMERT [15], Zhang *et al.* [10] first encode images and textual queries by stacked self-attention modules and subsequently employ cross-modal guided attention to obtain robust image and query representations. In [11], Ding *et al.* also leverage transformer-based encoder to represent images and textual queries, but they formulate the retrieval as a multi-task problem where image captioning [16] is integrated as an extra task to constrain the learned image representation. Although the above methods utilize stacked encoders, only the high-level visual and textual representations are used for similarity measurement. Besides, they only perform cross-modal matching on single granularity, *e.g.,* image-query matching [9, 10], image-word matching [11]. Considering product images are of diverse characteristics, we argue that such single-granularity similarity based on the representation of a specific level is sub-optimal.


Query: Korean small fresh women's blouse
Hierarchical Similarity
Level-1 Representation
... ...
Level-L Representation
Object-Granularity Similarity
Image-Granularity Similarity
Object-Granularity Similarity
Image-Granularity Similarity

(a)

(b)

Fig. 1. (a) An example of language-based product image retrieval. (b) Our proposed HSL learns cross-modal similarities between images and textual queries in a hierarchical manner.

In this paper, we propose a Hierarchical Similarity Learning (HSL) network which simultaneously exploits multi-level representations of images and textual queries, and multi-granularity similarities at each level. As shown in Fig. 1(b), we exploit the multiple similarities in a hierarchical manner. The framework of our proposed HSL is illustrated in Fig. 2. For both images and queries, multiple encoders are stacked to progressively learn the multi-level representations. These representations, generated by distinct encoders, are complementary to each other, which allows us to obtain an effective cross-modal similarity measurement. Moreover, considering product images usually include a number of objects, we propose to use both object-granularity similarity and image-granularity similarity to measure the relevance between images and queries. In summary, the main contributions of this paper are:

- We propose a Hierarchical Similarity Learning (HSL) network, which jointly exploits multi-level representations and multi-granularity similarities. Based on the multi-level representations and multi-granularity similarities, the final cross-modal similarity between images and queries is computed in a hierarchical manner.
- Our proposed HSL consistently performs better than general cross-modal retrieval models and models particularly designed for product image retrieval. Experiments on a large-scale language-based product image retrieval dataset demonstrate the effectiveness of our proposed method.

*Corresponding author: Jianfeng Dong

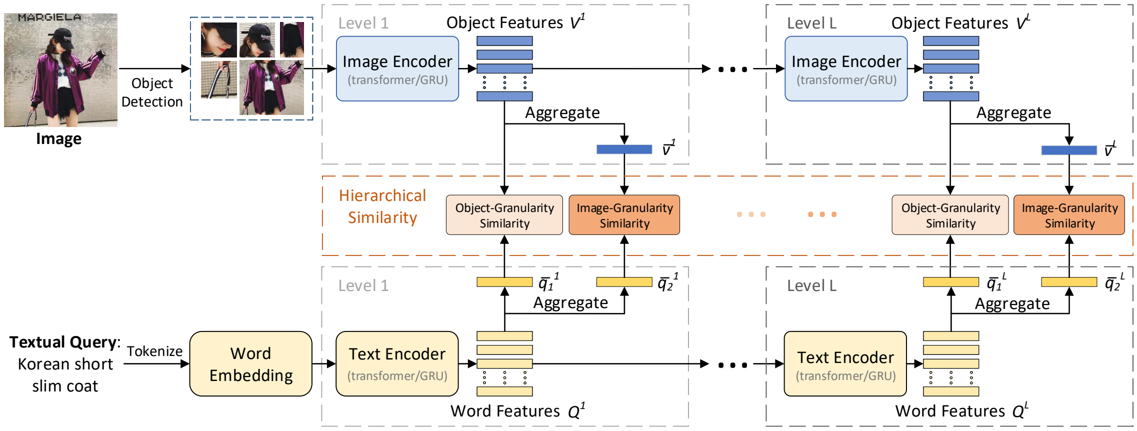


Fig. 2. Framework of our proposed Hierarchical Similarity Learning (HSL) network.

2. HIERARCHICAL SIMILARITY LEARNING

2.1. Multi-level Representation

Given a product image I and a textual query T, we first learn multi-level representations of image and text such that cross-modal similarities of different levels can be simultaneously learned. Through the multi-level encoding network, the image I and the sentence T are represented by a sequence of multi-level image features $\{V^1, V^2, ..., V^L\}$ and a sequence of multi-level text features $\{Q^1, Q^2, ..., Q^L\}$, L is the number of total levels.

