GSAIC: GeoScience Articles Illustration and Caption Dataset

Rui Shi

School of Information Science and Technology, Fudan University, Shanghai, China

19307130025@fudan.edu.cn

Abstract. The scientific investigation of geoscience includes data collection, sample classification and semantic, consisting of a large number of images. An image-text search model that can well assist the research work of geoscience. However, the existing image-text datasets are mainly in the field of daily life and lack academic image-text datasets. In order to help geoscience researchers to investigate through the image and text, and to provide a new benchmark for researchers in the fields of data mining and information retrieval, this paper proposes a novel parallel material of geoscience academic illiustrateion and caption (GSAIC) based on GAKG, which contains over 900,000 illustrations of earth science papers and the corresponding captions. GSAIC filters out high-quality illustrations and captions through a classifier, and with the support of experts annotations. The GSAIC will support several tasks Including text search for images, retrieving corresponding images or papers based on academic image descriptions and academic illustration classification tasks, for geoscience scenarios Finally, both the GSAIC benchmark and classifier are publicly accessible.

Keywords: Enter key words or phrases in alphabetical order, separated by commas.

1. Introduction

In real-life scenarios, we need to search for pictures by text and search text by pictures. Furthermore, these two tasks rely on the machine's understanding of text and pictures. In understanding the text, the natural language large model like BERT[1] and Transformer[2] has derived much work to embed text into space for representation so that the coordinates in the semantic space represent the inherent semantics of natural language information. In the understanding of images, from the earliest CNN [3] to the later large-scale pre-trained image feature extraction models VGG [4] and ResNet [5], characterizing the picture. Therefore, matching the semantics of the text with the features of the images is the core idea of retrieving between these two modes. This image-text matching problem is a classic multimodal information retrieval and feature engineering problem. As a result, many multimodal datasets have emerged [6-15]. These datasets are roughly divided into large-scale cross-domain datasets, small-scale single-domain datasets, and datasets with special processing for specific domains. And GAKG [49] has the potential to contribute the multimodal community, setting a multimodal knowledge graph for academic scenarios.

A large-scale cross-domain dataset is a dataset that includes various small objects in life. Such datasets are large in number, easy to obtain, and more suitable as benchmarks. The following are widespread examples:

Recipe1M+ [7] contains 1 million recipes and 1,300,000 food images, and labels dishes from the perspectives of ingredients and processes. It is widely used as a benchmark, including image retrieval tasks [21, 22, 23] (especially food image retrieval tasks) and text generation tasks [24].

MSCOCO [6] contains 91 object categories, 328K images, and 2.5M instances with labels. It pre-positions and segments objects in images, and is used in a variety of multimodal tasks. Among them, some works use it to train generative models [16] over it, while others use it as a benchmark for horizontal baselines comparison [17, 18]. In retrieval tasks [19, 20], most of the comparisons are based on precision, recall.

The AIC-ICC [8] dataset consists of an image dataset ICC with Chinese descriptions (including 300K images) and corresponding Chinese explanations. At present, it seems that most of the models using this dataset are generative task scenarios in the Chinese [25, 26] and serve as benchmarks. Not only that, it also drives more research on Chinese multimodal datasets [27].

MIT_STATE[9] contains 63440 pictures and 245 types of objects, each type of object has an average of 9 adjectives to describe, these adjectives emphasize the state of the object and the state transformation between pictures of similar objects, such as "old" and "new". According to the citations, most of the ones exploiting this dataset are image retrieval models [28, 29, 30], which are used as benchmarks to compare with other models on R@k.

Conceptual Captions[10] contains 3.3M pairs of pictures and descriptions, in which the description is generated from the Alt-text of the original image, so the style is changeable. Conceptual Captions are mainly used as a result benchmark for text generation models [31, 32, 33].

A small-scale single-domain dataset can be obtained by simply extracting and crawling a certain dataset for a specific life scene. This kind of dataset is more suitable for large-scale training before the project is implemented for a specific small-scale field application. Three common examples are listed:

Oxford 102 Flowers [12] contains 8189 images of 102 kinds of flowers. The poses and lighting of the flowers in the images vary. On this dataset, many multimodal tasks including image generation tasks [16, 38], text generation tasks [35, 39] and retrieval tasks [36, 40] are benchmarked and compared with each other.

