Heterogeneous Collaborative Refining for Real-Time End-to-End Image-Text Retrieval System

Nan, Guo*
Institute of Computing Technology, Chinese Academy of Sciences

Min, Yang
Institute of Computing Technology, Chinese Academy of Sciences

Xiaoping, Chen
Institute of Computing Technology, Chinese Academy of Sciences

Xiao, Xiao
Institute of Computing Technology, Chinese Academy of Sciences

Chenhao, Wang
Institute of Computing Technology, Chinese Academy of Sciences

Xiaochun, Ye
Institute of Computing Technology, Chinese Academy of Sciences

Dongrui, Fan
Institute of Computing Technology, Chinese Academy of Sciences

ABSTRACT

The image-text retrieval task currently suffers from high search latency due to the cost of image feature extraction and semantic alignment calculation. We propose a real-time image-text retrieval system for edge-end servers with low-power AI accelerator cards. The procedure is conspicuously sped up by selectively placing part of the deep learning calculation on accelerator devices with a heterogeneous collaborative computation scheme. We also design a lightweight GCN optimization method, which directly transfers the correlation between the image detection areas in projection to reduce computational redundancy. Our other contributions include performance analyses of models with different weights for industrial reference in practical applications. It is the first GCN-based image-text retrieval system to perform a real-time end-to-end search to the best of our knowledge. Experiments show that the system can process 20 image-to-text retrievals per second with high accuracy.

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Computer vision; Computer vision tasks; Visual content-based indexing and retrieval.

KEYWORDS

Image-text retrieval, GCN, Accelerator card, Edge server

ACM Reference Format:
Nan, Guo*, Min, Yang, Xiaoping, Chen, Xiao, Xiao, Chenhao, Wang, Xiaochun, Ye, and Dongrui, Fan. 2022. Heterogeneous Collaborative Refining for Real-Time End-to-End Image-Text Retrieval System. In *2022 the 6th International Conference on Innovation in Artificial Intelligence (ICIAI) (ICIAI 2022), March 04–06, 2022, Guangzhou, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3529466.3529486

ICIAI 2022, March 04–06, 2022, Guangzhou, China

ACM ISBN 978-1-4503-9550-2/22/03.
https://doi.org/10.1145/3529466.3529486

1 INTRODUCTION

Cross-modal retrieval is to find the relationship between different modal samples and realize the use of a specific modal piece to search for other modal instances with similar semantics. Image-text retrieval is one of the primary themes of cross-modality. It refers to inputting an image to retrieve several texts or using one text to retrieve several images with similar semantics. The image-text retrieval task faces two challenges: improving the retrieval recall and reducing the calculation load without sacrificing accuracy. There are a large number of scenarios for lightweight applications of cross-modal retrieval. Especially in the field involving data security, these scenarios require deploying models on the edge-end under actual requirements, not allowed to cloud servers. The servers at the edge end do not have high-power GPUs as cloud servers but are superior to the embedded system on the things end computing. However, existing image-text retrieval works have not thoroughly studied the problem of latency and accuracy at the edge servers.

We consider optimizing the system design from hardware and software to improve the retrieval performance. Hardware researchers commit to studying deep learning accelerator devices to strengthen the inference speed of neural networks, such as general parallel computing devices: GPU, TPU, and specified AI accelerate cards: VPU, Cambrian, Huawei Atlas, etc. Dedicated cards for edge-end servers with low power consumption serve an excellent purpose for image classification, object detection, face recognition, and so on. Still, they have limited support for complex applications with several modules recently, for example, cross-modal process.

The software researches for Deep Neural Network (DNN) include lightweight model acceleration. One fundamental problem of the image-text retrieval task is to measure the visual-semantic similarity. It needs to align the global context of the image with the words in the text and also to align different objects in the regional features. The image-text retrieval system with Graph Convolution Network (GCN) achieves these correspondences through feature extraction and semantic reasoning. Existing image-text retrieval systems have not studied the latency of cross-modal systems based on GCN, and the general GCN implementation has redundancy in time complexity and memory usage. Therefore, we accelerate the semantic reasoning process based on a lightweight GCN, a simple and efficient design to meet actual application requirements.