2.1.1. Image Representation

For the given product image I, a pre-trained object detection model is employed to detect main objects in the image, producing n detected objects $V^0 : \{v_k^0\}_{k=1}^n$, where $v_k^0 \in \mathbb{R}^{d_0}$ is the extracted feature vector of k-th object. Before feeding them into the first encoder, we use a linear layer to project object features to d_c-dimensional feature vectors which fit the input size of the following encoders. In order to obtain sufficient representation, we stack multiple encoders and utilize all their outputs as the multi-level representations. More concretely, by employing an encoder on the initial object features $V^0 : \{v_k^0\}_{k=1}^n$, we obtain the level-1 representation:

$$V^1 = \phi_1(V^0), \tag{1}$$

where ϕ_1 denotes the first image encoder. To capture the dependencies between objects, we utilize transformer [14] as our fundamental encoder which has been found effective in various tasks [17, 18] due to its superior ability to model sequential relation. Note that other sequential models such as GRU [19] or LSTM [20] can also be used as the encoder. Similarly, the outputs of the following encoders are represented by:

$$V^l = \phi_l(V^{l-1}), l = 2, ..., L, \tag{2}$$

where ϕ_l indicates l-th image encoder, L is the number of stacked encoders. Finally, through L stacked encoders, the multi-level representations of image I are obtained as a sequence of feature groups $\{V^l : \{v_k^l\}_{k=1}^n\}_{l=1}^L$.

2.1.2. Text Representation

Given a textual query T of m words, we first embed each word into a word vector space by GloVe word2vec [21], resulting in a sequence of word features $Q^0 : \{q_k^0\}_{k=1}^m$. Similar to the image encoder, a sequence of word features are then fed into a linear layer to change its dimension, followed by L stacked encoders:

$$Q^l = \psi_l(Q^{l-1}), l = 1, 2, ..., L, \tag{3}$$

where ψ_l denotes l-th text encoder. Finally, through L stacked encoders, the query T can be represented as $\{Q^l : \{q_k^l\}_{k=1}^m\}_{l=1}^L$.

2.2. Multi-granularity Similarity

Given both multi-level representations of images and textual queries, we propose to use multi-granularity similarity to measure their cross-modal similarity. For image and query representations at each level, we compute their object-granularity similarity and image-granularity similarity respectively. In what follows, we describe how to compute these two similarities based on the level-l image representation $V^l : \{v_k^l\}_{k=1}^n$ and query representation $Q^l : \{q_k^l\}_{k=1}^m$.

2.2.1. Object-granularity Similarity

To compute the object-granularity similarity, we learn to map object and query features into a common object-query embedding space, where the object-query similarity can be directly measured. The final object-granularity similarity between an image and a query is obtained by aggregating the object-query similarities of all objects in the image. Specifically, we first aggregate the word representations $\{q_k^l\}_{k=1}^m$ of level l into a query-level feature vector by mean pooling. A linear layer is further employed to project it into the object-query common embedding space, denoted as $\bar{q}_1^l$. For n object features $\{v_k^l\}_{k=1}^n$ of level l in the given image, we also employ a linear layer to transform them into the object-query embedding space, denoted by $\{\bar{v}_k^l\}_{k=1}^n$. Finally, the object-granularity cross-modal similarity between the image I and query T is computed as the average of similarities of all object-query pairs:

$$\sigma_{obj}^l(I, T) = \frac{1}{n}\sum_{k=1}^{n} r(\bar{v}_k^l, \bar{q}_1^l), \tag{4}$$

where $r(,)$ denotes the similarity function. In our implementation, as we perform text-to-image retrieval, we utilize projection length of the textual query feature onto the visual feature [22]:

$$r(\bar{v}_k^l, \bar{q}_1^l) = \bar{v}_k^l \cdot \bar{q}_1^l / \|\bar{v}_k^l\|. \tag{5}$$

2.2.2. *Image-granularity Similarity*

As for image-granularity similarity, we focus on global similarity between images and queries. Similarly, we map image and query representations to a common image-query embedding space. For textual queries, we aggregate the word-level feature vectors $\{q_k^l\}_{k=1}^m$ by mean pooling, followed by another linear layer to obtain the sentence-level feature $\bar{q}_2^l$. Note that the linear layer here does not share weights with that in the object-granularity similarity computation, as we are actually learning two distinct embedding spaces for two similarities. As product images generally contain abundant objects, it is necessary to capture salient features. In such consideration, we employ an attention module to aggregate object features into an image-level feature vector. Specifically, the attention weight for each object is computed by a multi-layer perceptron with one hidden layer. Formally, the attention weight for k-th object feature v_k^l at level l is computed as:

