

An Ensemble Approach for Wikipedia Image/Caption Matching Competition

Youwon Jang, Jaewon Kim, Dongkyu Cho, Uicheon Koh
Seoul National University

{sharifa, rlawodnjs017, kulupapa1127, hskoh87}@snu.ac.kr

Abstract

This paper deals with the Wikipedia-Image/Caption Matching task [1] introduced in the Kaggle competition site. Most of the images in Wikipedia, the world's largest online encyclopedia, do not have adequate captions or alt texts to describe the images or show a lack of sufficient understanding on them. The task covered in this paper is meant for a learning process based on millions of data (images and information of them). The task to be finally evaluated is to match the text for each image when tens of thousands of images and texts describing the images are given in random order. By doing so, it becomes possible to increase the understanding of the images by finding appropriate information for them on Wikipedia where their explanations are missing. Furthermore, it can be used for image search. To do the task, this paper proposes a novel method using Text-Image Matching. As for the text-image matching/retrieval method, CLIP [2] is used, while various modifications were made to match our designated task. In conclusion, our method earned a score of 0.33685, ranking 18th out of a total of 105 teams.

1. Introduction

Vision and natural languages are two significant main areas for understanding the real world, and many studies have tried to connect these areas [3], [4], [5]. Among the trials, the Image-Text Matching task, which means matching suitable texts with certain images, is crucial in that it can directly bridge the two domains. The task can be understood as measuring the visual-semantic similarity between natural language sentences and images. It can also be extended to problems such as finding an appropriate text description from an image query and vice versa.

Captions or alt texts can increase understanding of images and show better search results. However, most of the images on Wikipedia do not have or show a significant lack of this sort of additional information. Against this backdrop, the Kaggle competition was proposed.

Problem Configuration.

- **Goal.** Given some texts amounting to 92k images (URLs), participants find and submit up to five texts that are most relevant (and most relevant) for each image.
- **Ranking.** Before the end of the competition period, the result from the 15% (corresponding to 92k pieces) disclosed to participants among all test data is only reflected in the ranking, but the actual evaluation is executed based on the remaining 85% (about 520k) test data which remains private until the end of the competition.
- **Evaluation.** Evaluation is carried out using the NDCG%5 metric, which is mainly used in ranking-based recommendation systems.
- **Competition Period.** The Wikipedia Image/Text Matching competition started in mid-September 2021 and ends in mid-December. This team will participate in the mid-way and will participate for about 7 weeks.

In this paper, we are presenting an ensemble method for the problem proposed by Kaggle. The candidate sentences obtained by each method are selected in the order of similarity, and the top-5 description with the highest similarity per test image is finally selected. Our ensemble method contains:

1. Text/Image matching is performed by Contrastive Language-Image Pre-Training Model (CLIP [2]) according to Wikipedia Image/Caption dataset.
2. Calculate the similarity between descriptions and image URLs.

Contributions in this study is as follows:

- For the Wikipedia Image/Caption Matching Problem proposed in the Kaggle competition, Text/Image Matching was selected. Initially, in the midterm paper, we presented two candidate methods, Text/Image Matching and Image Captioning + Text Matching. However, due to certain conditions, the team decided to implement the first method.
- Considering the english-only context in original CLIP, we tried to replace visual encoder of CLIP with multilingual pretrained model (XLM-R [6]) and finetuned the model.

- We propose an ensemble method that extracts candidates from CLIP and URL-based approach simultaneously. The performance of our approach is about double than using just single method.
- In the competition that ended in mid-December, our ensemble model got a score of 0.33685 and placed 18th out of a total of 108 teams.

2. Related works

A few recent years have witnessed great advances in many computer vision fields such as Cross-modal Retrieval, Image captioning, and Vision & Multilingual modeling.

2.1. Image-Text Matching

Many existing studies have focused on learning the image-text correspondence based upon co-occurrence. Generally, there have been two approaches. Global correspondence learning method ([7], [8]) aims to capture correspondence between the entire image and text. The local correspondence learning method, on the other hand, learns the correspondence through the inspection of each local region-text pair.