This paper proposes a low-power deployment and GCN-based low-latency feature extraction network to achieve a real-time end-to-end image-text retrieval system. Compared with commercial GPUs, AI accelerator cards have the advantages of low energy consumption, small size, and low cost. Thus, we propose that the image feature detection part is inferred on the accelerator card to improve the system's overall efficiency. By directly transferring the correlation between the image detection areas in projection, the lightweight GCN achieves lower latency than the original back-propagation design on learning multiple weight parameters for relationship expression. Refine-ITR (heterogeneous collaborative Refining for real-time end-to-end Image-Text Retrieval system) is the first end-to-end image-text retrieval system designed for edge servers to the best of our knowledge. It achieves a 6x reduction in reference time with satisfactory accuracy. The main contributions of this paper are:

- Design a real-time end-to-end image-text retrieval system on heterogeneous edge-end servers;
- Propose a lightweight GCN optimization method for the semantic reasoning process;
- Give a detailed analysis of accuracy and speed comparisons before and after acceleration of the existing image-text retrieval system based on GCN, obtained as an end-to-end system speed of more than 6x and reaches satisfying accuracy.

2 RELATED WORK

2.1 Image-Text Retrieval

The image-text relations in retrieval systems usually are implemented by feature extraction and semantic relational reasoning. Feature extraction includes text feature and image feature extractions. Text features can be encoded by Long Short-Term Memory (LSTM) [1], Gated Recurrent Unit (GRU) [2], and transformer-based BERT [3], among which GRU has relatively lower computation cost. Image features mainly include global level, regional level, and multi-level representations. In general, Convolutional Neural Networks(CNN) trained in the image classification task, such as VGG [4] and ResNet [5], are employed to extract global descriptors[6] [7] [8]. Pre-trained object detectors, for example, Faster R-CNN [9], are utilized for region-level features [10] [11] [12]. Multi-level methods use global and local features to improve accuracy, such as MDM [13]and GSLS [14], which design two paths for separate global and local similarity calculations. The required calculation of the above image feature extraction processes based on CNN is considerably time-consuming. The selection of different models greatly impacts the retrieval recall.

Graph Neural Network (GNN) [15] can process the graph structure directly to promote a recursive neural network. VSRN [11] applies GNN to visual reasoning and learns the relations among image regions that correspond to relations of words in the sentences, which significantly improves the accuracy of cross-modal retrieval. Subsequent development on visual semantic reasoning, such as DSRAN [16], which uses multiple graph attention networks (GAT) to enhance object-wise relations and object-global-wise relations, and transformer-based methods [17] [18] [19] [20] further improve the recall rate. However, the computational cost is also increased evidently at the same time. GAT uses self-attention mechanisms, and a transformer encoder is composed of several multi-head attention layers, which contain a considerable amount of attention modules. The atom operation of the self-attention mechanism [21] is canonical dot-product, which causes the time complexity and memory usage per layer to be $O(L^2)$. On the other hand, GCN only uses the dot product in information transmission with much less the attention part. It is imperative to optimize the semantic reasoning process, which takes up considerable computing resources, and GCN-based relational reasoning is the best choice for less computation.

While most image-text retrieval researches focus on improving accuracy with complex algorithm designs, a few have proposed efficient computing. LightningDOT [18] offers real-time image-text retrieval by extracting feature indexes offline to accelerate the inference time, which is not an end-to-end usage. [19] presents a real-time retrieval model specialized for text-to-image, which does not extract image features online. No study has attempted to analyze the computational efficiency of end-to-end image-to-text retrieval systems. So we propose end-to-end research with a series of hardware and software optimizations.

2.2 System Acceleration

[22] presents an alternative approach to enable the efficient execution of DNNs on embedded devices. It dynamically determines which DNN to use by the desired accuracy and inference time. [23] trades off the execution time and memory consumption for ahead-of-time domain-specific optimization of CNN models. It uses integer linear programming for selecting primitive operations to implement convolutional layers. However, current architecture searching methods [24] do not consider the model transformation difficulty and the supporting characteristics of high-throughput data. So we use the manual design at the present stage and will automate the search in the future.

Optimization methods for improve computing efficiency include network pruning [25], quantization [26], calculation unit optimization [27] etc. The above methods mainly accelerate the general deep learning methods. Meanwhile, it also comes at the cost of reducing model accuracy. Differently, we aim to eliminate computational redundancy in the execution process so that the experimental accuracy is almost unaffected.

