Cross-Modal Similarity-Based Curriculum Learning for Image Captioning

Hongkuan Zhang[1] **Saku Sugawara**[2] **Akiko Aizawa**[2]
Lei Zhou[1] **Ryohei Sasano**[1] **Koichi Takeda**[1]
[1]Nagoya University [2]National Institute of Informatics
{zhang.hongkuan.k5,zhou.lei.e1}@s.mail.nagoya-u.ac.jp
{saku,aizawa}@nii.ac.jp {sasano,takedasu}@i.nagoya-u.ac.jp

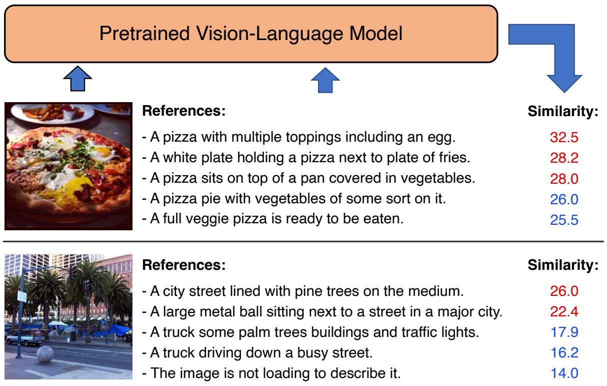
Abstract

Image captioning models require the high-level generalization ability to describe the contents of various images in words. Most existing approaches treat the image–caption pairs equally in their training without considering the differences in their learning difficulties. Several image captioning approaches introduce curriculum learning methods that present training data with increasing levels of difficulty. However, their difficulty measurements are either based on domain-specific features or prior model training. In this paper, we propose a simple yet efficient difficulty measurement for image captioning using cross-modal similarity calculated by a pretrained vision–language model. Experiments on the COCO and Flickr30k datasets show that our proposed approach achieves superior performance and competitive convergence speed to baselines without requiring heuristics or incurring additional training costs. Moreover, the higher model performance on difficult examples and unseen data also demonstrates the generalization ability.

1 Introduction

Image captioning has been widely investigated in computer vision and language research. However, most current methods treat image–caption pairs for training indistinctively, thus neglecting the difference in terms of learning difficulty. As illustrated in Figure 1, an image is annotated with multiple references with diverse styles and complexity levels. Such diversity can introduce different levels of learning difficulty, and undertrained captioning models can be misled by wrong gradients when training on the difficult data (Dong et al., 2021).

Curriculum learning (CL) has demonstrated improvements in model performance and training speed by presenting data sorted according to the learning difficulty (Bengio et al., 2009). Existing image captioning approaches using CL have drawbacks in their difficulty measurements: 1) Requiring domain-specific knowledge or heuristics (Liu et al., 2021); 2) Adding up the mono-modal difficulty scores without considering the cross-modal features (Alsharid et al., 2021); and 3) Requiring additional computational resources to train models on the target data (for the cases of bootstrapping methods) (Liu et al., 2021; Dong et al., 2021).



References:	Similarity:
- A pizza with multiple toppings including an egg.	32.5
- A white plate holding a pizza next to plate of fries.	28.2
- A pizza sits on top of a pan covered in vegetables.	28.0
- A pizza pie with vegetables of some sort on it.	26.0
- A full veggie pizza is ready to be eaten.	25.5

References:	Similarity:
- A city street lined with pine trees on the medium.	26.0
- A large metal ball sitting next to a street in a major city.	22.4
- A truck some palm trees buildings and traffic lights.	17.9
- A truck driving down a busy street.	16.2
- The image is not loading to describe it.	14.0

Figure 1: Example of cross-modal similarity score for the caption data calculated by the pretrained VL model CLIP. Numbers with RED and BLUE colors denote the higher and lower scores respectively.

We propose a simple yet efficient difficulty measurement using a pretrained vision–language (VL) model. Most VL models are pretrained with image–text matching tasks, which involve the calculation of the cross-modal similarity. The similarity reveals the model confidence in the image–text data relevance and a lower score indicates hard-to-determine or low-quality data (Lee et al., 2021; Hessel et al., 2021). As shown in Figure 1, the VL model assigns higher scores to the highly relevant image–caption pairs usually with simple images and appropriate captions, while assigns lower scores to less relevant pairs often with complex images and low-quality captions. We consider the pairs with higher scores to be easier to learn and train the captioning model with training examples presented from easy to hard.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7599 - 7606
December 7-11, 2022 ©2022 Association for Computational Linguistics

In our experiments on the COCO and Flickr30k datasets, models trained with our similarity-based CL achieve superior performance and convergence speed without domain-specific heuristics or additional pretraining cost. Moreover, using a VL pretrained model possessing the knowledge of the target data can further improve the performance. We also evaluate the trained model on difficult examples and unseen data, and the better performance of our method demonstrates the generalization ability. Last, our method brings higher improvement when applied to a smaller model, suggesting its applicability to scenarios in which fine-tuning large models is unfeasible.