Among the local learning approaches, SCAN [5] uses Stacked Cross Attention to learn full latent alignments using image region-word as context and infer image-text correspondence from it. Despite the remarkable improvement brought by such attention-based approaches, the neglect of delicate details is inevitable during the process. GSMN [9] tackles this problem with the implementation of graph structures, minimizing the loss of information by focusing on learning fine-grained phrase correspondence between image-text nodes. The recently publicized model, CLIP [2] is also a much anticipated answer to this question, and we will discuss the model in-depth in the latter part of this paper.

2.2. Data Augmentation

Data augmentation is key to designing a successful metric learning model. Having selected a metric learning-based model as our method, we needed to perform research on data augmentation policies to maximize our model performance. Over the years, researchers have focused on developing effective data augmentation policies that would thrust deep learning into a variety of fields. AutoAugment [10] used reinforcement learning to select a sequence of augmentation processes. AutoAugment was only the first among a series of automated augmentation models, each significantly improving generalization in the varied dataset. Fast AutoAugment [11] showed that an augmentation policy trained for density matching boosts generalization accuracy. In our task, we used RandAugment [12], a model designed to deal with previous issues (increased train complexity, or incapability to adjust regularization strength) trig-

gered by a separate search phase. RandAugment deals with such issues by removing the separate search phase, and replacing it with an optimal, simple search space. This allowed us to perform automated data augmentation at a low cost. This may not be considered as a novelty, but still functions as a trick to increase model performance.

2.3. Image Captioning

Image captioning is an active research area in vision and language. A series of developments have been proposed to boost image captioning by learning joint representations between the vision and language modalities via attention mechanism [13]. In particular, [14] exploits visual attention at object level via a bottom-up mechanism, and all salient image regions are associated with the output words through a top-down mechanism for image captioning. After that, Oscar [15], the state-of-the-art image captioning model, used the object tags detected in images as *anchor points* to ease the learning of alignments between vision and language modalities. Oscar is pre-trained on the massive corpus of text-image pairs for learning of joint representations and fine-tuned on downstream tasks such as image captioning and text-image retrieval, creating the new standard of vision-language understanding and generation methods. In this competition, we aborted the Image Captioning Model in favor of Image-Text Matching. Nonetheless, we thought image-captioning was worth mentioning.

2.4. Multilingual modeling

After the emergence of an epoch-making model Natural Language Processing, such as Transformer [16] and BERT [17], a lot of trials have been conducted not only to boost their performances, but also to extend and transfer the single language models into multi-language formats in various ways. When it comes to TextVQA problems, a new answering model based on a multimodal transformer architecture was proposed [18]. In order to implement multi-modality, a machine translation-augmented framework was devised, using Masked Region-to-Token Modeling and Visual Translation Language Modeling pre-training tasks [19].

In addition, pre-trained multi-language models under the current limitations of single modal representation are followed by a multi-lingual and multi-modal scenario learning universal representation [20].

2.5. Multilingual BERT(M-BERT)

After BERT(Bidirectional Encoder Representations from Transformers) was proposed, its multi-lingual version soon came out to share its variation across languages. It was pre-trained with monolingual corpora in 104 languages and showed relatively decent performances when fine-tuned for evaluation in another language [21]. The appearance of the so-called 'M-BERT'(Multilingual BERT) also led to em-

pirical evaluation of its performances in many different aspects [22, 23].

2.6. Normalized Discounted Cumulative Gain

When submitted, our captioning prediction is evaluated with the nDCG%5 metric. The nDCG metric is widely used to measure the ranking quality. In other words, it computes the accuracy of the predicted ranking.

Computing the nDCG starts by computing the DCG scores of the 'ideal'(given) ranking and the predicted ranking. A DCG(Discounted Cumulative Gain) is the sum of all the relevance scores in a recommendation set, divided by the log of the corresponding position.

$$DCG = \sum_{i=1}^n \frac{\text{relevance}_i}{\log_2(i+1)}$$

Where n is the number of options with above zero ground-truth relevance. In our case, we set n as 5 (nDCG%5), as we select 5 candidate captions from the given list.

Then we calculate nDCG score, which is the ratio of the DCG score over the ideal ranking's DCG score.

$$NDCG = \frac{DCG}{IDCG}$$

3. Method

3.1. Wikipedia Image/Caption Dataset

Train Dataset.

