Exploring the Relationship between Dataset Size and Image Captioning Model Performance

Tomáš Železný[1], Marek Hrúz[1]

[1]Department of Cybernetics and New Technologies for the Information Society, Technická 8, 301 00 Plzeň, Czech Republic

Abstract

Image captioning is a deep learning task that involves computer vision methods to extract visual information from the image and also natural language processing to generate the result caption in natural language. Image captioning models, just like other deep learning models, need a large amount of training data and require a long time to train. In this work, we investigate the impact of using a smaller amount of training data on the performance of the standard image captioning model Oscar. We train Oscar on different sizes of the training dataset and measure its performance in terms of accuracy and computational complexity. We observe that the computational time increases linearly with the amount of data used for training. However, the accuracy does not follow this linear trend and the relative improvement diminishes as we add more data to the training. We also measure the consistency of individual sizes of the training sets and observe that the more data we use for training the more consistent the metrics are. In addition to traditional evaluation metrics, we evaluate the performance using CLIP similarity. We investigate whether it can be used as a fully-fledged metric providing a unique advantage over the traditional metrics; it does not need reference captions that had to be acquired by human annotators. Our results show a high correlation between CLIP with the other metrics. This work provides valuable insights for understanding the requirements for training effective image captioning models. We believe our results can be transferred to other models, even in other deep-learning tasks.

Keywords

Image captioning, deep learning, computer vision, machine learning, data size analysis

1. Introduction

Image captioning is a task in computer vision that involves generating a textual description of an image. The goal is to provide a comprehensive and human-like description of the content of an image, which can be useful for a variety of applications, such as enabling individuals with visual impairments to better understand visual information, improving the accuracy and relevance of image search results, etc. It is a complex task because it requires the identification and interpretation of visual information, as well as the generation of grammatically correct and fluent sentences. This requires a combined effort of computer vision and natural language processing methods.

The scientific community has been interested in this task for over a decade [1]. The methods used for this task were relying on hand-crafted features and rule-based algorithms. Recent advances in machine learning and artificial intelligence have enabled the development of more effective image captioning models, which are able to generate high-quality captions for a wide range of images.

An important feature of image captioning is that there is not only one correct caption for an image. This is because different individuals may consider different aspects of an image to be important, and they may therefore describe the image in different ways. Because of this, there is not one ideal evaluation metric that can be used to measure the quality of a generated caption, as different metrics may be better suited for evaluating different attributes of the caption.

The general problem of deep learning is that it requires a large amount of data and the training process can be computationally intensive. In this work, we investigate the relationship between the size of the training dataset and the performance of a standard image captioning model, Oscar [2]. We train Oscar on different sizes of the training dataset and measure the performance by means of accuracy and also computational complexity. We expect this dependency to have linear behavior, where increasing the size of the training dataset will result in a corresponding increase in computational time. This research is important because it can help us understand the limitations of deep learning models and the computational resources required to train them effectively. Additionally, our results can provide valuable insights for future research on image captioning and other applications of deep learning.

Our contribution in this work is an experiment that confirms the expected behavior of the Oscar model, i.e., linear dependence. We also provide insight into the relationship between the size of the training dataset and

26th Computer Vision Winter Workshop, Robert Sablatnig and Florian Kleber (eds.), Krems, Lower Austria, Austria, Feb. 15-17, 2023

zeleznyt@kky.zcu.cz (T. Železný); mhruz@ntis.zcu.cz (M. Hrúz)

0000-0002-0974-7069 (T. Železný); 0000-0002-7851-9879 (M. Hrúz)

its performance on selected metrics. Furthermore, we measure the consistency of the data for each of the metrics used, and we expect that smaller subsets of the data will have higher variance than larger subsets. Our results will help to better understand the requirements for training effective image captioning models and the potential trade-offs between dataset size and performance. Additionally, our findings may be useful for researchers and practitioners who are interested in optimizing the training of deep learning models in general.