2 Related Work

Curriculum Learning (CL) CL is a method to train a model with sorted data to improve generalization and accelerate convergence. It has been explored in neural machine translation (Platanios et al., 2019; Liu et al., 2020), relation extraction (Huang and Du, 2019), and natural language understanding (Xu et al., 2020) in the language field, or image classification (Wang et al., 2019; Xiang et al., 2020) and semantic segmentation (Wei et al., 2016; Huang and Du, 2019) in the vision field. We focus on the efficient data difficulty measurement of CL methods for image captioning.

Mono-Modal Difficulty Measurement Difficulty measurement can be classified into predefined and automatic methods. Predefined methods require heuristics based on the data feature, such as the variety (Bengio et al., 2009) or number (Wei et al., 2016) of objects in an image and the length (Spitkovsky et al., 2010) or word rarity (Platanios et al., 2019) in a sentence, to measure the data complexity for image classification and language generation. In contrast, automatic methods usually adopt a teacher model for difficulty measurement based on cross entropy (Weinshall et al., 2018; Xu et al., 2020) or perplexity (Zhou et al., 2020b) to determine the model confidence and uncertainty.

Cross-Modal Difficulty Measurement Difficulty measurement for caption data requires to consider the visual and language modalities. Alsharid et al. (2021) directly added visual and textual difficulty scores measured by Wasserstein distance and TF-IDF respectively for ultrasound image captioning. Similarly, Liu et al. (2021) used both domain-specific heuristics and entropy from the bootstrapping model trained on the target data as difficulty measurements for medical report generation. In addition, Dong et al. (2021) trained several bootstrapping models to evaluate generated captions with the BLEU score as the image difficulty. Unlike these works, we propose an efficient measurement based on cross-modal similarity to improve the model performance in the general domain.

3 Methodology

3.1 Cross-modal Similarity

To calculate the cross-modal similarity, we use either CLIP (Radford et al., 2021) pretrained on image–text pairs from the web or ViLT (Kim et al., 2021) pretrained on labelled image-caption pairs for comparison. Specifically, given an image $X = (x_1, ..., x_P)$ with P patches and a text $Y = (y_1, ..., y_T)$ with T tokens, CLIP encodes each modality with individual encoder to obtain the visual feature $\mathbf{x}$ and textual feature $\mathbf{y}$ respectively. Then the similarity is calculated as:

$$D_{\text{CLIP\_sim}} = \cos(\mathbf{x}, \mathbf{y}). \tag{1}$$

While for ViLT, image and text inputs are concatenated with a prepended $[\text{class}]$ token, and the inputs are encoded by a cross-modal encoder as:

$$\text{ViLT}(X, Y) = x'_{[\text{class}]}, x'_1, ..., x'_P, y'_1, ..., y'_T. \tag{2}$$

The joint representation $x'_{[\text{class}]}$ is then given to a pretrained fully-connected layer which is denoted as FFN to calculate the similarity as:

$$D_{\text{ViLT\_sim}} = \text{sigmoid}(\text{FFN}(x'_{[\text{class}]})). \tag{3}$$

3.2 Training Schedule

With the sorted dataset, we need to schedule when and how much harder data should be given during the training. Here we use the Baby Step learning (Spitkovsky et al., 2010) as our training schedule. The sorted dataset is equally divided into L buckets, and the model is trained with the easiest bucket first. When the model performance on the validation set does not improve over several epochs, we consider the model has converged and then merge the harder bucket with current buckets to continue the training. Training terminates when all the buckets are used and the maximum number of training epochs is reached. In the experiments, we apply this training schedule to all the CL methods, and adjust the optimal number of buckets based on the model performance on the validation set. We use the notation Simi-CL for our proposed similarity-based CL.