The training data is provided in a tsv format, but in slightly different ways. Each tsv file is comprised of 18 columns, containing various information including the image URL, caption title and the description of the image, and the written language type and others. The image pixels and the ResNet embeddings are provided separately in a compressed csv format(200 image pixel files/ 215 ResNet embedding files). The sheer size of the data makes it expensive to compute for this challenge. Thus, the research team has to find a more efficient way to read and compute the data. Total Train Data Size is 346.15GB.

Test Dataset.

The test data is also provided in both tsv and csv formats. The tsv data contains the id and its corresponding image URL. The csv data, like its train counterpart, contains image pixel information and the ResNet embeddings. Most importantly, a list of the test caption candidates are provided, which makes this challenge more similar to a text-image match/retrieval task. (Total Test Data Size: 3.27GB)

Data Preprocessing.

The raw train data consists of $\sim 37M$ rows. As depicted above, it is composed of text(caption), and image(pixel) subparts. However, a cross-examination showed that many of the training data have missing values from each subpart. Excluding such partially missing data, the train set shrinks to the size of $\sim 27M$ rows. Additionally, some of the data included damaged byte arrays, thus excluding the damaged byte arrays. Finally, we performed a train/valid split upon the remaining data. In the end, the training data is of the size of $\sim 26.9M$ caption+image sets, and the valid data is of the size of 134K rows.

3.2. A simple method: Levenshtein distance with image URL

The simple baseline would be to use the image URLs. This method does not require large computation. First, we cleanse the URLs of test images, and only the 'meaningful' words are left as a result. Then, we measure the similarity between the extracted URLs and the candidate list by calculating Levenshtein Distance. Finally, we selected the top 5 candidates per image, in order of similarity. Even though URL-based assumption is simple and naive, it is still noteworthy to refer. Thus, we will partially utilize the method in our final model.

Also, we tried to measure the similarity with different metadata and submit it to Kaggle. But there was no significant result, so we omit the explanation about other metadata.

3.3. Contrastive Learning-based Text-Image Matching

In the simple framework for contrastive learning of visual representations (SimCLR) [24], the purpose is to maximize agreement between differently augmented views stemmed from the same data example in the latent space. Two separate data augmentation operators sampled from the same family of augmentations ($t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$) and applied to each data example to obtain two correlated views. Thus, a minibatch of N examples is augmented to $2N$ data examples ($\tilde{x}_{2k-1} = t(x_k)$ and $\tilde{x}_{2k} = t'(x_k)$). SimCLR treats augmented images from the same data example as positive pairs and the other $2(N-1)$ as negative examples. The loss function for contrastive learning is defined as

$$\mathcal{L} = \frac{1}{2N} \sum_{k=1}^N [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$$