CUB-200-2011 [11] contains 11,788 photos of 200 bird species, all images are annotated with bounding boxes, bird body part locations and attribute labels. This dataset has been widely used in a variety of multimodal scenarios, including image generation [16, 17, 18], text generation [34, 35] and retrieval tasks [36, 37].

Stanford Online Products [13] contains 120k pictures of online products in 23k categories. At present, it appears that the models using this dataset mostly retrieve scenes [37, 41, 42, 43] as result benchmarks.

Datasets that are specially processed for specific fields are the most difficult and professional datasets to acquire. The process and goal of acquisition are to solve specific problems in typical scenarios, which are rare at present.

The WIT [14] dataset is derived from Wikipedia and contains 11.5M images involving 37.6M entities and 108 languages. This dataset is mainly used for benchmarks [44, 45] for retrieval tasks.

CoDraw [15] contains 138k pieces of information, forming 10k dialogues and cartoon images corresponding to each sentence. This dataset is mostly used in picture generation scenarios, especially in the task of “generative visual painting” (GeNeVA) [46, 47, 48].

It is easy to find that the most suitable scenarios for this type of dataset are in life, but in academic life, there is also a great demand for searching images by text and text by images. In the process of reading the paper, we can understand the picture through the accompanying text and better understand the things stated in the paper through the picture. Although we can search for pictures by text, we can search by retrieving the description text of the pictures, but when we want to understand the pictures in detail, the obscurity of these words will cause certain obstacles. Graph, the function of returning related graphs to solve the system of finding related work in scientific research. Of course, when we get a picture of an article, we want to understand the picture's information. If we have a system that can express the textual information of this picture, then our scientific research efficiency will be significantly improved.

In response, we constructed the GSAIC dataset for the geoscience academic community, a rigorously screened and categorized dataset of matching images and captions in geoscience academic papers. We further enhance the influence of GAKG in the earth sciences by searching GAKG and constructing GSAIC and also provide a service for the earth science community to search for text by image and image by text. And our main contributions can be listed as follows:

The GSAIC dataset is proposed. To our knowledge, this is the only dataset currently used for academic multimodal information mining, and we will update this dataset regularly.

The current mainstream multi-modal datasets are listed and counted to facilitate the selection of datasets by various multi-modal researchers.

We provide a three-label classifier for the classification of the illustration after doing the OCR over research papers' PDF.

The rest of the paper is organized as below. In section 2 we introduce the benchmarks exist currently and point out what gap the GSAIC has filled up. In section 3, we will introduce the construction of GSAIC and conclude the whole paper in the final section.

2. Dataset Overview

For the current mainstream multimodal datasets, we have made an overall summary:

Table 1. Introduction of text-image benchmarks.

Dataset name	*Scale(image numbers)*	*Object classes*	*Object numbers*
ms coco	328,000	-	2,500,000
CUB-200-2011	11,788	200	200
Oxford 102 Flowers	8,189	103	103
Multi-Modal-CelebA-HQ	30,000	-	-
CoDraw	~10,000	-	-
Recipe1M+	13,000,000	-	-
Flickr30k	31,000	-	-
AIC-ICC	300,000	-	-
Fashion200K	200,000	5	200,000
MIT_STATE	63,440	245	-
MIRFLICKR	25,000	24	-
NUS_WIDE	269,648	81	-
Visual Genome	100,000	-	2,100,000
ImageNet	3,200,000	5247	-
WIT	11,500,000	-	-
Conceptual Captions	3,300,000	-	
Stanford Online Products	120,053	22,634	120,053
CARS196	16,185	196	16,185
LAION-400M	400,000,000	-	-
GSAIC	~900,000	3	-

For these datasets, we quantified from dataset size, image complexity, and text complexity. We abstracted each complexity into a numerical value of -5~+5, and invited several companions to make a qualitative evaluation together. The evaluation criteria are as follows:

Dataset size: The larger the dataset size, the higher the evaluation value. For example, the scale of the LAION-400M dataset has reached 100 million levels, and the scale evaluation is the highest (point 5); Oxford flowers only have more than 8,000 images, and the scale evaluation is the lowest (point -5). In general, the larger the dataset size, the larger the divided training set, the more multimodal features contained in the training set, and the less likely the trained model is overfit.