$$w_k^l = W_2\delta(W_1 v_k^l + b_1) + b_2, \tag{6}$$

where $\delta(\cdot)$ is the ReLU nonlinear activation, W_1, W_2 and b_1, b_2 denote the transformation matrices and biases respectively. Besides, we use a softmax layer to normalize the attention weights. With the learned attention weights $\{w_k^l\}_{k=1}^n$, the image-level feature is obtained as a weighted sum of all object features:

$$\bar{v}^l = \sum_{k=1}^{n} w_k^l v_k^l. \tag{7}$$

Finally, we define the image-granularity similarity between the query T and the image I at level l as:

$$\sigma_{img}^l(I,T) = r(\bar{v}^l, \bar{q}_2^l). \tag{8}$$

2.3. Training and Evaluation

2.3.1. *Model Training*

To train the model, we use the cross-modal projection matching loss [22], which aims to learn a common space where the similarity between relevant pairs are forced to be greater than irrelevant pairs in a contrastive manner. Different from [22] which only utilizes the loss over the final output of models, we employ the loss not only over the multi-level representations but also over the multiple granularities. Specifically, at each level, we employ the loss on both object-granularity similarity σ_{obj}^l and image-granularity similarity σ_{img}^l. Formally, given a mini-batch with N relevant image-query pairs, the loss on the object-granularity similarity at the level l is:

$$\mathcal{L}_{obj}^l = \frac{1}{N}\sum_{i=1}^{N} -log\frac{exp(\sigma_{obj}^l(I_i,T_i))}{\sum_{j=1}^{N} exp(\sigma_{obj}^l(I_i,T_j))}, \tag{9}$$

where (I_i, T_i) indicates relevant image-query pair, (I_i, T_j) denotes irrelevant image-query pair if $i \neq j$. Similarly, the loss on the image-granularity similarity is defined as:

$$\mathcal{L}_{img}^l = \frac{1}{N}\sum_{i=1}^{N} -log\frac{exp(\sigma_{img}^l(I_i,T_i))}{\sum_{j=1}^{N} exp(\sigma_{img}^l(I_i,T_j))}. \tag{10}$$

As we employ the loss over all the representations of distinct levels, the final overall loss of a model with L-level representations is:

$$\mathcal{L} = \sum_{l=1}^{L} \lambda_l(\mathcal{L}_{obj}^l + \mathcal{L}_{img}^l). \tag{11}$$

where λ_l is a hyper-parameter of level l which control the balance between different levels.

2.3.2. *Evaluation*

After the model being trained, we measure the similarity between product images and textual queries in terms of multi-level representations and multiple granularities. Concretely, given a product image I and a textual query T, their hierarchical similarity $s(I,T)$ is obtained by:

$$s(I,T) = \sum_{l=1}^{L} \lambda_l(\sigma_{obj}^l(I,T) + \sigma_{img}^l(I,T)). \tag{12}$$

With the hierarchical similarity, given a textual query, all candidate product images are ranked according to their similarities with the given query in descending order.

3. EXPERIMENTS

3.1. Setup

3.1.1. *Dataset and Metric*

We conduct experiments on the dataset of KDD Cup 2020 Challenges for Modern E-Commerce Platform: Multimodalities Recall [23], a large-scale dataset for language-based product image retrieval. The dataset is comprised of a training set of 3 million product images, a validation set of 9177 product images, and two test sets. As the ground-truth of the two test sets are not open-released, we report performance on the validation set. Each training image is annotated with a textual phrase or a sentence which describes the specific product in the image. In the validation set, there are 496 textual queries, and each query is along with a candidate pool of around 30 product images.

Following the evaluation protocol of the dataset, we use the nDCG@5 as the performance metric. We also report the performance of nDCG@k (k=10,15,20,25,30) to obtain a more comprehensive evaluation.

3.1.2. *Implementation Details*

We use PyTorch as our deep learning environment. For object detection, we directly use the detected objects and extracted features provided by the dataset. For the image encoder, we utilize a 3-layer transformer with 4 heads and hidden dimension of 512. For textual query pre-processing, we first convert all words to the lowercase and then replace words that occur less than 5 times in the training set with a special token $\langle unk \rangle$. For the text encoder, we use a 2-layer transformer with the same structure as the image encoder. We also use one-layer bi-GRU as per-level encoder with dimension of 512. Both images and queries are embedded into space of 1024 dimension. For the model loss, we empirically set λ_1 to 0.5 and λ_2 to 1. To train the model, we utilize Adam optimizer[24]. The initial learning rate is set to 2e-4, decayed by 0.1 every epoch. The network is totally trained for 30 epochs. At each iteration, 256 query-product pairs are randomly sampled.