3.3 Baseline Approaches

Addup-CL This method simply adds the difficulty scores of two modalities, and we use pretrained models to measure the difficulty score of each modality for a fair comparison. Specifically, we use pretrained object detector BUTD (Anderson et al., 2018) for visual difficulty D_v and language model GPT-2 (Radford et al., 2019) for textual difficulty D_t, and take the weighted sum to obtain the adding up difficulty D_{addup}:

$$\begin{aligned} D_v &= -\sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n}\log p_{k,n}, \\ D_t &= -\sum_{t=1}^{T}\log p(y_t|y_{<t}), \\ D_{\text{addup}} &= \lambda \times D_v + (1-\lambda) \times D_t, \end{aligned} \tag{4}$$

where K denotes the top-K detected boxes with the highest confidence score, N denotes the detected object classes, $p_{k,n}$ is the probability of the n-th class for the k-th box, and λ denotes the weight.

Bootstrap-CL This method requires training a model with target data in advance to provide the difficulty score. Specifically, we train the captioning model on each dataset with the regular strategy, then calculate the cross-entropy loss using the trained model as follows:

$$D_{\text{bootstrap}} = -\sum_{t=1}^{T}\log p(y_t|y_{<t}, X). \tag{5}$$

4 Experiments

4.1 Settings

We performed experiments on the COCO and Flickr30k datasets and adopted the Karparthy splitting strategy (Karpathy and Fei-Fei, 2015), obtaining 113k/5k/5k and 29k/1k/1k for the training/validation/test sets, respectively. We implemented the captioning model as a vanilla Transformer based on the publicly available codes (Luo et al., 2018), and set the batch size to 10, learning rate to 3e-4, and dropout rate to 0.4 for all the experimental settings.

For CL-related settings, the split numbers L for the Baby Step learning were empirically determined to be 5 and 3 for COCO and Flickr30k, respectively. About the hyperparameters in Addup-CL, the number of the object detection classes N is 1600 and we use the top-10 confident boxes to calculate the difficulty, and the weight λ was set to 0.6 after the parameter tuning. For the similarity calculation, we used the base version of CLIP as the default model, and compare it with two versions of ViLT models which were fine-tuned on the COCO (`ViLT-CC`) and the Flickr30k (`ViLT-FL`) respectively by ViLT authors. We evaluated the performance with the COCO API and focused on four metrics: BLEU-4, METEOR, CIDEr, and SPICE.

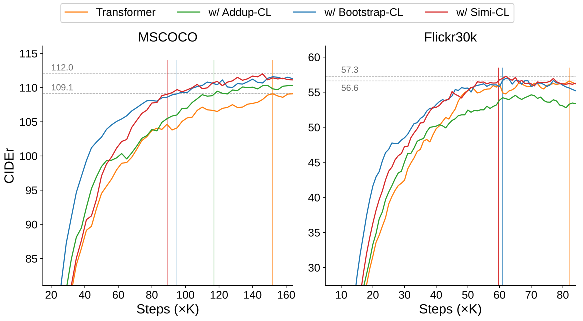


Figure 2: Model performance variation on the validation sets of both datasets during the training.

4.2 Main Results

The performance on the validation set is shown in Figure 2. For COCO, all the CL methods improve the performance, and accelerate the convergence speed towards the best vanilla model performance. Particularly, Simi-CL achieves better performance than Bootstrap-CL without additional training cost, and both methods outperform Addup-CL. For the Flickr30k dataset, we observe a similar phenomenon but with smaller improvement and Add-CL fails to improve the performance, which indicates that CL method is more efficient for the larger dataset. Since the vocabulary size of COCO is larger than Flickr30k (9,487 vs. 7,000), we suppose the difficulty measurement is efficacy for more diverse data, which requires further investigations.

For the model performance on the test sets listed in Table 1, the performance of CL-based models is consistent with that for the validation sets, achieving similar performance to existing Transformer baselines. Among the Simi-CL settings, using the ViLT model without fine-tuning can bring improvements similar to Bootstrap-CL but lower than the CLIP model that has a larger size. While using the ViLT models fine-tuned with in-domain and non-target data, the model achieves similar performance to CLIP with fewer parameters, and using ViLT models fine-tuned with the target data can further outperform CLIP, which reveals the efficacy