3 THE IMAGE-TEXT RETRIEVAL SYSTEM OPTIMIZATION

Current image-text retrieval works mainly focus on algorithmic processing after image features. The input is image features, which are often obtained through offline models and then online reasoning. However, in practical applications, image-text retrieval requires high end-to-end efficiency, low power consumption, and reliable accuracy during the inference process. Therefore, we propose to implement the end-to-end implementation of image-text retrieval tasks on the AI accelerator card and CPU, which generally meet the configuration requirements of the edge server. We design a heterogeneous collaborative scheme to selective assign computation to low-power accelerator cards and present a lightweight GCN encoding method.

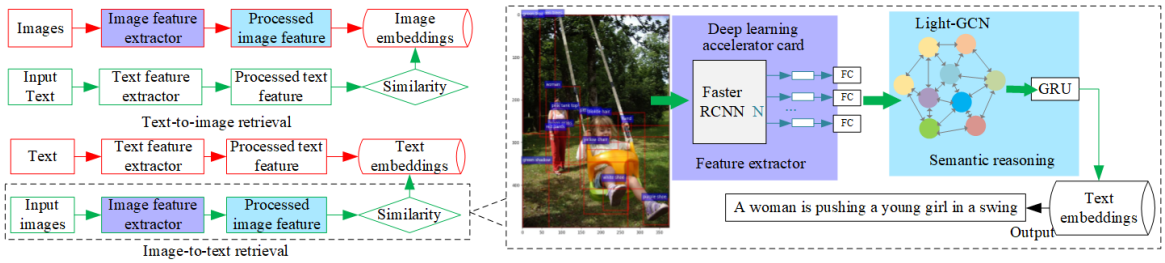

Figure 1: An overview of the proposed image-text retrieval system. The red box represents the offline processing, and the green box represents the online retrieval process. The purple-filled area is the image feature extractor module processed on AI accelerator cards. The blue area is the semantic reasoning process.

3.1 Heterogeneous Collaborative Design for Inference System

Local and global feature analysis [16] shows that local features are better than global parts. Under an appropriate semantic relation learning module, the use of the two features only improves the accuracy slightly but increases quite a lot of computation. Therefore, our system only uses local object-level detected features according to Bottom-to-up [28], which the attention module uses Faster R-CNN to achieve. Object detection generates the features of multiple detection areas in a single image. Next is the semantic reasoning module, including the following three parts: 1) to establish the connection between the elements of these detection areas using the lightweight GCN proposed in this paper. 2) to establish a relationship between vision and text and use a GRU-based text encoder to map text representations to the exact dimensions as image features. 3) the two kinds of features jointly match after embedding the text and the image separately. Finally, sorting the matching results obtains the optimal image and text alignment.

Figure 1 shows the end-to-end image-text retrieval system framework in this paper. The red box represents the offline process, and the green box represents the online retrieval process. The purple-filled area is the image feature extraction module, usually with significant computation. The blue-filled site is the semantic reasoning module for the image embedding, which also occupies a large proportion of calculation and is second only to the purple area. The image embedding is extracted offline in the text-to-image search flow, while the image-to-text retrieval process needs to compute image descriptors online. Due to the heavy image embedding calculation, the image-to-text retrieval has a noticeable time delay.

Our framework combines the AI accelerate card and CPU cooperative computing scheme to allocate the two structures in image-text retrieval: CONV and without CONV.

The object detection module based on Faster-RCNN has a fair amount of classical convolution calculation, in line with the design philosophy of AI cards. Considering the advantages of low energy consumption, small size, and low cost of existing AI accelerator cards on the market compared with available GPUs, this system's image feature extraction is performed on the AI accelerator card in priority. The detection model is trained on the cloud and transferred to the inference card.

The system implements computing elements except for image feature extraction on the CPU. The accelerator card is dedicatedly designed for neural networks. The support for other aspects of operators is less concise due to power and area limitations, for example, canonical dot-product, one of the primary computing units of GCN and GRU. What's more, in practical applications with high-throughput data processing, image feature extraction requires large quantities of computation and can occupy almost all the resources of accelerator devices. Therefore, the modules that populate fewer computational resources, such as text feature processing and semantic reasoning, are generally executed on the CPU.