Image complexity: The more entity classes and entity features contained in the image, the higher the evaluation of image complexity. Images of multi-entity categories can allow the model to learn the main features of entities more accurately because the multi-entity image analysis process implicitly distinguishes between entities; multi-feature images allow the model to learn more detailed features of entities. In the evaluation process, the number of entities takes precedence over the number of entity features because the model trained on a single entity dataset is less generalized. For example, each image in MSCOCO specifically contains multiple entities and is pre-segmented, which can train a model with good generalization performance; when using CUB simultaneously, it is often necessary to fine-tune the model.

Text complexity: The more the text matches the image, the higher the text complexity evaluation. The text that matches the image can accurately describe the image, more comprehensively include the entities and features in the image so that the model can better correspond to the feature information of the image and text, and better realize multimodal learning. The primary way we evaluate match is to look at the text and then try to find the corresponding image in a set of images. For example, our MSCOCO contains five paragraphs of description text to describe a picture, and the matching is strong, while a flower name in Oxford flowers corresponds to the pictures of all such flowers, and the matching is weak.

After quantification, we get the following two distribution maps—qualitative comparison results of the complexity of 20 graphic multimodal datasets, including GSAIC. The abscissas of the two figures are the scale of the dataset, and the ordinates are the image complexity and text complexity, respectively.

