The Solution for the CVPR2023 NICE Image Captioning Challenge

Xiangyu Wu, YiGao, Hailiang Zhang, YangYang, Weili Guo, Jianfeng Lu Nanjing University of Science and Technology

Abstract

In this paper, we present our solution to the New frontiers for Zero-shot Image Captioning Challenge. Different from the traditional image captioning datasets, this challenge includes a larger new variety of visual concepts from many domains (such as COVID-19) as well as various image types (photographs, illustrations, graphics). For the data level, we collect external training data from Laion-5B, a large-scale CLIP-filtered image-text dataset. For the model level, we use OFA, a large-scale visual-language pre-training model based on handcrafted templates, to perform the image captioning task. In addition, we introduce contrastive learning to align image-text pairs to learn new visual concepts in the pre-training stage. Then, we propose a similarity-bucket strategy and incorporate this strategy into the template to force the model to generate higher quality and more matching captions. Finally, by retrieval-augmented strategy, we construct a contentrich template, containing the most relevant top-k captions from other image-text pairs, to guide the model in generating semantic-rich captions. Our method ranks first on the leaderboard, achieving 105.17 and 325.72 Cider-Score in the validation and test phase, respectively.

1. Introduction

Zero-shot image captioning [3, 4] requires joint modeling for vision and language, which aims to generate a concise textual summary for a given image. However, in real-world scenarios, high-quality human-annotated data are always difficult to obtain. Therefore, it is crucial and feasible that learn aligned vision and language representations from large-scale web-crewed data and transfer knowledge from pre-training models to downstream tasks.

NoCaps [2] is an extensive attended dataset for zero-shot image captioning, which contains nearly 400 object classes seen in test images. However, different from NoCaps, the competition dataset includes a larger variety of novel visual concepts as well as various image types (an example is shown in Figure 1).

To accomplish this task, the models need to broadly un-

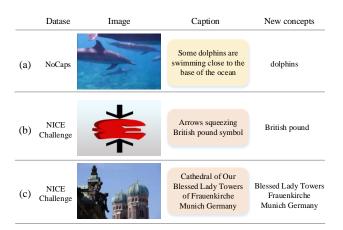


Figure 1. (a): NoCaps dataset, which always includes common objects such as animals, plants and furniture, etc. (b)(c): NICE Challenge dataset, which includes many novel visual concepts and various image types, such as famous historic, cultural and graphics, etc.

derstand vision-language relations and simultaneously learn how to combine language components for a new concept of image. Therefore, we use OFA [13], a large-scale Task-Agnostic and Modality-Agnostic pre-training framework, as our base model, and Laion-5B [12], a dataset of 585 billion CLIP-filtered image-text pairs, as the main source of our base dataset.

In addition, we explore some strategies to enhance zeroshot capacity: (1) we introduce contrastive learning on the OFA model to align a wide range of visual conceptual representations. (2) we propose a similarity-bucket strategy to incorporate image-text similarity into the template and guide model to generate high-quality and more matching captions. (3) we employ a retrieval-augmented [15] strategy to enrich the textual information of the template. As a result, our method ranks first on the leaderboard, producing 105.17 and 325.72 Cider-Score in the validation and test phase, respectively. In the remainder of this technical report, we will introduce the detailed architecture of our solution for this challenge.

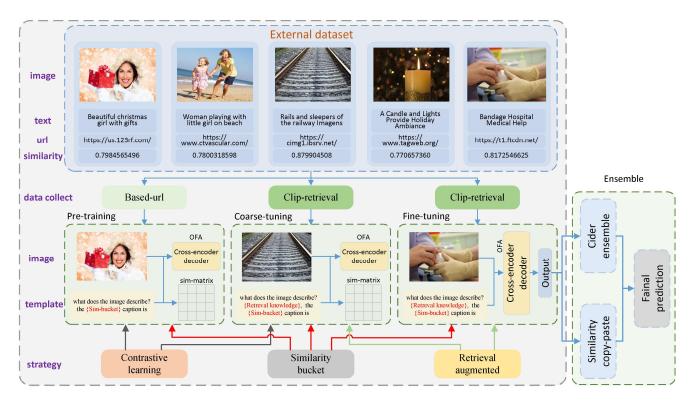


Figure 2. Overall Architecture. Our solution consists of four main stages, which includes Pre-training, Coarse-tuning, Fine-tuning and Model-ensemble. The training data for the first three stages are all collected from the large-scale Laion-5B dataset.

2. Related Work

2.1. Vision-language Pre-training

Vision-language pre-training (VLP) [8, 9] aims to improve the performance of downstream tasks by pre-training the model on large-scale image-text pairs. The single-stream models refer to models where the text and visual features are concatenated together, then fed into a single transformer block. The dual-stream models refer to models where the text and visual features are not concatenated together but sent to two different transformer blocks independently. Different from the above two architectures, the OFA [13] model formulates both pretraining and finetuning tasks in a unified sequence-to-sequence abstraction via handcrafted instructions to achieve Task-Agnostic.