After allocating computing modules on different hardware, our heterogeneous collaborative design puts forward mean time equalization and batch size adjustment to use computing resources fully. DNN inference executes on the AI accelerator card and CPU sequentially. It can result in wasted computing resources when calculations are waiting. Making full use of computing resources is critical to improving system performance. To achieve high hardware parallelism, first, it is necessary to choose appropriate hardware and distribute computing to make the average calculation time of the feature extraction module and relationship alignment module roughly equal. Second, The closer and smaller the batch sizes of the two modules are, the smaller the total system delay is.

3.2 Lightweight GCN

The image-text retrieval system VSRN [11] first establishes a relationship reasoning model with semantic relevance between image detection areas. The method adopted is to calculate the distance of the features in the embedded space. Specifically, a fully connected relationship graph is constructed by the distance between each pair of feature vectors, as shown in the following equations,

$$R\left(v_i, v_j\right) = \phi(v_i)^T \theta\left(v_j\right) \tag{1}$$

$$V^* = W_r\left(RVW_g\right) + V \tag{2}$$

where V = $\{v_1, \cdot s, v_k\}$, $v_i \in R$, $i = 1, 2, \cdot s, k$. is the set of detected regions. $\phi(v_i) = W_\phi v_i$, $\theta(v_j) = W_\theta v_j$ are the relationship enhanced representations. The reasoning Equations 1) and (2) are contained in the embedding stage of the image-text retrieval system. The whole reasoning stage suffers from time-consuming. We analyze through experiments that the time for processing each set of region features in one image is about 1.3 milliseconds, mainly because it contains multiple 2048x2048 high-dimensional non-sparse matrix multiplication operations.

Inspired by this, we construct a lightweight reasoning mechanism to solve the problem of the image feature reasoning module taking up a too long time. Therefore, we propose to directly measure the projection and distance of the paired feature vectors of the image areas to describe their relationship, which can enhance the semantic representation based on image detection areas more targeted with fewer parameters. The newly proposed measurement halved the image feature embedding time is less than the sub-milliseconds shown in experiments. The specific Equation is as follows,

$$V^*_{light} = W_r \left(\sigma \left(VV^T \right) VW_g \right) + V \tag{3}$$

where $\sigma(VV^T) = W_\sigma VV^T$ is the feature embedding between pairs of image regions, and W_σ can be learned by backpropagation of parameters. W_σ is the weight matrix for learning the cross-relationship between image regions, and the dimension is K x K. The dimension of the vector representation of the image detection areas V is K x D. W_g is the weight parameter that the graph convolution layer needs to learn, and the weight dimension is D x D . W_r is the weight matrix that the residual structure needs to learn. $\sigma(VV^T)$ is a lightweight affinity matrix. The output $V^*_{light} = \{v^*_1, \cdot s, v^*_k\}$, $v^*_i \in R$ is the node feature representation that is enhanced by the lightweight feature reasoning network for the relationship among image detection regions.

The relationship graph G is established by building a fully connected relationship between dozens of detection areas in an image. That is, each detection area serves as the nodes in the graph, and $\sigma(VV^T)$ represents the connected relationship between the nodes in the forward propagation equation. The connection between the detected regions V and the area relationship $\sigma(VV^T)$ is set up further through the above formula to establish an enhanced feature expression that fuses the relationship between fully connected adjacent regions.

Equation 3) helps to achieve a low-latency feature semantic reasoning network. The low-latency feature semantic processing directly transfers the correlation between the image detection areas in projection to the GCN instead of the original backpropagation to learn the feature relationship expression form with multiple weight parameters.

4 EXPERIMENTS AND ANALYSIS

4.1 Platform Parameters

We evaluate our method on the Flickr30K [29] and MS-COCO [30] datasets. Flickr30K consists of 31,783 images collected from the Flickr website. Five human-annotated text descriptions accompany each image. We use the standard training, validation, and testing split, containing 28,000 images, 1000 images, and 1000 images, respectively. MS COCO is a large-scale multi-task dataset. We use the image captioning dataset split. MS-COCO consists of 123,287 images, and every image has five description captions. We use the training, validation, and testing splits, containing 113,287 images, 5000 images, and 5000 images, respectively. The final results are obtained by averaging over five folds of 1000 test images.

We measure the performance by the recall at K (R@K) for the evaluation matrix, defined as the fraction of queries for which the correct item is retrieved in the closest K points to the query. We develop the system with an AMD CPU EPYC 7702 and 6 Huawei Atlas 300 accelerator card, which works with the FP-32 model.