where $\ell(i, j)$ (loss function for a positive pair of examples) is defined as

$$\ell(i, j) = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}$$

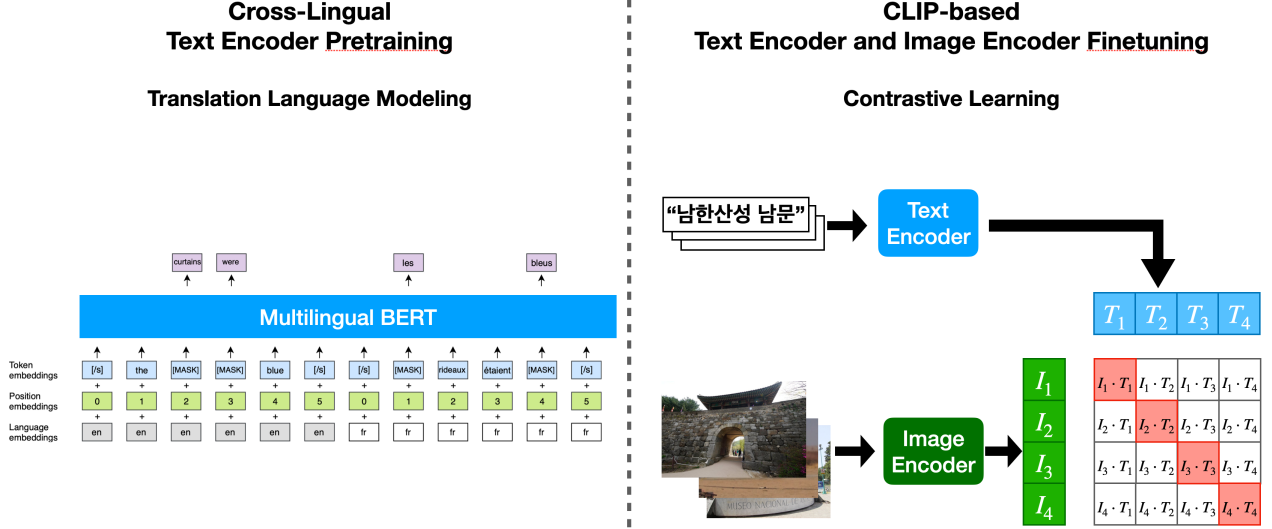


Figure 1. An overview of multilingual CLIP matching text and images. The training phase is divided into: 1. Teacher learning-based text encoder finetuning; 2. Contrastive learning-based text encoder and image encoder finetuning.



Figure 2. A sample image of the test data.

	caption_title_and_reference_description
0	Dodge Challenger [SEP] Heckansicht
1	Oregon State Police [SEP] A Dodge Charger of the Oregon State Police in Portland on I-5 in September 2012.
2	Streetcars in New Orleans [SEP] Streetcars on the Canal Street line.
3	Capitol Corridor [SEP] A typical Capitol Corridor train with a Charger locomotive and California Cars and a Superliner car.
4	Streetcars in New Orleans [SEP] Streetcars on the Canal Street line.

Figure 3. A sample of the test caption candidates. The task is to select five candidates(similar to the above image) from the caption list.

Let $\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v} / \|\mathbf{u}\| \|\mathbf{v}\|$ denote the dot product between ℓ_2 normalized \mathbf{u} and \mathbf{v} , $\mathbb{1}_{[k \neq i]} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $k \neq i$, τ denote a temperature parameter, and z_i denotes the projected features of augmented data \tilde{x}_i (i.e. $z_i = g(f(\tilde{x}_i))$), where f is a data encoder and g is a projection network.

In the CLIP [2], the contrastive learning extends to the multimodal contexts: images and text. Rather than considering two differently augmented views, the image and corresponding caption are considered as positive pairs, and the other combinations of images and captions are negative examples. To adapt original denotement of contrastive loss function, let z_{2k-1} correspond with projected features of image x_i (i.e. $z_{2k-1} = g_{\text{img}}(f_{\text{img}}(x_i))$) and z_{2k} correspond with projected features of corresponding caption y_i (i.e. $z_{2k} = g_{\text{text}}(f_{\text{text}}(y_i))$). By computing cosine similarity between projected features of images and caption, we can easily find the most matched caption for the image.

It is well-known that contrastive learning takes advantage of the large number of batch size [24], and it approximately took 255k GPU hours to train ResNet50-based CLIP in the original work. Obviously, finetuning the pre-trained network is the only choice. The challenge is that pre-trained text encoder in CLIP is trained in english-only contexts while our task is meant to match images with multilingual captions in language-imbalanced dataset. To solve this problem, our multilingual image/text matching architecture is shown as Fig 1. Firstly, text encoder is finetuned to minimize the loss

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N \text{MSE}(f_{\text{text}}(x_i), f_{\text{text}}(\mathcal{M}_{l(x_i), t}(x_i)))$$

Method	Score	Rank
Simple method: just using URL	0.18064	91
Vanilla CLIP	0.17924	93
Ensemble Method	0.33685	18

Table 1. **Score Result for each method** – A total of 105 teams participated.

where $\mathcal{M}_{l(x_i),t}$ denotes pretrained models that translates language of x_i to target language t , and f_{text} is the text encoder, which is inspired from knowledge distillation [25]. After finetuning text encoder in the multilingual dataset, the image encoder and text encoder are trained simultaneously by CLIP-based contrastive learning.