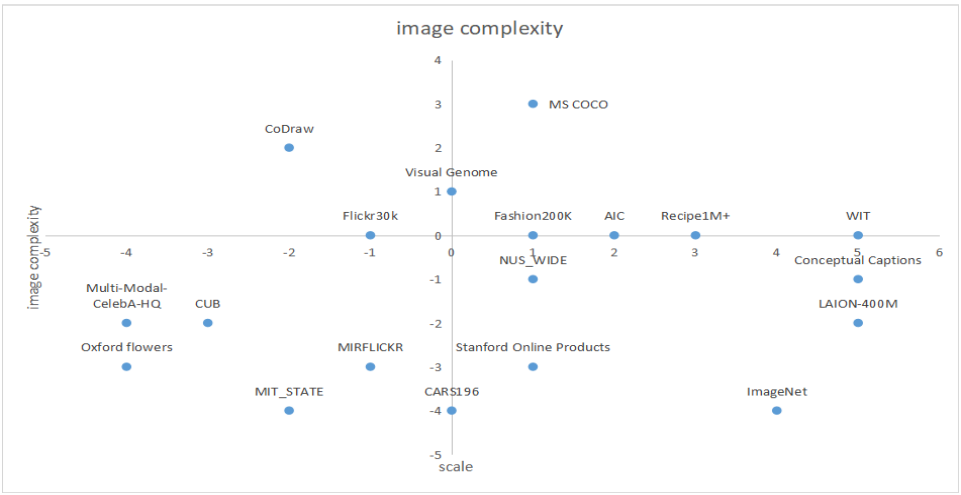


FIG.1 image complexity

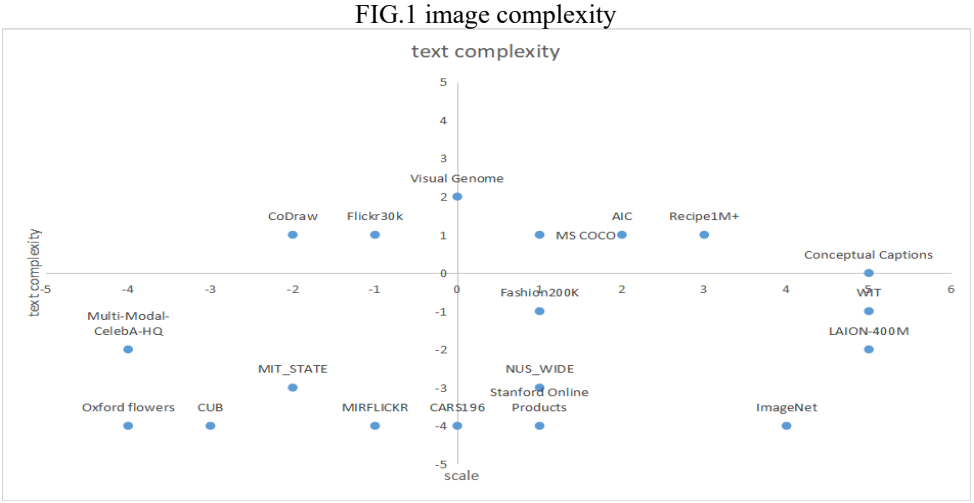
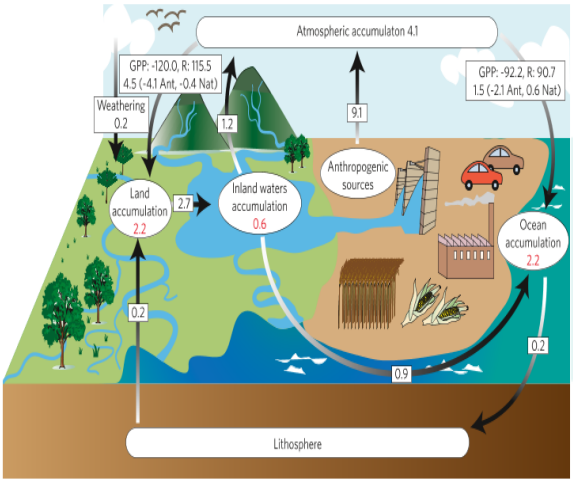
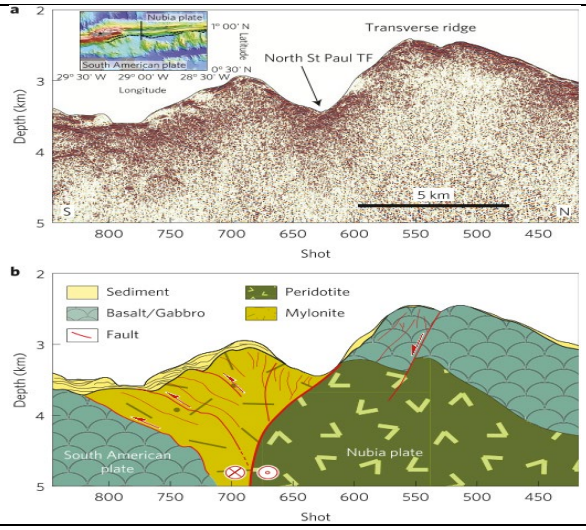
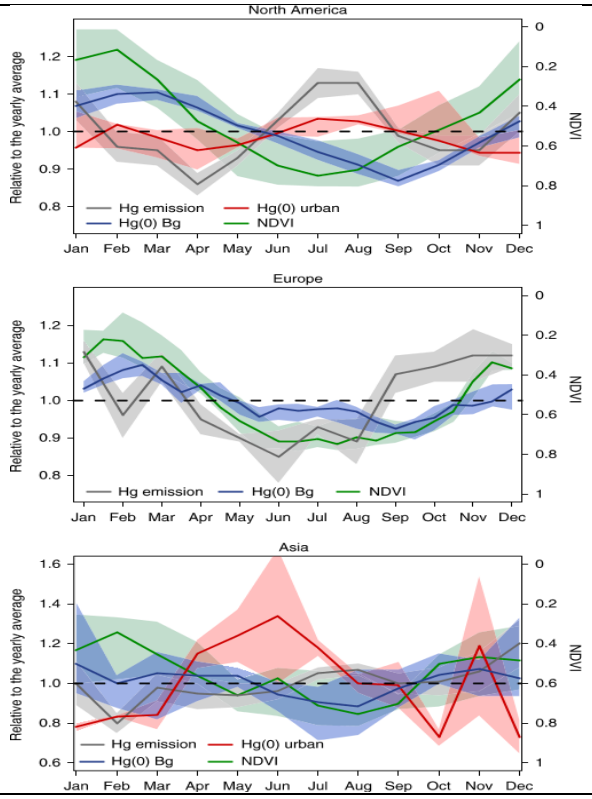
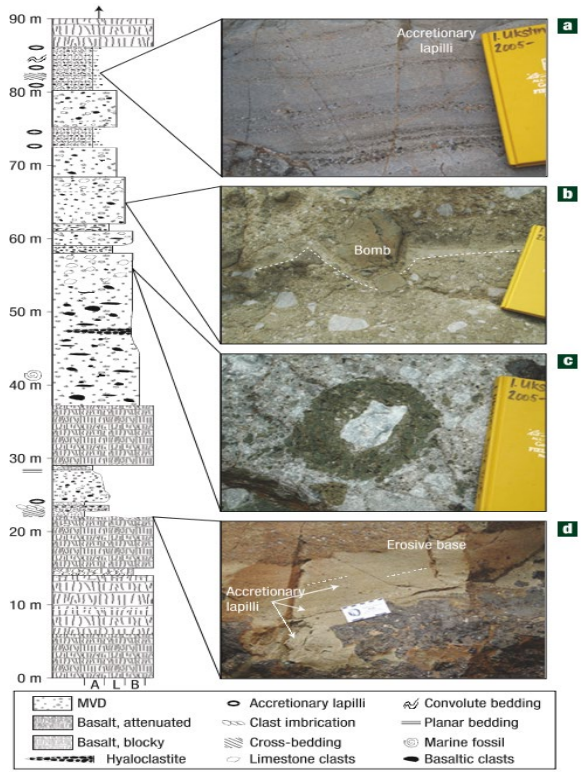