4.2 GCN Acceleration

We set the word embedding size to 300 and the dimension of the joint embedding space D to 2048. The GCN is trained for 30 epochs with Adam optimizer, and the initial learning rate is 0.0002. We use a mini-batch size of 128 and update the learning rate at every ten epochs.

To demonstrate the efficiency of the lightweight GCN (Light-GCN), we compare it with the GCN in VSRN, which is the most similar method to our system. Table 1 shows the quantitative evaluation of the VSRN-GCN and Light-GCN on the image-text retrieval tasks. We test the recalls and time costs with the above AMD CPU. With the lightweight GCN optimization, the image feature extraction time is reduced from 44.6ms to 27.8ms, with a 37.7% reduction. At the same time, all four recall results (R@1, R@5, R@10, and average recall) of the two retrieval tasks are almost equal to the VSRN-GCN.

Accuracy comparison with state-of-the-art methods. The image feature extraction method plays an essential role in retrieval. The compared state-of-the-art methods include global image features, local region features, and both. All these excellent approaches do not involve attention mechanisms, which are time and calculation consuming that are not suitable for applications on the edge servers at the current stage of attention research. Experiments show that our system performs well when dealing with large datasets such as the MS-COCO and small datasets such as the Flickr30K. Table 2 presents that all 12 recall values of our Light-GCN are among the highest in both text-to-image and image-to-text search tasks.

The quantitative results show that the recall rate of text-to-image retrieval is lower than that of image-to-text searching. The reason is that each image in the dataset corresponds to 5 texts. Images are more accessible to match ground-truth texts. Figure 2 shows the Light-GCN qualitative results of image retrieval given text queries and text retrieval given image queries on the Flickr30K dataset. We outline the corresponding images in red and unmatched images in green boxes. And the match texts are in red, and the unmatched texts are in green.

4.3 The Influence of Image Feature Size

We apply several parameter configurations to reduce memory and speed up inference to mimic the real-life scenario, which needs to meet different precision, memory, and time delay. We test three dimensions of the image feature: 512, 1024, and 2048, to balance accuracy and runtime on the Flickr30K dataset. Features with large sizes have more description data for images, which can improve matching accuracy and increase the computational load of the semantic reasoning module. As Table 3 shows, the feature size of 2048 corresponds to the largest model and the highest accuracy. When the embedding space dimension is 1024, it can reduce the model size by 72.0% and time cost by 72.9% while sacrificing only 3% of average recall. The 1024 dimension image feature setting meets a good balance between precision and speed favorably to the edge AI application. The feature size of 512 can reach the model with only

Table 1: QUANTITATIVE EVALUATION RESULTS OF THE IMAGE-TEXT RETRIEVAL ON FLIKER30K IN TERMS OF RECALL@K(R@K, K=1,5,10) AND EMBEDDING TIME(MS)

Model	Image-to-Text				Text-to-Image				Image relationship reasoning time(ms)
	R@1	R@5	R@10	Ave	R@1	R@5	R@10	Ave	
VSRN-GCN	71.3	90.6	96.0	86.0	54.7	81.8	88.2	74.9	44.6
Light-GCN	71.6	91.3	95.7	86.2	53.5	79.8	87	73.4	27.8

Table 2: ACCURACY COMPARISON WITH STATE-OF-THE-ART METHODS ON MSCOCO AND FLICKR30K. THREE MULTI ROWS DENOTE THE GLOBAL DESCRIPTORS, REGION-LEVEL FEATURES, AND BOTH FEATURES.

	MSCOCO 1K						FLICKR-30K 1K					
	Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
GLOBAL												
VSE++ [6]	64.6	90.0	95.7	52.0	84.3	92.0	52.9	80.5	87.2	39.6	70.1	79.5
MFM [7]	58.9	86.3	92.4	47.7	81.0	90.9	50.2	78.1	86.7	38.2	70.1	80.2
MTFN [8]	74.3	94.9	97.9	60.1	89.1	95.0	65.3	88.3	93.3	52.0	80.1	86.1
LOCAL												
SCAN [10]	72.7	94.8	98.4	58.8	88.4	94.8	67.9	89.0	94.4	43.9	74.2	82.8
CAMP [31]	72.3	94.8	98.3	58.5	87.9	95.0	68.1	89.7	95.2	51.5	77.1	85.3
VSRN [11]	76.2	94.8	98.2	62.8	89.7	95.1	71.3	90.6	96.0	54.7	81.8	88.2
SGM [12]	73.4	93.8	97.8	57.5	87.3	94.3	71.8	91.7	95.5	53.5	79.6	86.5
BOTH												
MDM [13]	54.7	84.1	91.9	44.6	79.6	90.5	44.9	75.4	84.4	34.4	67.0	77.7
GSLS [14]	68.9	94.1	98.0	58.6	88.2	94.9	68.2	89.1	94.5	43.4	73.5	82.5
Light-GCN	75.0	94.9	98.3	60.7	89.1	94.7	71.6	91.3	95.7	53.5	79.8	87.0