In addition to using state-of-the-art evaluation metrics, we also evaluate our image captioning methods on CLIP (Contrastive Language-Image Pre-training) similarity [3]. We investigate whether CLIP can be used as a full-fledged evaluation metric for image captioning. We find that it has a major advantage over traditional metrics: it does not require reference labels from annotators. This means that CLIP can be used to evaluate image captioning models in an unsupervised or self-supervised manner, which can be useful in situations where annotated data is not available or is too expensive to obtain.

2. Related Work

2.1. Datasets

Image captioning models are trained on large datasets consisting of pairs of images and captions. These datasets may differ in terms of the domain they cover, the number of image-caption pairs they contain, and the number of captions per image.

One well-known dataset for image captioning is Flickr30k [4], which includes approximately 31,000 images of everyday scenes, each described by five independent annotators, resulting in 155,000 image-caption pairs. Another popular dataset is COCO Captions [5], which contains over 164,000 images of everyday scenes, with five annotations per image, for a total of over 820,000 image-caption pairs. The Conceptual Captions dataset [6] comprises images collected from a large number of web pages, with one caption per image extracted from the alt-text HTML attribute. This dataset contains over 3,000,000 image-caption pairs. Conceptual12m [7] is a similar dataset, also extracted from web pages, with a total of over 12,000,000 image-caption pairs.

Each of these datasets has its own advantages and disadvantages. For instance, the Flickr30k dataset has a good consistency and is well-suited for evaluation due to the multiple reference captions provided for each image. It is a valuable feature because a single image can often be described in multiple ways, and it is useful to have a diverse set of captions for each image to better capture the range of possible descriptions. However, the quality of datasets containing images collected from the internet, such as Conceptual Captions and Conceptual12m, may depend on the filtering applied during collection, and their consistency may be harder to guarantee. These datasets, however, offer a larger number of images and a greater variance. As a result, state-of-the-art image captioning models often utilize a combination of multiple datasets in order to achieve the best performance. In this work, we chose to use the COCO Captions dataset for our experiments due to its suitable size for training and also dividing into subsets. The COCO Captions dataset also has a sufficient number of images to allow for a robust evaluation of the model's performance.

2.2. Evaluation

The evaluation of image captions is a challenging task due to the inherent subjectivity of language and the multiple ways in which an image can be correctly described. Most evaluation metrics for image captioning compute the difference between a candidate caption and a reference caption provided by human annotators. Traditional metrics, such as BLEU [8], ROUGE [9], METEOR [10], and CIDEr [11], are based on the positions of n-grams in the candidate and reference captions. More advanced metrics, such as SPICE [12], measure the semantic similarity between the captions using graph-based representations.

Individual metrics may be suitable in different situations. For example, BLEU is a simple and inexpensive metric to compute, but it does not perform well when compared to other metrics [13]. On the other hand, CIDEr is considered to be the best-performing metric that compares n-grams in candidate and reference captions. However, it requires the entire dataset to be computed, making it computationally expensive for larger datasets. SPICE is a popular metric that compares the semantics of the captions rather than their syntax. However, it requires a complex model to accurately capture semantic relationships, making it computationally expensive.

In tasks of image generation, the Fréchet inception distance (FID) [14] is used to evaluate the quality of images generated by a generative model, such as a generative adversarial network (GAN) [15]. Similarly, CLIP [3] can be used to assess the similarity between an image and text. CLIP is a deep learning model developed by OpenAI that is able to encode the image and text into a common semantic space. The cosine similarity can then be used to compute the agreement between the input text and the image. Also, diffusion models for generating images use CLIP [16] to evaluate the generated image based on text input. In image captioning, CLIP can be used to evaluate the generated caption. Although CLIP has not been considered a standard evaluation metric for image captioning, in this study we present it as such. In this study, we present it as a potential fully-fledged metric that thoroughly assesses the semantic quality of candi-