FIG.2 text complexity

As can be seen from the above coordinate diagram, the large-scale and complex multimodal datasets of graphics and text in the first quadrant are scarce. This kind of phenomenon mainly results from high-complexity datasets often needing to label text and filter images manually, so the scale of the model is limited. At the same time, due to the large scale of large-scale datasets, it is not realistic to manually label text and filter images, which limits the complexity of the dataset. Spend.

It can be seen from the above analysis that high-complexity and large-scale datasets are beneficial to the development of multimodal models, but the current real-life datasets have the disadvantage of being extensive but not precise or precise but not big, so based on geosciences. The emergence of multimodal image and text datasets in scientific research papers has filled this gap very well. The illustrations in earth science research papers are of high pixel quality, rich in scientific information, and come with explanatory text that closely describes their content without manual annotation.

Table 2 Some examples in GSAIC, including their images and descriptions, are not exhaustive due to space limitations.

	
The 'boundless carbon cycle'. The schematic highlights carbon fluxes through inland waters5, and also includes pre-industrial2 and anthropogenic3 fluxes. Values are net fluxes ...	Reflection seismic line crossing the St Paul shear zone. a, Fully migrated profile (location shown in the inset; black star, St Peter and St Paul islets). b, ...
	
Seasonal variation of Hg emissions, vegetation activity and atmospheric Hg(0) concentrations. Grey, seasonal variation of Hg...	Stratigraphy of mafic volcaniclastic deposits, Daqiao. Rock lithologies are: lava and mafic volcaniclastic deposits (MVDs), with clast size distribution of...

At the same time, the scale of data can also be massive in the way of crawling, so we believe that the geography graphic dataset has great application potential.

3. Construction of GSAIC

In this section, we will detail the construction of GSAIC, and all the data will be share on the final draft:

3.1 Data preprocessing:

We crawled a lot of illustrations and descriptions on the release pages of geoscience papers from the paper website, and also referred to the GAKG dataset, and then divided out clear pictures through crowdsourcing methods such as human annotation, and obtained 15,260 pairs of illustration and captions. Since these data are extracted by OCR, there are still invalid data such as plain text after removing the low-pixel data, so we decided to design a classification model to assist us in classification.