3.2. Comparison with the State-of-the-art

To verify the viability of our proposed model, we compare it with two groups of works: one group consists of general image-text retrieval methods, the other are methods particularly designed for language-based product image retrieval. The results are summarized in Table 1. Our proposed model HSL with transformer encoder obtains nDCG@5 score of 0.7488, which outperforms the two groups

Table 1. Performance comparison with state-of-the-art models. Our proposed HSL with the transformer as encoders performs the best. For a fair comparison, all the scores are obtained by single model without model ensemble.

Method	nDCG@5
SCAN[25]	0.5609
VSE0[26]	0.6381
VSE++[26]	0.6494
LXMERT+LightGBM [10]	0.6200
MCAN[12, 9]	0.6900
VisualBERT[13, 9]	0.7100
Word-based Model[11]	0.7290
HSL(bi-GRU)	0.6846
HSL(transformer)	**0.7448**

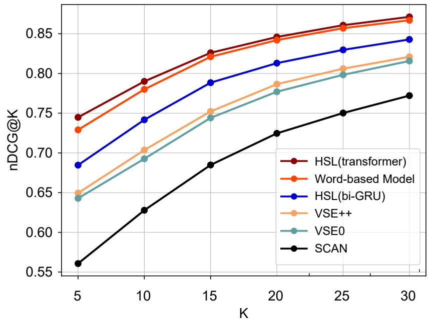


Fig. 3. Performances comparison in terms of varied k of nDCG@k.

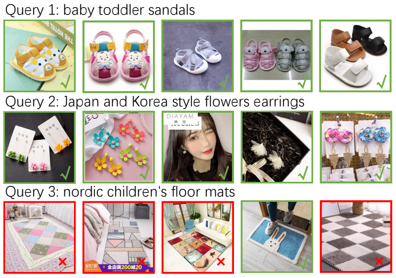

Query 1: baby toddler sandals
Query 2: Japan and Korea style flowers earrings
Query 3: nordic children's floor mats

Fig. 4. For each textual query, top-5 product images sorted in descending order of similarity are presented. Green bounding box indicates correct one, while red ones are incorrect.

of methods with a clear margin. The result shows the effectiveness of HSL for language-based product image retrieval.

Among the first group, SCAN measures the image-text similarity in terms of the object granularity, while both VSE0 and VSE++ only consider the image-granularity similarity. As these three models utilize GRU-based text encoding, we replace the transformer encoders in HSL with bidirectional GRU (bi-GRU) to make the comparison fairer. HSL equipped with bi-GRU outperforms these three methods with a clear margin. It shows the importance of multi-granularity similarity for product image retrieval. Among the second group, all models employ transformers to encode input data, but they only consider specific granularity similarity at high level. The better performance of HSL verifies the effectiveness of our hierarchical similarity learning framework. Additionally, we also report $nDCG@k$ of varied k in Fig. 3. Our HSL consistently outperforms the other counterparts.

Fig. 4 displays some qualitative results of our proposed HSL. HSL performs well for top 2 queries, while bad for Query 3 of *nordic children's floor mats*. Although top-5 retrieved images of Query 3 are all mat products, our model can not distinguish which mats are for children and which are not thus give the relatively bad result.

Table 2. Ablation study of HSL. The ✗symbol indicates model without multiple-level representation or multi-granularity similarity, and used single level or granularity are specified in parenthesis.

Multi-level Representation?	Multi-granularity Similarity ?	nDCG@5
✗(Level 1)	✗(Object)	0.7275
	✗(Image)	0.7240
	✓	0.7339
✗(Level 2)	✗(Object)	0.7351
	✗(Image)	0.7273
	✓	0.7362
✓	✗(Object)	0.7412
	✗(Image)	0.7381
	✓	**0.7448**

3.3. Ablation Study

Table 2 summarizes the results of the ablation study. It can be seen that HSL suffers the performance degeneration when a single-level representation is used, which shows the effectiveness of our multi-level representations. Among the similarity granularity, using the object-granularity similarity performs better than the image-granularity one, and multi-granularity similarities considering both granularities achieve the best result. The result shows the complementarity of the object-granularity similarity and the image-granularity similarity. Moreover, our full model with both multi-level representations and multi-granularity similarities gives the best nDCG@5 score of 0.7448, which further verifies the importance of our proposed hierarchical similarity learning for language-based product image retrieval. We also try HSL with more than two levels, while find that using more levels does not lead to better performance.