Table 3: QUANTITATIVE EVALUATION WITH DIFFERENT DIMENSIONS OF THE IMAGE-TEX RETRIEVAL ON FLIKER30K TEST SET. WE TEST THE TIME ON THE CPU.

Model-emb-size	Image-to-Text				Text-to-Image				Encode-time (ms)	Weights (MB)
	R@1	R@5	R@10	Ave	R@1	R@5	R@10	Ave		
2048_our	71.6	91.3	95.7	86.2	53.5	79.8	87	73.4	48.4	433.2
1024_our	68.8	88.5	93.9	83.7	50.3	76.8	84.8	70.6	13.1	121.5
512_our	64.5	88.3	93.1	82.0	47.4	75.3	83.3	68.6	6.1	40.6

Table 4: RUNNING TIME OF END-TO-END IMAGE-TO-TEXT RETRIEVAL SYSTEMS

Time(ms)	VSRN-GPU	Refine ITR-GPU	Refine ITR- Atlas
Image feature extraction	222.2	222.2	16.8
Semantic reasoning	29.2	21.3	21.3
Whole image-to-text retrieval	253.1	245.2	39.8

less than 50M, which is suitable for scenes with no requirement for high accuracy, but the memory is strictly limited.

4.4 Experiments on End-to-End Image-to-Text Retrieval System

The proposed end-to-end image-to-text retrieval system is tested in two parts: image feature extraction and the rest of image feature-to-text retrieval processing. We use accelerator cards to generate the image detection features and CPU to run other modules. We show the total running time of the end-to-end image-text retrieval system on two different platforms: GTX2080Ti+CPU and Atlas300+CPU, as shown in Table 4. The input images size is 256x256.

We test the image feature extraction time on the Flickr30K test set on two kinds of accelerating devices: 2 GTX2080Ti and 6 Atlas300. The power consumption of 2 GTX2080Ti is 500w, and 6 Atlas300 is 402w, approximately equal power. The average time of one image on 6 Atlas300 cards is 16.8ms compared with 222.2ms on 2 GTX2080Ti,