1% sub_{01}	a group of brown cows standing in a field	25% sub_{01}	a cow that is laying down in the grass.
1% sub_{02}	a group of cows that are standing together.	25% sub_{02}	a cow is standing in a field with another cow behind it.
1% sub_{03}	a group of cows are standing in the grass.	25% sub_{03}	a cow is standing in a field with another cow.
1% sub_{04}	a herd of black and white cows in a field.	25% sub_{04}	a cow with a red ear tag standing in a field.
1% sub_{05}	a group of cows stand together in a grassy area.	25% sub_{05}	a black and white cow standing in a field.
1% sub_{06}	a herd of cows standing in a field.	25% sub_{06}	a cow is standing in the grass with another cow behind it.
1% sub_{07}	a group of cows grazing on a field.	25% sub_{07}	a cow is standing in a field with another cow behind it.
1% sub_{08}	a group of brown cows laying in a field	25% sub_{08}	a cow is standing in a field of grass.
1% sub_{09}	a couple of cows standing together in a field.	25% sub_{09}	a cow is standing in a field with other cows.
1% sub_{10}	two cows in a field with a fence surrounded by green grass.	25% sub_{10}	two cows are laying down in a field.

Figure 1: Examples of generated captions for the same image. On the left side, there are captions from different models trained on the 1% subset of data. On the right, there are captions from models trained on the 25% subset. We see that there is greater variability of the captions from the 1% subset, while the semantics are mostly correct.

date captions. We compute the correlation between CLIP and other metrics and investigate whether CLIP can be used in this manner. A previous research study [17] has conducted similar experiments, but focused on computing the correlation with human judgment and comparing it to correlations with other metrics, whereas we compute correlations with other metrics directly.

2.3. Image Captioning Methods

Recent advances in image captioning have seen the widespread adoption of deep learning techniques. Early methods used convolutional neural networks (CNNs) as encoders, such as the model proposed by [18]. More recent approaches have used Faster R-CNN [19] for object detection in images, leading to improved performance. The latest methods employ transformer architectures [20], which have achieved state-of-the-art performance on a variety of tasks. Among the best-performing methods are transformer-based methods Oscar [2], VinVL [21] and OFA [22], which use multimodal input. mPLUG [23] is another image captioning method that uses two unimodal encoders, one for images and one for text. These encoders are then combined using a cross-modal skip-connected network, which consists of multiple skip-connected fusion blocks.

3. Experiments

In this work, we investigate the performance and efficiency of the image captioning method Oscar [2]. Our motivation for using this specific method is that we have previously used it in our own experiments and found it to be a convenient method to use. While it may not currently be the best-performing model, Oscar is a transformer-based method and we believe that the results of our experiments may be generalizable to other transformer-based or deep-learning models in the field.

To assess the performance of Oscar, we conducted two main experiments. The first experiment involved measuring the time needed to train the model using various amounts of data while tracking the performance on a set of chosen evaluation metrics. In addition to traditional metrics, we also evaluated the model using CLIP similarity [3]. In the second experiment, we measured the correlation between the various metrics used in order to determine the potential use of CLIP as a fully-fledged metric in the image captioning field.

3.1. Method

Our experiments are based on the training and evaluation of the image captioning model Oscar [2]. Oscar is a transformer-based model, which uses a multimodal input. The input consists of feature vectors and tags of objects detected in the source image by an external object detector. The output is the predicted caption describing the source image.