2.2. Image Captioning

Zero-shot image captioning aims to generate textual descriptions without human-annotated data. This involves developing algorithms and models that can analyze the visual content and generate a corresponding textual description. Image captioning has a wide range of applications, including image and video search, assistive technologies for the visually impaired, and automated content generation for social media and marketing. Feng et al. [5] propose unsupervised captioning without using paired image-caption su-

pervision. Kim et al [7] focus on learning efficiency and improve the data efficiency by learning from auxiliary unpaired image-caption data.

2.3. Vision-Language Retrieval

Vision-Language Retrieval [6, 14] aims to learn consistent representations of different modalities, to retrieve instances of one modality based on queries from another modality. The goal of vision-language retrieval is to create more intuitive and effective ways for humans to interact with machines, such as through image and video search, automated image captioning, and visual question answering. For example, [4] proposed a similarity graph reasoning module relying on a graph convolution neural network. With the developments in Transformer-based language understanding, large-scale vision-language transformers have inspired deeper modal interaction in retrieval models.

3. Methodology

3.1. Overall Architecture

Figure 2 illustrates the overall architecture of our solution, which contains four components: Pre-training, Coarse-tuning, Fine-tuning, and Model-ensemble.

Pre-training stage collects specific *url* data from Laion-5B, which can align a wide range of image-text concepts

and store sufficient vision-language knowledge through contrastive learning, image captioning pre-train objectives, and handcrafted templates with similarity-bucket strategy.

Coarse-tuning stage utilizes the *Clip-retrieval* [1] library to retrieve small-scale datasets with simple data cleaning, which can learn a large variety of vision-language novel concepts similar to the competition domain. In this stage, we introduce a retrieval-augmented strategy to retrieve abundant textual knowledge and integrate it into the template. As same as the pre-training stage, we utilize contrastive learning, image captioning, and similarity-bucket to coarse tuning.

Fine-tuning stage further compresses the retrieved dataset in the coarse-tuning stage and adds it to the competition validation dataset. By applying similarity-bucket and retrieval-augmented strategies, the zero-shot performance of the model can be effectively improved.

Model-ensemble is the last stage, which uses similarity-copy-paste and cider-ensemble tricks to improve the generalization ability of the model on the zero-shot image captioning task.

Next, we will specifically introduce contrastivelearning, similarity-bucket, retrieval-augmented strategies, and model-ensemble tricks, as well as how to integrate these strategies into the vision-language pre-training model OFA.

3.2. Contrastive-learning

Contrastive-learning aims to learn better uni-modal representations before fusion. A similarity function is learned in which the parallel image-text pairs are assigned higher similarity scores. We take the [CLS] embeddings output by image and text encoders as joint representations for prediction. Then the cross-modal contrastive learning can be formulated as:

$$\ell_{itc} = \frac{1}{2} \mathbb{E}_{(\mathbf{i}, \mathbf{t})} [CE(\mathbf{y}^{i2t}(\mathbf{i}), \mathbf{p}^{i2t}(\mathbf{i})) + CE(\mathbf{y}^{t2i}(\mathbf{t}), \mathbf{p}^{t2i}(\mathbf{t}))]$$

$$p_b^{i2t}(\mathbf{i}) = \frac{exp(s(\mathbf{i}, \mathbf{t}_b)/\tau)}{\sum_{b=1}^{B} exp(s(\mathbf{i}, \mathbf{t}_b)/\tau)}, p_b^{t2i}(\mathbf{t}) = \frac{exp(s(\mathbf{t}, \mathbf{i}_b)/\tau)}{\sum_{b=1}^{B} exp(s(\mathbf{t}, \mathbf{i}_b)/\tau)}$$
(1)

where $p_b^{i2t}(\mathbf{i})$ and $p_b^{t2i}(\mathbf{t})$ denote softmax-normalized image-to-text and text-to-image similarity with batch size B and temperature scale parameter τ . CE denotes crossentropy loss. We follow ALBEF [10] and use the momentum model to generate pseudo-targets as additional supervision.