The data we obtained above can be divided into three parts: GeoScience domain illustrations, charts, and plain text. Plain text pictures need to be eliminated because their literal content is basically the same as the description, and lack of multimodal information; GeoScience domain illustrations can match their descriptions well, and they also have multimodal research value in other fields; as for charts, its statement in captions and diagrams also needs the relevant mathematical knowledge base. Concequently, it is of significance to divide them into a separate category. According to the above analysis, we will classify these 15,260 image-text pairs into the above three categories, so as to improve the validity of the dataset. In this way, it can help researchers studying multimodal model to train their models for. Based on the experimental experience, we decided to use VGG as the backbone of the classification model.

3.2 Classification:

In image classification, we choose to use an image classification mechanism based on VGG. The VGG model is a commonly used model for image classification. This article uses VGG16 because it has a simpler architecture and better performance than other versions of VGG. It has 13 convolutional layers, 5 pooling layers, and 3 fully connected layers. The convolutional layer and pooling layer are partly responsible for feature extraction, and the fully connected layer is responsible for classification. The dimension of feature extraction is up to 4096 dimensions. We use VGG trained on ImageNet for the pretrain model, and use two 2-class models for classification, the first model classifies text/invalid images from useful illustrations like GeoScience domain illustrations and charts, and the second model make GeoScience domain illustrations and charts as two parts. Experiments show that this classification method is more accurate than three-classification model.

Information about the classification model

Input and Output of VGG:

The VGG input is the image of GSAIC, which is converted to 224*224 size using torch.resize, for the purpose of matching ImageNet-based pre-training model.

VGG outputs two 2-dimensional vectors [o1, o2]/[o3, o4], which represent their results on two 2-classification tasks, respectively, and merges them into a 3-dimensional vector [x1, x2, x3], each numerical value represents the likelihood that the graph belongs to three categories.

The specific combination method is as follows:

$$[O_1, O_2] = soft\max([o_1, o_2])$$

$$[O_3, O_4] = soft\max([o_3, o_4])$$

$$[x_1, x_2, x_3] = [O_1, O_2 * O_3, O_2 * O_4]$$

Where o1, o2 are the outputs of the first binary classification model, and o3, o4 are the outputs of the second binary classification model.

After merging in this way, the relative size of each output is preserved. The second value of the first model is equivalent to the weight of the data classified as valid. If a picture is invalid data, its o1 is more significant, and o2 is smaller. x2 and x3 will also be smaller, thus realizing the fusion of the two-category results into the three-category results.

Loss function

In general, the model, first calculate the loss of each image corresponding to the two outputs in each batch:

$$loss(o, class) = -\log\left(\frac{\exp(o[class])}{\sum_j \exp(o[j])}\right) = -o[class] + \log(\sum_j \exp(o[j])) \tag{1}$$

In the above equation, o is the value output by VGG, and class is the category number. First, use the internal softmax method of logarithmic symbols to convert the two values into a probability form with a sum of 1, without changing the relative size, and then take the negative logarithm, so that the larger the numerator, the smaller the loss.

$loss = \hat{y}_i * loss(o_i, class_i)$[2]

Where $\hat{y}$ is the one-hot code of the current image category. Among the losses of the three categories in expression [1], we only select the loss of the corresponding category

$$L = \frac{1}{N}\sum_{N}^{k=1} loss_k$$

Finally, the N losses of N pictures in a batch are averaged to get the total loss of this batch.

By building a three-label classifier, we can well judge which illustrations from the PDF should be keep and which should not be preserved so that the illustrations quality can be guaranteed. Finally, it is shown that the accuracy rate of our classification model is 86.03%, and the accuracy rate of separating invalid data is 98.30%. The main errors are concentrated in the classification between GeoScience domain illustration and charts because small charts or chart elements often appear on geoscience images. Under this circumstance, we manually reclassify or mix this two-class illustrations in practical work. At the same time, the accuracy of direct three-classification is 84.34%, it is evident that the accuracy of classification invalid data also decreases simultaneously. So, in the end, we use two binary classification models.