The authors of Oscar provide a demonstration dataset of feature vectors and object tags that can be used as input to Oscar, but do not specify the method by which these object detections are obtained. In order to generate captions for custom images outside of the demonstration

dataset, we developed a full pipeline that takes a source image as input and produces a caption as output. According to [2], Oscar's input is a 2054-dimensional vector for each detected object, where the first 2048 dimensions are image features extracted from a detection network and the remaining 6 values contain the coordinates and size of the bounding box for the detected object. We used the Faster R-CNN detection network implemented in the Detectron2 [24] framework as the object detector. We used the R50-C4 backbone, which meets the requirements of having a 2048-dimensional vector in the final layer. We use the feature vector from this layer together with the predicted class as the input to Oscar. The Faster R-CNN model was pre-trained on the COCO dataset [25] and is used without any further fine-tuning for our task. The quality of our pipeline is definitely restricted by the quality of the detector. In our case, we are able to detect only 80 possible classes (COCO classes), which may hinder the expressivity of the model.

Analysis of the demonstration dataset provided by Oscar revealed that there are always at least 10 detections per image, with confidence scores higher than 0.2. Based on this finding, we configured the object detector in our pipeline to generate detections with confidence scores higher than 0.2, and to include detections with lower confidence scores if there are fewer than 10 detections in total. This ensures that the input to Oscar matches the format of the demonstration dataset.

3.2. Dataset

In this work, we conducted experiments using the COCO Captions [5] dataset. It consists of 164,062 images with 5 captions each, divided into the train, validation, and test sets. The annotation for the test set is not publicly available, so we redistributed the original train+val sets into our own train+val+test sets for evaluation on the COCO Captions dataset.

The demonstration dataset provided by Oscar also consists of images from the COCO Captions dataset, which is split into train+val+test sets that originally belonged to the original train+val COCO Captions dataset. We decided to follow this distribution, resulting in final train+val+test sets of 5,000+5,000+113,287 images.

3.3. Impact of Different Volumes of Data on Model Performance

In this experiment, we evaluate the performance of the Oscar image captioning model on the COCO Captions dataset. As described in Section 3.2, the dataset was split into training, validation, and test sets, with the validation and test sets remaining unchanged for evaluation purposes.

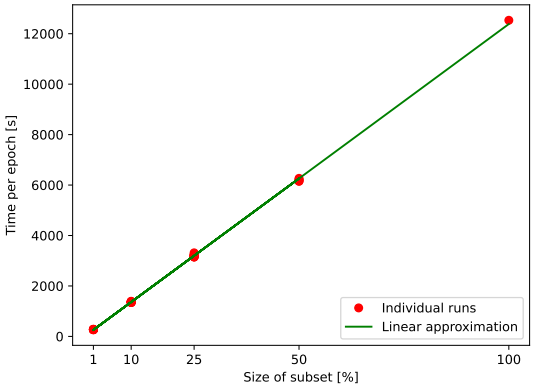

Figure 2: Relationship between average time elapsed per epoch and the subset size used for training. We see that the measured data confirm the expected behaviour, i.e. linear dependence.

1 %	a dog laying on top of a bed.
10 %	a dog is laying on a bed in a room.
25 %	a dog sitting on a bed next to a person.
50 %	a dog sitting on a bed with clothes and a book.
100%	a dog sitting on a bed with a blanket and a pillow.

Figure 3: Examples of captions generated from the best models of each subset of the data. We can see the improvement of the caption as we add more data.

To assess the effect of training data size on model performance, we selected various amounts of data from the training set to train Oscar. The sizes of the training subsets were 100%, 50%, 25%, 10%, and 1% of the original train set. For each subset size, multiple random selections were made from the full training set to measure the consistency of the selected data. The number of random selections for each subset size is shown in Table 1. The number of data selections was chosen to provide a sufficient number of samples to measure variance while also considering the computational resources available.