3.3. Similarity-bucket

Similarity-bucket strategy provides different similarity-prompts to the vision-language model in the pre-training, coarse-tuning, and fine-tuning stages. During the training part, we define *n* buckets according to the similarity of training datasets in each stage. The similarity of the image-text

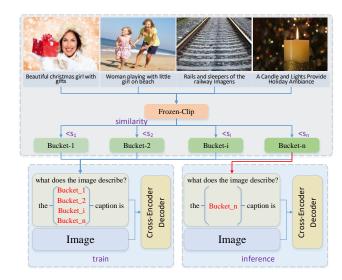


Figure 3. Similarity-bucket is utilized in pre-training, coarsetuning, and fine-tuning stages.

pair in the first bucket is very low, and may be noisy data. The n-th bucket indicates that the similarity and quality of the image-text pair are very high. Through the similarity prompt, the model can learn the representation of training data with different qualities independently. In the inference part, the similarity-prompt is fixed to the n-th bucket, forcing the model to generate higher quality and more matching captions.

As shown in figure 3, for example, in the pre-training stage, for each image-text pair, we utilize Frozen-Clip [11] to predict the similarity of the image-text pair, and then we can obtain a similarity collection, which includes the similarity of all training data. With n denoting the number of buckets, then based on the size of the dataset, we divide different similarity thresholds to represent each bucket. The result of splitting buckets satisfies the condition that the number of image-text pairs for the first and $n\!-\!th$ buckets is less than the number of image-text pairs for the remaining buckets.

In the training part, each image-text pair belongs to only one bucket, therefore, each image-text pair corresponds to a certain similarity-prompt. We insert this similarity-prompt into the template of the OFA model, resulting in the template: "What does the image describe? The {bucket_i} caption is". Through this similarity-prompt, the model can learn the representation of data with different qualities. In the inference part, we fix the similarity-prompt to the *n-th* bucket and control the model to generate the most matched and high-quality caption for a given image.

3.4. Retrieval-augmented

Retrieval-augmented [15] strategy provides a mini knowledge-base for each image-text pair during the train-

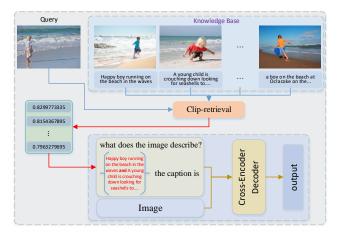


Figure 4. Similarity-bucket is utilized in pre-training, coarsetuning, and fine-tuning stages.

ing part. The model can not only extract visual features such as objects, attributes, and relationships of the image but also explicitly align the information of the image with the knowledge in the knowledge-base.

As shown in figure 4, for example, in the coarse-tuning stage, for each image query, we utilize Clip-retrieval [1] library to retrieve top-k image-text pairs according to the similarity between image query and each image of the Laion-5B dataset, and then we extract k texts from the top-k retrieved image-text pairs. These k texts are concatenated into a mini knowledge-base and inserted into the template of the OFA model, resulting in the template: "What does the image describe? {retrieval knowledge}, the caption is". Through this retrieved knowledge, the model can generate content-rich and diverse captions.

3.5. Model-ensemble

The last stage is model-ensemble, we use similarity copy-paste and cider-ensemble tricks. Similarity copy-paste trick directly copies and paste the caption through a fast shortcut. Specifically, in the training dataset, we compute the similarity between test data and training data. Then, we set a larger similarity threshold to select the most relevant image-text pair as the candidate pair. Finally, we define the similarity between the candidate caption and the candidate image as c_1 , and the similarity between the model prediction and the candidate image as c_2 , respectively, The caption corresponding to $max(c_1, c_2)$ is used as the final prediction result. Cider-ensemble trick aims to find the best prediction result. Specifically, we fine-tune *n* models to obtain n prediction results, each prediction is calculated as a cider-score along with other n-1 predictions, and the caption corresponding to the highest score is used as the final prediction result.

4. Experiments

4.1. Implementation Detail

The training data at all stages are collected from Laion-5B, a large-scale CLIP-filtered imagetext dataset. Each image-text pair in Laion-5B includes an image, a text, an url, and the similarity between the image and text. In the pre-training stage, we collect 6M image-text pairs based on specific url (thumbx.shutterstock.com, editorial01.shutterstock.com, etc.), and extract top-1M imagetext pairs based on similarity. In the coarse-tuning stage, we use all the competition images to retrieve external data from Laion-5B through Clip-retrieval library. For each image query, we retrieve top-30 image-text pair based on the similarity between the image query and Laion-5B, and retain 120k image-text pairs through simple data cleaning, such as filtering out too long, too short and non-English image-text pair. In the fine-tuning stage, we also use all the competition images to retrieve external data from Laion-5B through Clip-retrieval library. For each image query, we retrieve top-10 image-text pair with the only url www.tscdn.net, and retain 12k image-text pairs. In addition, we also add the 5kvalidation dataset to the fine-tuning stage.

Model. The base image captioning model we used is OFA, which is a large-scale visual-language pre-training model based on handcrafted templates. In each training stage, the number of similarity buckets is 4, and these four different similarity-prompts are "noise", "low quality", "high quality", "best match". In the coarse-tuning and fine-tuning stages, the numbers of retrieved relevant imagetext pairs are 1,2 and 4 so that we can obtain different models to execute the model-ensemble trick.