Model	COCO				Flickr30k			
	B@4	M	C	S	B@4	M	C	S
Transformer Baseline w/ pretraining								
LEMON (Hu et al., 2022)	40.3	30.2	133.3	23.3	-	-	-	-
Transformer Baselines w/o pretraining								
VL-BART (Cho et al., 2021)	33.8	28.5	112.4	21.4	-	-	-	-
Unified VLP (Zhou et al., 2020a)	35.5	28.2	114.3	21.0	27.6	20.9	56.8	15.3
AoANet (Huang et al., 2019)	37.2	28.4	119.8	21.3	-	-	-	-
Our Implemented Baselines								
Transformer	35.7	27.9	113.0	20.9	27.7	21.8	58.5	16.0
Transformer + Addup-CL	35.2	27.9	114.2	21.0	26.5	21.5	56.6	16.0
Transformer + Bootstrap-CL	36.1	28.0	115.8	21.1	27.6	21.9	59.1	16.0
Our Proposed Methods								
Transformer + Simi-CL (ViLT)	35.9	28.0	115.6	21.2	27.3	21.9	59.0	16.0
Transformer + Simi-CL (CLIP)	36.3	28.1	116.2	21.2	27.0	**22.1**	59.6	16.2
Transformer + Simi-CL (ViLT-CC)	**36.4**	**28.2**	**117.1**	**21.4**	27.5	**22.1**	61.0	**16.3**
Transformer + Simi-CL (ViLT-FL)	36.0	28.0	115.9	21.0	**28.5**	**22.1**	**61.8**	16.2

Table 1: Overall performance of CL-based methods and existing state-of-the-art models on COCO and Flickr30k. B@4, M, C, and S represent BLEU-4, METEOR, CIDEr, and SPICE, respectively.

Model	Difficulty Level							
	Level-1		Level-2		Level-3		Level-4	
	B@4	C	B@4	C	B@4	C	B@4	C
Transformer	**71.0**	**199.5**	36.5	120.9	4.5	84.0	0.7	47.8
+ Bootstrap-CL	57.1	172.0	**37.8**	122.5	27.0	97.8	18.3	71.2
+ Simi-CL	57.0	172.5	37.1	**122.8**	**27.8**	**100.4**	**18.6**	**72.6**

Table 2: Model performance on the divided COCO test sets with different difficulty levels.

of teacher models possessing the target data knowledge for better curricula design. More details about the measured difficulty score distributions can be found in Appendix A. We also conduct the significance test for measuring the improvements which are described in Appendix B.

4.3 Quantitative Analysis

Performance on Divided Sets To understand how CL contributes to the vanilla captioning model, we equally divide the COCO test set into four subsets based on the BLEU scores of captions generated by the vanilla model for the test images. The results in Table 2 show that the vanilla model performance is unbalanced among data with different difficulty levels, while both CL methods improve the performance on the harder subsets, and Simi-CL achieves the best performance.

Model	B@4	M	C	S
Transformer	15.8	17.0	35.8	10.9
+ Bootstrap-CL	18.1	17.5	38.3	11.4
+ Simi-CL	**18.6**	**18.2**	**39.8**	**11.7**

Table 3: Cross-dataset performance evaluation using the best-performing COCO model for Flickr30k.

Model Generalization To evaluate the model generalization ability, we test the model with cross-dataset evaluation referencing the former work (Torralba and Efros, 2011). Specifically, we use the best-performing model trained with COCO to generate captions for the unseen test set from Flickr30k, obtaining the results listed in Table 3. The model performance maintains similar trends, with Simi-CL achieving the highest improvement and thus the best generalization ability.