CLIP is trained on english-only context, however, this competition consists of multilingual captions. Thus, we tried to replace original text encoder with multilingual text encoder. We implemented the XLM [26]/XLM-R [6] model that achieves SOTA performance in cross-lingual understanding, which is pretrained by Translation Language Modelling (TLM). TLM can be considered as the extension of Masked Language Modelling (MLM), which predicts masked words by referring paired but different languages.

We’ve done 2 stages of finetuning (using multimodal contrastive learning). We first finetuned the visual encoder while freezing the language model. Lastly, we finetuned both visual and language models.

Additionally, we used data augmentation so that it can help the contrastive learning process. The RandAugment [10] allowed automated the data augmentation task with significantly low computational cost. We have searched some combinations of the number of augmentation (n) and the magnitude of augmentation (m). Although the decreases of loss are similar, we chose $n = 3$ and $m = 1$ for our data augmentation, which performs slightly quicker decrease in our model.

To maximize the score, we partially used the ’url-based’ method as a trick. As one image can have up to 5 candidate captions, we chose 3 candidate captions from our model-based predictions and 2 from the url-based predictions.

4. Results

4.1. Final Score

Table 1 shows the experimental results evaluated by the Kaggle estimator. The range of score is 0 to 1.

First, using the simple method and calculating the Levenshtein distance with the image URLs, Kaggle estimated the score to be about 0.18064. It would be a baseline score.

The Vanilla CLIP model gets 0.17924 point, which is

even slightly lower than the simple method. This is presumably because the data domain used in pre-training and the Wikipedia data domain is very different from the vanilla clip model, although it is a very good model in Image-Text representation learning.

Our method, the ensemble method, recorded 0.33685 points, earning 18th place of a total of 150 teams. As mentioned above in the method section, this method uses URLs as clues. We had not expected the score to jump significantly compared to pure CLIP-based predictions. From this, we could assume that URLs serve as effective clues in this task.

4.2. Qualitative Result

In Figure 4, the Image-Text Matching output of our model is represented. Each of the 4 images is about Scottish Gaelic [27], Battle of Thermopylae [28], Kykkos watermill [29], and Kansai International Airport [30] in order.

1. The 1st image is about the distribution of Gaelic speakers in Scotland, in 2011. As shown in the matching result, it can be seen that the model captures the content which is the proportion of certain language usage in Scotland (scots) well.
2. The 2nd image is about the site of the Battle of Thermopylae. In every five sentences selected by our model, "Battle" or "Thermopylae"(also, in other languages) is appearing. Thus, it can be seen that not only did it find the right place for the image, but it also had the ability to process multilingual sentences.
3. The 3th image is about the Kykkos watermill. The model found the correct answer as the first candidate. The 4th and 5th sentences are both contain information about the stone bridge, so it is not the correct answer. However, it is consistent with what is shown in the image, so it can be confirmed that the model understands the content of the image.
4. The 4th image is about the Kansai International Airport, so our model did not find the correct answer. However, it is clear that the image is an airport photo, and the model managed to find a related sentence.

4.3. Discussion

Here, we would like to explain why our model does not perform as expected.

As is well known, most of the Wikipedia pages contain proper nouns. However, since our general-purpose model like CLIP does not have appropriate methods of embeddings or representations for these proper nouns, it was not easy to learn enough from the given data alone. Although the amount of training data is huge, it is difficult to learn from the pages containing proper nouns or quite a specific range of topics unless the same entity name is exactly present in the training data.