The Oscar model was trained using various sizes of training subsets for a total of 30 epochs, and the

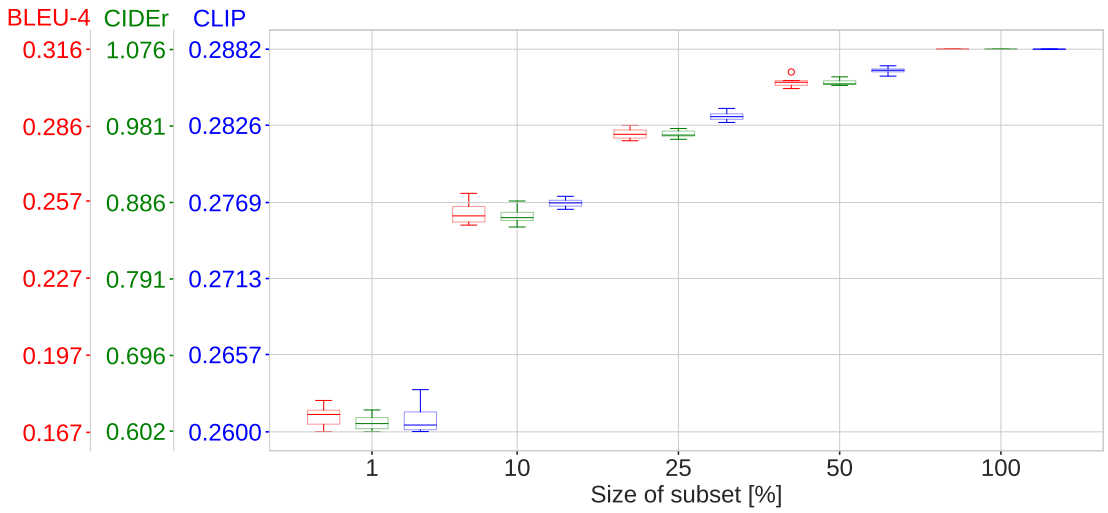

Figure 4: Relationship between the size of the training set used to train Oscar [2] and the score of BLEU-4, CIDEr and CLIP metrics obtained by evaluating trained Oscar on the test set. We use different axis for each metric to better visualize the trends in individual metrics for a clearer comparison. The variance of individual sets of given sizes is visualized by boxplots. We can see that the upper quartile of the smaller set does not intersect with the lower quartile of the larger set. Note that there is no variance for the 100% split because there was only one selection.

Table 1
Number of selections per subset size.

Subset size	100 %	50 %	25 %	10 %	1 %
Selections	1	5	10	10	10

elapsed time was recorded. Training was conducted using NVIDIA GeForce GTX 1080 Ti GPUs. The relationship between elapsed time and training subset size is shown in Figure 2. As expected, this relationship follows a linear dependence between data size and computational time.

During training, the model was evaluated on the validation set after every 5th epoch, and the best-performing checkpoint was saved. The CIDEr metric was used for this evaluation because it has been found to correlate well with human judgment [17] and Oscar uses it as its default output score. After training, the best-performing checkpoint was selected based on its performance on the validation set and then evaluated on the test set. The resulting score on the test set is shown in Figure 4.

In order to assess the consistency of evaluation results, we measured the variability of the metric scores for each subset size. The variability is visualized in the figure using boxplots, which allow us to see the variance of different metrics across the individual subsets. The non-overlapping quarters of the boxplots indicate that there is a statistically significant difference in the scores depending on the subset size. This highlights the importance of carefully considering the subset size in order to obtain reliable results. For qualitative assessment of this experiment see Figures 1 and 3.

3.4. Evaluating Image Captioning with CLIP

In the second experiment, we investigate whether CLIP similarity can be used as a fully-fledged metric for evaluating image captioning tasks. Our analysis of the data, as depicted in Figure 4, revealed that CLIP exhibits behavior similar to that of other metrics. To further investigate this relationship, we calculated Pearson's correlation coefficient between all metrics across all subsets of the data. The resulting correlations are presented in Figure 5.

Our findings show that all metrics are highly correlated. This indicates the correct, consistent, and expected behavior of all the metrics. In addition, we observed that the BLEU, METEOR, ROUGE, and CIDEr metrics tend to be on average more correlated with each other than with SPICE or CLIP. This trend is likely due to the fact that the former group of metrics compares the placement of n-grams in candidate and reference captions, while the latter two metrics do not consider syntactic content but rather focus on semantics.