4. Conclusion

Based on GAKG, this paper proposes a novel GeoScience illustrations and captions parallel resource benchmark named GSAIC, which contains 15,620 geoscience paper illustration data and corresponding captions. To our knowledge, this is the first parallel material, used for academic multimodal data mining, and we will update this dataset regularly. The current mainstream multimodal datasets are introduced and counted to facilitate the selection of datasets by various multimodal researchers. We believe that the GSAIC dataset will support the community of GeoScience and the researchers in CV, since the dataset with complex captions and professional illustrations are rare, which sets a strong stumbling block to the existing methods.

References

[1] Devlin, Jacob et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” ArXiv abs/1810.04805 (2019): n. pag.

[2] Vaswani, Ashish et al. “Attention is All you Need.” ArXiv abs/1706.03762 (2017): n. pag.

[3] LeCun, Yann et al. “Gradient-based learning applied to document recognition.” Proc. IEEE 86 (1998): 2278-2324.

[4] Simonyan, Karen and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” CoRR abs/1409.1556 (2015): n. pag.

[5] He, Kaiming et al. “Deep Residual Learning for Image Recognition.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 770-778.

[6] Lin, Tsung-Yi et al. “Microsoft COCO: Common Objects in Context.” ECCV (2014).

[7] Salvador, Amaia et al. “Learning Cross-Modal Embeddings for Cooking Recipes and Food Images.” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017): 3068-3076.

[8] Wu, Jiahong et al. “AI Challenger : A Large-scale Dataset for Going Deeper in Image Understanding.” ArXiv abs/1711.06475 (2017): n. pag.

[9] Isola, Phillip et al. “Discovering states and transformations in image collections.” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): 1383-1391.

[10] Sharma, Piyush et al. “Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning.” ACL (2018).

[11] Wah, Catherine et al. “The Caltech-UCSD Birds-200-2011 Dataset.” (2011).

[12] Nilsback, Maria-Elena and Andrew Zisserman. “Automated Flower Classification over a Large Number of Classes.” 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing (2008): 722-729.

[13] Song, Hyun Oh et al. “Deep Metric Learning via Lifted Structured Feature Embedding.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 4004-4012.

[14] Srinivasan, Krishna et al. “WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning.” Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2021): n. pag.

[15] Kim, Jin-Hwa et al. “CoDraw: Collaborative Drawing as a Testbed for Grounded Goal-driven Communication.” ACL (2019).

[16] Zhang, Han et al. “StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks.” 2017 IEEE International Conference on Computer Vision (ICCV) (2017): 5908-5916.

[17] Ramesh, Aditya et al. “Zero-Shot Text-to-Image Generation.” ArXiv abs/2102.12092 (2021): n. pag.

[18] Xu, Tao et al. “AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks.” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018): 1316-1324.

[19] Yuan, Li et al. “Central Similarity Quantization for Efficient Image and Video Retrieval.” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020): 3080-3089.

[20] Messina, Nicola et al. “Transformer Reasoning Network for Image- Text Matching and Retrieval.” 2020 25th International Conference on Pattern Recognition (ICPR) (2021): 5222-5229.

[21] Xie, Zhongwei et al. “Cross-Modal Retrieval between Event-Dense Text and Image.” Proceedings of the 2022 International Conference on Multimedia Retrieval (2022): n. pag.

[22] Fain, Mikhail et al. “Dividing and Conquering Cross-Modal Recipe Retrieval: from Nearest Neighbours Baselines to SoTA.” ArXiv abs/1911.12763 (2019): n. pag.

[23] Fontanellaz, Matthias et al. “Self-Attention and Ingredient-Attention Based Model for Recipe Retrieval from Image Queries.” Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management (2019): n. pag.

[24] Wang, Hao et al. “Structure-Aware Generation Network for Recipe Generation from Images.” ECCV (2020).

[25] Lu, Huimin et al. “Chinese Image Captioning via Fuzzy Attention-based DenseNet-BiLSTM.” ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 17 (2021): 1 - 18.

[26] Song, Yuqing et al. “Unpaired Cross-lingual Image Caption Generation with Self-Supervised Rewards.” Proceedings of the 27th ACM International Conference on Multimedia (2019): n. pag.