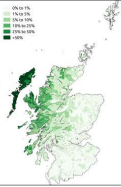



Image	Top-5 caption_title_and_reference_description
	<ul style="list-style-type: none"> • Northern Ireland [SEP] Percentage of people aged 3+ claiming to have some ability in Ulster Scots in the 2011 census • Languages of Northern Ireland [SEP] Percentage of people aged 3+ claiming to have some ability in Ulster Scots in the 2011 census • Scots language [SEP] The proportion of respondents in the 2011 census in Scotland aged 3 and above who stated that they can speak Lowland Scots • Catholic Church in Scotland [SEP] Percentage claiming to be Catholic in the 2011 UK Census in Scotland • Шкотски гелски језик [SEP] Процент распрострањености шкотског гелског језика у Шкотској према попису из 2011.
	<ul style="list-style-type: none"> • "Thermopylae [SEP] View of the Thermopylae pass from the area of the Phocian Wall. In ancient times, the coastline would have been much closer to the mountain, near the road to the right. This is a result of sedimentary deposition." • Thermopylés [SEP] Thermopylés eli Thermopylai. • Thermopylae [SEP] Thermopylae o ardal Mur y Phociaid. Yn yr hen amser roedd y môr yn cyrraedd bron at droed y mynydd. • "Bataille des Thermopyles [SEP] Le site de la bataille. Dans l'Antiquité, le rivage était approximativement au niveau du chemin ..." • "Battle of Thermopylae [SEP] The site of the battle today. Mount Kallidromon on the left, and the wide coastal plain formed ..."
	<ul style="list-style-type: none"> • Kykkos watermill [SEP] Kykkos Watermill • "Onoz, Namur [SEP] Watermill in Onoz" • Khust [SEP] War memorial • Puente Colgante de São Vicente [SEP] Puente Colgante de São Vicente 1910-1915 • Καστανιάνη Ιωαννίνων [SEP] Πέτρινο γεφύρι στην Καστανιά
	<ul style="list-style-type: none"> • Belfast International Airport [SEP] Belfast International Airport • Hong Kong International Airport [SEP] Sky pier • Marshall Islands International Airport [SEP] Welcome • Hercilio Luz International Airport [SEP] The new terminal's check-in hall in 2019 • Copernicus Airport Wrocław [SEP] Interior of Main Terminal

Figure 4. Image-Text Matching Result over first 4 images in test data.

In addition, since the overall quality of the data was poor as seen in many missing columns and the given period of time to this team is too short compared to the amount of data, we did not have enough opportunities to learn enough. And, considering the nature of contrastive learning, restraints from the limited computation resources on increasing the batch size has a great effect on learning and final performance.

5. Conclusions

We have presented an ensemble method with Text-Image Matching, contrastive learning, and calculating similarity between metadata and a candidate list as a method to solve the Wikipedia image/caption matching task. For Text-Image Matching, the CLIP model is employed for the base model. In the Wikipedia Image/Caption Matching task, we scored 18 out of 105 with a score of 0.33685. Although the score is not very high, considering the relatively shorter participation period, it is a remarkable performance. If we had had enough time to study longer, it would be expected that we could get a higher score.

References

- [1] Kaggle wikipedia - image/caption matching. <https://www.kaggle.com/c/wikipedia-image-caption/data>. Accessed: 2021-12-09. 1
- [2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 2, 4
- [3] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives, 2018. 1
- [4] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction, 2019. 1
- [5] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching, 2018. 1, 2
- [6] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020. 1, 5
- [7] Yu Liu, Yanming Guo, Erwin M. Bakker, and Michael S. Lew. Learning a recurrent residual fusion network for multi-modal matching, 2017. 2
- [8] Jiuxiang Gu, Jianfei Cai, Shafiq Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models, 2018. 2
- [9] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching, 2020. 2

- [10] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data, 2019. 2, 5
- [11] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment, 2019. 2
- [12] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space, 2019. 2
- [13] Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, 17(11):1875–1886, Nov 2015. 2
- [14] 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, 2018. 2
- [15] 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, 2020. 2
- [16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 2
- [17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 2
- [18] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa, 2020. 2
- [19] Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training, 2021. 2
- [20] Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Tarooh Bharti, Lijuan Wang, Jianfeng Gao, Dongdong Zhang, and Nan Duan. M3p: Learning universal representations via multitask multilingual multimodal pre-training, 2021. 2
- [21] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. 2
- [22] Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual bert: An empirical study, 2020. 3
- [23] Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert, 2019. 3
- [24] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 3, 4
- [25] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 5
- [26] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining, 2019. 5
- [27] Scottish gaelic. https://en.wikipedia.org/wiki/Scottish_Gaelic. 5
- [28] Battle of thermopylae. https://en.wikipedia.org/wiki/Battle_of_Thermopylae. 5
- [29] Kykkos watermill. https://en.wikipedia.org/wiki/Kykkos_watermill. 5
- [30] Kansai international airport. https://en.wikipedia.org/wiki/Kansai_International_Airport. 5