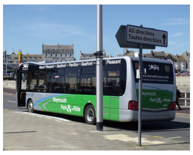
High-Score Examples from Hardest Subset		Low-Score Examples from Easiest Subset	
	Transformer : A girl playing a video game in a living room. (82.4) **Transformer + Bootstrap-CL :** A woman standing in a living room holding a remote. (127.2) **Transformer + Simi-CL :** A woman standing in a living room holding a wii remote. (**170.8**) **References:** 1. A woman standing next to a couch holding a wii controller. 2. Young woman playing wii in a furnished living room.		**Transformer :** A group of sleep laying on top of hay. (**206.1**) **Transformer + Bootstrap-CL :** A couple of sheep laying on top of dry grass. (185.5) **Transformer + Simi-CL :** A group of sheep laying in hay next to each other. (131.2) **References:** 1. A couple of sheep laying on top of a pile of dry grass. 2. Sheep laying and eating hay in an enclosure.
	Transformer : A stone building with a stone wall in front of it. (22.0) **Transformer + Bootstrap-CL :** A stone wall with a brick wall and a vase. (42.0) **Transformer + Simi-CL :** A bunch of vases sitting on a stone wall. (**112.3**) **References:** 1. A window on brick wall with vases in the sill. 2. Several colorful vases on a stone window ledge.		**Transformer :** A cat laying in a suitcase on the floor. (**293.5**) **Transformer + Bootstrap-CL :** A cat laying on top of a piece of luggage. (149.9) **Transformer + Simi-CL :** A cat sitting inside of a bag on the floor. (207.8) **References:** 1. A cat laying in a bag in a room. 2. A cat sitting in a suitcase on the floor.
	Transformer : A busy highway with a lot of traffic. (35.8) **Transformer + Bootstrap-CL :** A train travelling on a bridge over a street. (67.9) **Transformer + Simi-CL :** A train travelling down a highway next to traffic. (**126.1**) **References:** 1. A train travelling down tracks next to a highway. 2. A photo of a train heading down the tracks.		**Transformer :** A green and white bus driving down a street. (**252.6**) **Transformer + Bootstrap-CL :** A green and white bus parked at a bus stop. (126.5) **Transformer + Simi-CL :** A green and white bus parked next to a street sign. (162.4) **References:** 1. An empty bus travels down a city street. 2. A green and white bus is on the street.

Figure 3: Samples of generated captions on divided sets. Number in the parentheses indicates the CIDEr score.

Model	B@4	M	C	S
BUTD	35.2	27.2	109.9	20.1
+Simi-CL	**36.2**	**27.8**	**113.0**	**20.6**
AoANet	36.8	28.0	**117.2**	21.3
+Simi-CL	**37.3**	**28.2**	117.0	**21.4**

Table 4: Model performance on the COCO for applying Simi-CL to different model architectures.

Target Model Architecture To investigate the effect of CL on different model architectures, we applied it to the LSTM-based model BUTD and more advanced Transformer-based model AoANet. The results are shown in Table 4. The improvement achieved by CL is higher when applied to a simpler architecture, which reveals the small-size model can benefit more from the large pretrained model.

4.4 Qualitative Analysis

We compare the captions from the vanilla model and CL-based models on the aforementioned divided subsets to understand the differences of generated captions as shown in Figure 3. On the hardest subset, we observed captions from Simi-CL can recognize objects more accurately such as *wii remote* or *vases* and describe contents in detail, which reveals the improved generalization ability. While on the easiest subset, we found that even if both CL-based models generate captions with similar or higher quality, low scores are given since the matched n-grams are less based on the limited references, which indicates the model-based metrics should be considered for the reference-free evaluation, and we leave it to our future work.

5 Conclusion

In this paper, we propose an efficient cross-modal similarity-based difficulty measurement for image captioning. Our proposed Simi-CL method boosts the model performance and training speed especially for larger datasets, and the pretrained models fine-tuned with target data can lead to further improvement. The improvement for data with different difficulty levels and data from other dataset indicates that Simi-CL achieves the highest model generalization ability. We also apply Simi-CL to different model architectures, and the higher improvement for the simpler model shows its practicality when only small-size models can be implemented in real-world scenarios.

Limitations

The limitations of this paper are listed as follows:

Multiple Difficulty Measurements We mainly focus on the CL method with a single measurement, but ensemble multiple measurements for model training may improve the model performance further or disturb each other, which requires further investigations.

More Advanced Training Schedules There are other advanced continuous CL training schedules such as the competence-based learning (Platanios et al., 2019), which samples the data from easy to hard gradually. We think our study is a baseline for follow-up work, and we believe a better training schedule will further boost the model performance.

Challengeable Datasets There are several challengable image captioning datasets, such as the Novel Object Captioning (NoCaps) (Agrawal et al., 2019) and Conceptual Captions (CC) (Sharma et al., 2018). Since the model trained with the CL method can handle the harder data with better generalization ability, we believe its performance on these datasets will be improved.

Acknowledgements

We would like to thank the members of Aizawa Lab for their feedback on this work and the New Energy and Industrial Technology Development Organization (NEDO) for supporting this project.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019)*, pages 8948–8957.

Mohammad Alsharid, Rasheed El-Bouri, Harshita Sharma, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. 2021. A course-focused dual curriculum for image captioning. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI 2021)*, pages 716–720.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, pages 6077–6086.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML 2009)*, pages 41–48.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning (ICML 2021)*, pages 1931–1942.