[27] Li, Xirong et al. “COCO-CN for Cross-Lingual Image Tagging, Captioning, and Retrieval.” IEEE Transactions on Multimedia 21 (2019): 2347-2360.

[28] Anwaar, Muhammad Umer et al. “Compositional Learning of Image-Text Query for Image Retrieval.” 2021 IEEE Winter Conference on Applications of Computer Vision (WACV) (2021): 1139-1148.

[29] Zhang, Feifei et al. “Joint Attribute Manipulation and Modality Alignment Learning for Composing Text and Image to Image Retrieval.” Proceedings of the 28th ACM International Conference on Multimedia (2020): n. pag.

[30] Vo, Nam S. et al. “Composing Text and Image for Image Retrieval - an Empirical Odyssey.” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019): 6432-6441.

[31] Mokady, Ron et al. “ClipCap: CLIP Prefix for Image Captioning.” ArXiv abs/2111.09734 (2021): n. pag.

[32] Li, Zhuowan et al. “Context-Aware Group Captioning via Self-Attention and Contrastive Features.” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020): 3437-3447.

[33] Guo, Dan et al. “Recurrent Relational Memory Network for Unsupervised Image Captioning.” IJCAI (2020).

[34] Vedantam, Ramakrishna et al. “Context-Aware Captions from Context-Agnostic Supervision.” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017): 1070-1079.

[35] Chen, Tseng-Hung et al. “Show, Adapt and Tell: Adversarial Training of Cross-Domain Image Captioner.” 2017 IEEE International Conference on Computer Vision (ICCV) (2017): 521-530.

[36] Wei, Xiu-Shen et al. “Selective Convolutional Descriptor Aggregation for Fine-Grained Image Retrieval.” IEEE Transactions on Image Processing 26 (2017): 2868-2881.

[37] Zhe, Xuefei et al. “Directional Statistics-based Deep Metric Learning for Image Classification and Retrieval.” ArXiv abs/1802.09662 (2019): n. pag.

[38] Yuan, Mingkuan and Yuxin Peng. “CKD: Cross-Task Knowledge Distillation for Text-to-Image Synthesis.” IEEE Transactions on Multimedia 22 (2020): 1955-1968.

[39] Zhao, Wei et al. “Dual Learning for Cross-domain Image Captioning.” Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (2017): n. pag.

[40] Lv, Xiaoming and Fa-jie Duan. “Metric learning via feature weighting for scalable image retrieval.” Pattern Recognit. Lett. 109 (2018): 97-102.

[41] Brown, A. et al. “Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval.” ArXiv abs/2007.12163 (2020): n. pag.

[42] Chen, Binghui and Weihong Deng. “Hybrid-Attention Based Decoupled Metric Learning for Zero-Shot Image Retrieval.” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019): 2745-2754.

[43] Chen, Binghui and Weihong Deng. “Energy Confused Adversarial Metric Learning for Zero-Shot Image Retrieval and Clustering.” ArXiv abs/1901.07169 (2019): n. pag.

[44] Jain, Aashi et al. “MURAL: Multimodal, Multitask Retrieval Across Languages.” ArXiv abs/2109.05125 (2021): n. pag.

[45] Messina, Nicola et al. “Transformer-Based Multi-modal Proposal and Re-Rank for Wikipedia Image-Caption Matching.” ArXiv abs/2206.10436 (2022): n. pag.

[46] El-Nouby, Alaaeldin et al. “Tell, Draw, and Repeat: Generating and Modifying Images Based on Continual Linguistic Instruction.” 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019): 10303-10311.

[47] Liu, Zhenhuan et al. “IR-GAN: Image Manipulation with Linguistic Instruction by Increment Reasoning.” Proceedings of the 28th ACM International Conference on Multimedia (2020): n. pag.

[48] Matsumori, Shoya et al. “LatteGAN: Visually Guided Language Attention for Multi-Turn Text-Conditioned Image Manipulation.” IEEE Access 9 (2021): 160521-160532.

[49] Deng, Cheng et al. “GAKG: A Multimodal Geoscience Academic Knowledge Graph.” Proceedings of the 30th ACM International Conference on Information & Knowledge Management (2021): n. pag.