4. CONCLUSION

This paper proposes a hierarchical similarity learning network for language-based product image retrieval task. Different from existing works that consider single-granularity similarity based on the representation of a specific level, we compute multi-granularity similarities based on multi-level representations in a hierarchical manner. Experiments on a large-scale product image retrieval dataset verify the viability of our model for language-based product image retrieval, and the ablation study shows the importance of both multi-level representations and multi-granularity similarities. In the future, we would like to explore our model for general cross-modal retrieval tasks, such as text-to-video retrieval [27, 28, 29].

5. ACKNOWLEDGMENTS

This work was partly supported by the National Key Research and Development Program of China under No. 2018YFB0804102 and No. 2020YFB2103802, NSFC under No. 61902347, No. 61772466, U1936215, and U1836202, the Zhejiang Provincial Natural Science Foundation under No. LQ19F020002 and No. LR19F020003, the Zhejiang Provincial Public Welfare Technology Research Project under No. LGF21F020010 and LGG20F020005, Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies, and the Fundamental Research Funds for the Central Universities (Zhejiang University NGICS Platform).

6. REFERENCES

[1] Yuxin Peng, Xin Huang, and Yunzhen Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Transactions on circuits and systems for video technology*, vol. 28, no. 9, pp. 2372–2385, 2017.

[2] Jianfeng Dong, "Cross-media relevance computation for multimedia retrieval," in *ACM Multimedia*, 2017, pp. 831–835.

[3] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *ACM MM*, 2010, pp. 251–260.

[4] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang, "Dual encoding for zero-example video retrieval," in *CVPR*, 2019, pp. 9346–9355.

[5] Zhe Ma, Jianfeng Dong, Yao Zhang, Zhongzi Long, Yuan He, Hui Xue, and Shouling Ji, "Fine-grained fashion similarity learning by attribute-specific embedding network," in *AAAI*, 2020, vol. 34, pp. 11741–11748.

[6] Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth, "Learning type-aware embeddings for fashion compatibility," in *ECCV*, 2018, pp. 390–405.

[7] Xun Yang, Xiangnan He, Xiang Wang, Yunshan Ma, Fuli Feng, Meng Wang, and Tat-Seng Chua, "Interpretable fashion matching with rich attributes," in *SIGIR*, 2019, pp. 775–784.

[8] Yanbei Chen, Shaogang Gong, and Loris Bazzani, "Image search with text feedback by visiolinguistic attention learning," in *CVPR*, 2020, pp. 3001–3011.

[9] KueiChun Huang, ChiYu Yang, and KenYu Lin, "KDD_WinnieTheBest," `https://github.com/steven95421/KDD_WinnieTheBest`, 2020, [Online; accessed 20-10-2020].

[10] Qi Zhang, Lixiang Wang, Changyu Li, Peng Zhang, Shengyuan Zheng, Kai Wang, Yudong Xu, Lei Zhong, Chenyu Jin, and Jingjie Li, "KDD2020_mutilmodalities," `https://github.com/miziha-zp/KDD2020_mutilmodalities`, 2020, [Online; accessed 20-10-2020].

[11] Yuhui Ding, Lei Wu, Chengxi Li, Jieyu Yang, and Yue Wang, "KDD-Cup-Multimodalities-Recall," `https://github.com/dingyh0626/KDD-Cup-Multimodalities-Recall`, 2020, [Online; accessed 20-10-2020].

[12] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian, "Deep modular co-attention networks for visual question answering," in *CVPR*, 2019, pp. 6281–6290.

[13] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang, "VisualBERT: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NeuralPS*, 2017, pp. 5998–6008.

[15] Hao Tan and Mohit Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019.

[16] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015, pp. 3156–3164.

[17] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeuralPS*, 2019, pp. 13–23.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[20] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "GloVe: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.

[22] Ying Zhang and Huchuan Lu, "Deep cross-modal projection learning for image-text matching," in *ECCV*, 2018, pp. 686–701.

[23] "KDD Cup 2020 Challenges for Modern E-Commerce Platform: Multimodalities Recall," `https://tianchi.aliyun.com/competition/entrance/231786/information`, [Online; accessed 20-10-2020].

[24] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He, "Stacked cross attention for image-text matching," in *ECCV*, 2018, pp. 201–216.

[26] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," in *BMVC*, 2018.

[27] Xirong Li, Chaoxi Xu, Gang Yang, Zhineng Chen, and Jianfeng Dong, "W2vv++ fully deep learning for ad-hoc video search," in *ACM Multimedia*, 2019, pp. 1786–1794.

[28] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua, "Tree-augmented cross-modal encoding for complex-query video retrieval," in *SIGIR*, 2020, pp. 1339–1348.

[29] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang, "Dual encoding for video retrieval by text," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.