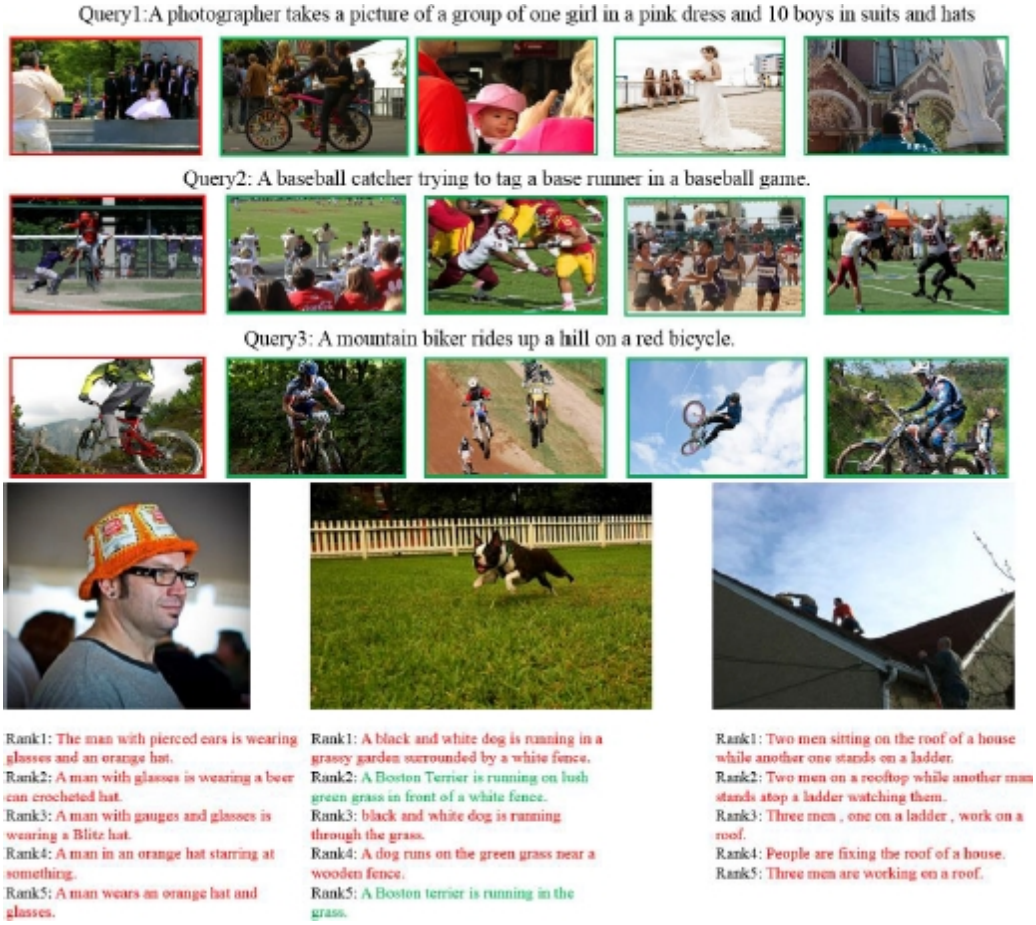

Figure 2: The Light-GCN qualitative results of image retrieval given text queries and text retrieval given image queries on the Flickr30K dataset.

with about 13.2x speedup. Extracting the image feature on Atlas makes the whole time of Refine ITR decrease from 245.2 on GPU to 39.8 on special accelerate devices, 6.2x speed faster. The batch size is 24, with the memory and computing resources fully utilized.

For the module worked on CPU, the feature dimension is 1024, and batch size is also 24, with the computing resource of CPU almost fully occupied. Our optimized system is 6.4x times faster than the VSRN running on the GTX2080Ti. When using multiple acceleration cards, the average processing time on a single server can be less than 40ms. The server can process 20 images within 1 second, realizing "real-time" processing. The searching time of the system is negligible, with 0.78ms for retrieving 1000 images, far lower than the running time of other modules. So it is suitable to implement on larger data sets. The delay time from the image input to retrieval corresponding texts is 0.9s, which most users can accept in the large data query. For keyframe retrieval in videos, the system can process multiple videos concurrently. The pursuit of large processing quantity per unit time is also the characteristic of deep learning computing except for absolute low latency. Since we do not add complex and unique calculation operators, the system can easily extend to similar AI accelerator cards, such as Cambrian v100, and realize the image-text retrieval application in the terminal, mobile vehicle-mounted devices in the future hopefully.

5 CONCLUSION

This paper aims to explore cross-modal applications on the edge server. We propose a real-time end-to-end image-text retrieval system that employs AI accelerator cards to calculate the time-consuming image feature extraction module and presents light-weight GCN optimization to speed up the semantic reasoning process. The heterogeneous collaborative design can improve the efficiency of the whole image-text search system and significantly reduce the online retrieval time of image-to-text processing. Experiments show the absolute speed advantage of the proposed end-to-end image-text search system without sacrificing accuracy compared with previous methods.

ACKNOWLEDGMENTS

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDC05000000), the National Natural Science Foundation of China (Grant No.

61732018, 61872335, and 61802367), the Austrian-Chinese Cooperative R&D Project (FFG and CAS) (Grant No. 171111KYSB20200002), and CAS Project for Young Scientists in Basic Research (Grant No. YSBR-029).

REFERENCES

[1] Sepp Hochreiter and J¨urgen Schmidhuber. Long short-term memory.Neural Computation, 9(8):1735–1780, 1997

[2] Kyunghyun Cho, Bart van Merrienboer, C¸ aglar G¨ulc¸ehre, Dzmitry Bahdanau,Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, October 25-29, Doha, Qatar, 2014.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, Minneapolis, MN, USA, June 2-7, Volume 1, 2019.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In 3rd International Conference on Learning Representations, ICLR, San Diego, CA, USA, May 7-9, 2015.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016.