Xinzhi Dong, Chengjiang Long, Wenju Xu, and Chunxia Xiao. 2021. Dual graph convolutional networks with transformer and curriculum learning for image captioning. In *Proceedings of the 29th ACM International Conference on Multimedia (ACM-MM 2021)*, pages 2615–2624.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, pages 7514–7528.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pages 17980–17989.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019)*, pages 4634–4643.

Yuyun Huang and Jinhua Du. 2019. Self-attention enhanced cnns and collaborative curriculum learning for distantly supervised relation extraction. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP 2019)*, pages 389–398.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 3128–3137.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning (ICML 2021)*, pages 5583–5594.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing (EMNLP 2004)*, pages 388–395.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. 2021. UMIC: An unreferenced metric for image captioning via contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, pages 220–226.

Fenglin Liu, Shen Ge, and Xian Wu. 2021. Competence-based multimodal curriculum learning for medical report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, pages 3001–3012.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 427–436.

Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, pages 6964–6974.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (NAACL-HLT 2019)*, pages 1162–1172.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML 2021)*, pages 8748–8763.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, page 9.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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 (ACL 2018)*, pages 2556–2565.

Valentin I Spitkovsky, Hiyan Alshawi, and Dan Jurafsky. 2010. From baby steps to leapfrog: How “less is more” in unsupervised dependency parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2010)*, pages 751–759.

Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, pages 1521–1528.

Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan. 2019. Dynamic curriculum learning for imbalanced data classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019)*, pages 5017–5026.

Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. 2016. STC: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI 2016)*, pages 2314–2320.

Daphna Weinshall, Gad Cohen, and Dan Amir. 2018. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *International Conference on Machine Learning (ICML 2019)*, pages 5238–5246.

Liuyu Xiang, Guiguang Ding, and Jungong Han. 2020. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision (ECCV 2020)*, pages 247–263.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 6095–6104.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020a. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2020)*, pages 13041–13049.

Yikai Zhou, Baosong Yang, Derek F Wong, Yu Wan, and Lidia S Chao. 2020b. Uncertainty-aware curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 6934–6944.

A Difficulty Score Distribution

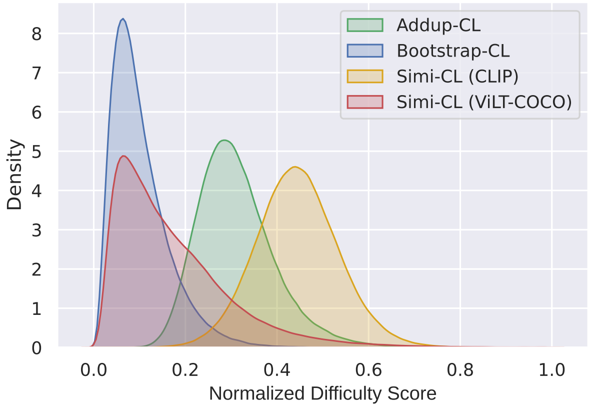


Figure 4: Difficulty score distribution for the COCO training data under different difficulty measurements.

We provide distributions of normalized difficulty scores for the COCO training data under different difficulty measurements as shown in Figure 4. We find distributions of scores measured by both VL models have higher dispersion, which indicates VL models can differentiate the difficulty levels of the training data better, thus can achieve higher model performance. However, although the distribution of Addup method has higher dispersion than Bootstrap, the latter one brings higher improvement. We suppose that it is because a good measurement requires not only strong differentiation ability, but also the rational data sorting order, and Addup method cannot provide the appropriate sorting order by simply adding up the mono-modal scores for the difficulty measurement.

B Significance Test

We adopt the Bootstrap Test (Koehn, 2004) that is widely used to compare two NMT systems' performance for our significance test. According to the sampling strategy in Bootstrap Test, we repeatedly sample images with replacements from the COCO test set to create 1000 sampled test sets (each contains 5000 images). Then we compute the metric scores for Bootstrap-CL and Simi-CL on all sampled sets. Finally we calculate the percentage of times that Simi-CL outperforms Bootstrap-CL to obtain the statistical significance as: Simi-CL is superior on BLEU-4, CIDEr, METEOR, and SPICE with p-value 0.032, 0.005, 0.001, and 0.001 respectively. If we set the significance level to 0.05, the improvements in all the metrics indicate the statistically significant improvement of our method.