Pre-training stage. The size of the pre-training dataset is IM, and we load the pre-trained ofa-large weights. Due to the large scale of the pre-training dataset, we only continued to pre-train 5 epochs. We use $NVIDIA\ RTX\ 3090\times 4$ with an initial learning rate of Ie-5, the input image size is 380×380 , the batch size is I6, and the input text length is I6. All other hyper-parameters follow the default setting of the OFA model.

Coarse-tuning stage. The size of the coarse-tuning dataset is 120k, For each image query, we retrieve top-30 image-text pairs based on similarity between the image query and Laion-5B. We load the pre-trained weight from the pre-training stage and coarse-tune 20 epochs. We use $NVIDIA\ RTX\ 3090\times4$ with an initial learning rate of 1e-5, the input image size is 480×480 , the batch size is 16, and the input text length is 30. All other hyper-parameters follow the default setting of the OFA model.

Fine-tuning stage. The size of the fine-tuning dataset is 17k, which includes 12k image-text pairs retrieved from Laion-5B and 5k competition validation dataset. We load the coarse-tuned weight from the coarse-tuning stage and

fine-tune 100 epochs. We use *NVIDIA RTX* 3090×4 with an initial learning rate of 1e-5, the input image size is 480×480 , the batch size is 16, and the input text length is 30. All other hyper-parameters follow the default setting of the OFA model.

Model-ensenble stage. We perform two random image augmentations for the test set and training set, obtaining 4 similarities, and taking the average as the final similarity. The range of similarity values is between $0.15\sim0.4$, and we set a large similarity threshold to 0.35 to perform the copypaste trick. By setting the number of different top-k retrieval image-text pairs, taking weights for different epochs, and so on, we ultimately fuse 20 models using the cider-score trick to obtain the final score.

4.2. Result

Table 1. We report the Cider score of our methods on the test set.

#	Method	Cider
1	OFA+Pre-training,Coarse-tuning,Fine-tuning	170+
2	Retrieval-augmented	290+
3	Similarity-bucket	310+
4	Model-ensemble	325+

As reported in table 1, with pre-training in 1M dataset and finn-tuning in 5k validation dataset, the performance of the OFA model reached a 170+ cider score. The most significant improvement strategy is the retrieval-augmented, which improves the cider score by 120+ points. By utilizing the similarity-bucket strategy and model-ensemble trick, we achieved the highest cider score on the leaderboard of 325+.

5. Conclusion

This report summarizes our solution for the New frontiers for Zero-shot Image Captioning challenge, which includes four core components: Pre-training, Coarse-tuning, Fine-tuning, and Model-ensemble. Our solution indicates that the similarity-bucket is an effective paradigm for controlling the model to generate high-quality results to vision-language downstream tasks. The final competition results show the effectiveness of our solution.

References

- [1] Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them. https://github.com/rom1504/clip-retrieval, 2022. 3, 4
- [2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale, Feb 2020. 1

- [3] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning, Feb 2021. 1
- [4] Zhengcong Fei, Xu Yan, Shuhui Wang, and Qi Tian. Deecap: Dynamic early exiting for efficient image captioning. 1, 2
- [5] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jan 2020. 2
- [6] Zhong Ji, Haoran Wang, Jungong Han, and Yanwei Pang. Sman: Stacked multimodal attention network for cross-modal image-text retrieval. <u>IEEE Transactions on</u> Cybernetics, page 1086–1097, May 2020. 2
- [7] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. In <u>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Nov 2019. 2</u>
- [8] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. 2
- [9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. 2
- [10] Junnan Li, RamprasaathR. Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and StevenC.H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation, Jul 2021. 3
- [11] Alec Radford, JongMin Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, JackA. Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, Feb 2021. 3
- [12] Christoph Schuhmann, §§°°romain Beaumont, Vencu Vencu, Ade Gordon, Wightman Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, °°jenia Jitsev, UC Berkeley, and Gentec Data. Laion-5b: An open large-scale dataset for training next generation image-text models. 1
- [13] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, Hongxia Yang, and Chang Zhou; ericzhou. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. 1, 2
- [14] Yabing Wang, Jianfeng Dong, Tianxiang Liang, Minsong Zhang, Rui Cai, and Xun Wang. Cross-lingual cross-modal retrieval with noise-robust learning. In <u>Proceedings of the</u> 30th ACM International Conference on Multimedia, Oct 2022. 2
- [15] Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, Ming-Yu Liu, Yuke Zhu, Mohammad Shoeybi, Bryan

Catanzaro, Chaowei Xiao, and Anima Anandkumar. Revilm: Retrieval-augmented visual language model for zero and few-shot image captioning. 1,3