[6] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improving visual-semantic embeddings with hard negatives. In British Machine Vision Conference, BMVC, Newcastle, UK, September 3-6, 2018.

[7] Lin Ma, Wenhao Jiang, Zequn Jie, Yu-Gang Jiang, and Wei Liu. Matching image and sentence with multi-faceted representations. IEEE Trans. Circuits Syst. Video Technol., 30(7):2250–2261, 2020.

[8] Tan Wang, Xing Xu, Yang Yang, Alan Hanjalic, Heng Tao Shen, and Jingkuan Song. Matching images and text with multi-modal tensor fusion and re-ranking. In Proceedings of the 27th ACM International Conference on Multimedia, MM, Nice, France, October 21-25, 2019.

[9] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6):1137–1149, 2017

[10] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In Computer Vision - ECCV - 15th European Conference, Munich, Germany, September 8-14, Proceedings, Part IV, volume 11208 of Lecture Notes in Computer Science, pages 212–228. Springer, 2018.

[11] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In IEEE/CVF International Conference on Computer Vision, ICCV, Seoul, Korea (South), October 27 - November 2, 2019.

[12] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In IEEE Winter Conference on Applications of Computer Vision, WACV, Snowmass Village, CO, USA, March 1-5, 2020.

[13] Lin Ma, Wenhao Jiang, Zequn Jie, and Xu Wang. Bidirectional image sentence retrieval by local and global deep matching. Neurocomputing, 345:36–44, 2019.

[14] Zhixin Li, Feng Ling, Canlong Zhang, and Huifang Ma. Combining global and local similarity for cross-media retrieval. IEEE Access, 8:21847–21856, 2020.

[15] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. IEEE Trans. Neural Networks, 20(1):61–80, 2009.

[16] Keyu Wen, Xiaodong Gu, and Qingrong Cheng. Learning dual semantic relations with graph attention for image-text matching. IEEE Transactions on Circuits and Systems for Video Technology, pages 1–1, 2020.

[17] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision language tasks. In Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020.

[18] Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021.

[19] Xiaopeng Lu, Tiancheng Zhao, and Kyusong Lee. Visualsparta: Sparse transformer fragment-level matching for large-scale text-to-image search. CoRR, abs/2101.00265, 2021, unpublished.

[20] Nicola Messina, Fabrizio Falchi, Andrea Esuli, and Giuseppe Amato. Transformer reasoning network for image- text matching and retrieval. In 25th International Conference on Pattern Recognition, ICPR, Virtual Event / Milan, Italy, January 10-15, 2020.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, Long Beach, CA, USA, 2017.

[22] Vicent Sanz Marco, Ben Taylor, Zheng Wang, and Yehia Elkhatib. Optimizing deep learning inference on embedded systems through adaptive model selection. ACM Transactions on Embedded Computing Systems, 19(1):2:1–2:28, 2020.

[23] Yuan Wen, Andrew Anderson, Valentin Radu, Michael F. P. O'Boyle, and David Gregg. TASO: time and space optimization for memory constrained DNN inference. In 32nd IEEE International Symposium on Computer Architecture and High Performance Computing, SBAC-PAD, Porto, Portugal, September 9-11, 2020.

[24] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. Mnasnet: Platform-aware neural architecture search for mobile. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Long Beach, CA, USA, June 16-20, 2019.

[25] Yang He, Ping Liu, Ziwei Wang, Zhilan Hu, and Yi Yang. Filter pruning via geometric median for deep convolutional neural networks acceleration. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Long Beach, CA, USA, June 16-20, 2019.

[26] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew G. Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer arithmetic-only inference. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Salt Lake City, UT, USA, June 18-22, 2018.

[27] Liqiang Lu, Yun Liang, Qingcheng Xiao, and Shengen Yan. Evaluating fast algorithms for convolutional neural networks on fpgas. In 25th IEEE Annual International Symposium on Field-Programmable Custom Computing Machines, FCCM, Napa, CA, USA, April 30 - May 2, 2017.

[28] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Salt Lake City, UT, USA, June 18-22, 2018.

[29] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans. Assoc. Comput. Linguistics, 2:67–78, 2014.

[30] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll´ar, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In Computer Vision - ECCV 13th European Conference, Zurich, Switzerland, September 6-12, 2014.

[31] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 5763–5772, 2019.