The main takeaway is that CLIP is a viable metric for image captioning evaluation which does not need reference captions. This outcome is essential since it enables hypothetical training of a captioning system without references in an unsupervised manner.

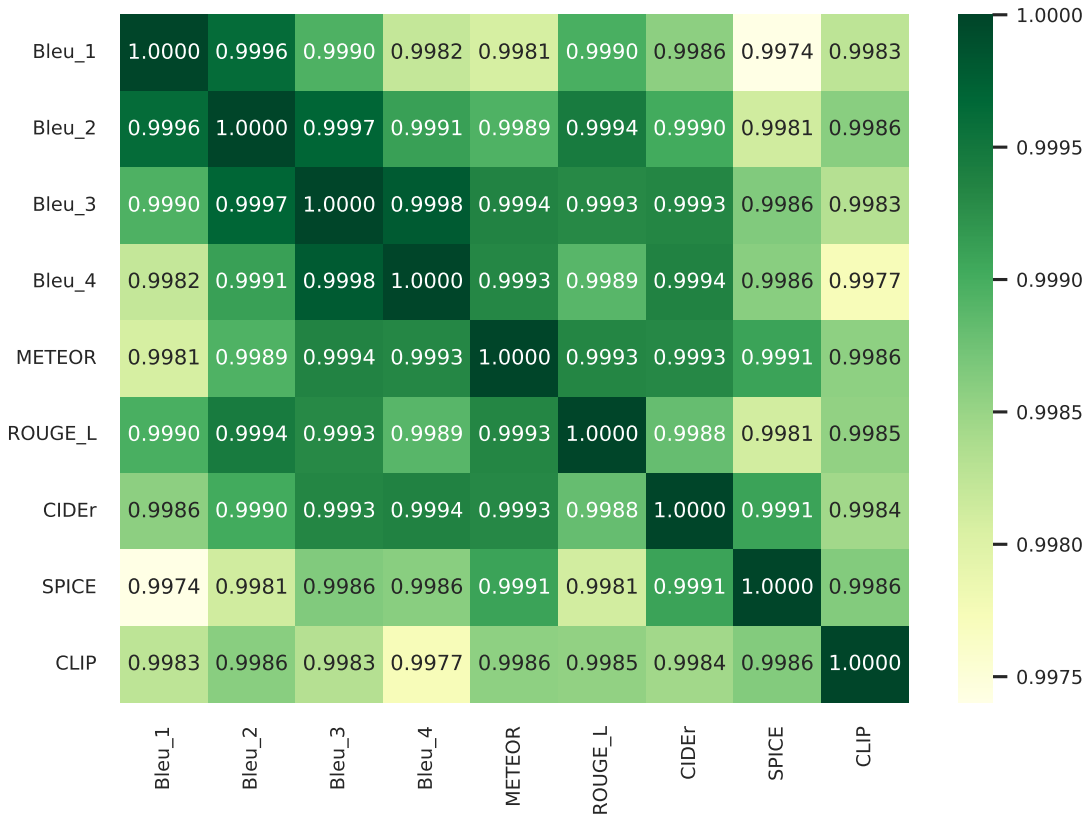

Figure 5: Pearson's correlation coefficient matrix computed pair-wise for all used metrics. We see that all the metrics highly correlate.

4. Conclusion

In our work, we conducted several experiments to analyze the training of the image captioning method Oscar. First, we trained the method on different sizes of training data. We measured the elapsed time of the training loop and the performance on given metrics. The training duration has a linear relationship with the volume of data that is used. Furthermore, we have measured the behavior of individual metrics based on the size of the training data. We measured the consistency of the data for individual subsets. We experimentally show that the models trained on smaller subsets have a higher variance of all the evaluation metrics than the models trained on larger sets. We observe that the scores converge to some value. However, the improvement of the individual metrics is not linearly dependent on the amount of data used for training. As we add more data for training, the improvement diminishes. This is affected by multiple phenomena: The first one is the capacity of the model itself, hence the convergence to a non-perfect value of the metrics. The second one is the quality of the dataset. We chose COCO Caption for multiple reasons. Because we believe it has good consistency - it contains scenes of everyday life with a limited variety of objects and because it has 5 annotations per image. Another reason is that it has good size - it is big enough to make an adequate 1% split from it, but it is also small enough for 36 training runs of 30 epochs to be computed in reasonable time on our GPUs. Lastly, the quality of the detector producing the detections and feature vectors affects the performance.

Based on our output, one can now decide to reduce the training data volume if the goal is to achieve a specific minimum score of a metric. It can be assumed, that the behavior will be similar to other models and datasets.

In our second experiment, we evaluated the correlation between various state-of-the-art metrics and the CLIP metric, which we believe, can be used as a fully-fledged metric for image captioning with its huge advantage - it does not need any reference captions. Our results showed that all the metrics including CLIP are highly correlated. This supports CLIP's potential use as a fully-fledged met-

ric for image captioning. Previous research [17] has also investigated the CLIP metric, focusing on the correlation with human judgment and comparing it to the correlation of other metrics. In comparing those results to ours, we found that the ranking of the correlation of individual metrics to human judgment corresponds to the ranking of the correlation of other metrics with CLIP.

Acknowledgments

The work has been supported by the grant of the University of West Bohemia, project No. SGS-2022-017. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic. Also, we would like to thank RNDr. Blanka Šedivá, Ph.D. for giving us the initial idea for this research.

References

[1] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. Forsyth, Every picture tells a story: Generating sentences from images, in: European conference on computer vision, Springer, 2010, pp. 15–29.

[2] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al., Oscar: Object-semantics aligned pre-training for vision-language tasks, in: European Conference on Computer Vision, Springer, 2020, pp. 121–137.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[4] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Transactions of the Association for Computational Linguistics 2 (2014) 67–78. doi:10.1162/tacl_a_00166.

[5] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, C. L. Zitnick, Microsoft coco captions: Data collection and evaluation server, arXiv preprint arXiv:1504.00325 (2015).

[6] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2556–2565.

[7] S. Changpinyo, P. Sharma, N. Ding, R. Soricut, Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, in: CVPR, 2021, pp. 3558–3568.

[8] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[9] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[10] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.

[11] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: CVPR, 2015, pp. 4566–4575.

[12] P. Anderson, B. Fernando, M. Johnson, S. Gould, Spice: Semantic propositional image caption evaluation, in: European conference on computer vision, Springer, 2016, pp. 382–398.

[13] Y. Cui, G. Yang, A. Veit, X. Huang, S. Belongie, Learning to evaluate image captioning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5804–5812.

[14] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Advances in neural information processing systems 30 (2017).

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM 63 (2020) 139–144.

[16] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, arXiv preprint arXiv:2204.06125 (2022).

[17] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, Y. Choi, Clipscore: A reference-free evaluation metric for image captioning, arXiv preprint arXiv:2104.08718 (2021).

[18] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: CVPR, 2015, pp. 3156–3164.

[19] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in neural information processing systems 28 (2015).

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, At-

tention is all you need, Advances in neural information processing systems 30 (2017).
[21] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, J. Gao, Vinvl: Revisiting visual representations in vision-language models, in: CVPR, 2021, pp. 5579–5588.
[22] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, H. Yang, Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, arXiv preprint arXiv:2202.03052 (2022).
[23] C. Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu, Z. Cao, et al., mplug: Effective and efficient vision-language learning by cross-modal skip-connections, arXiv preprint arXiv:2205.12005 (2022).
[24] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, R. Girshick, Detectron2, https://github.com/facebookresearch/detectron2, 2019.
[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.