

The Supplementary of Refining Visual Perception for Decoration Display: A Self-Enhanced Deep Captioning Model

1. Experiments

In supplementary materials, we give more analyses and further consider the experimental performance in different scenarios.

1.1. Complexity Analysis

In detail, the common components of CPRC are the multi-class classifier and the generator. For the multi-class classifier, cross entropy and KL divergence optimization are involved, which has overall time complexity $o(|v|n)$, where $|v|$ is the size of the class set, and n denotes the number of instances. The time complexity of the generator is $O\left(\sum_{i=0}^4 M_i d_i + r^2 D + T D^2\right)$, where M_i represents the input dimension of the full connection layer, d_i indicates the output dimension of the full connection layer, i denotes the index of layers, r represents the number of regions using Faster R-CNN, D denotes the dimension of regions, and T represents the time step of the recurrent neural network.

1.2. Influence of Unsupervised Data

Furthermore, we explore the influence of unsupervised data, i.e., we fix the supervised ratio to 1%, and tune the data ratio from unsupervised data in $\{10\%, 40\%, 70\%, 100\%\}$, the results are recorded in Table 1. We find that, as the percentage of unsupervised data increases, the performance of CPRC also improves in terms of all metrics. This indicates that CPRC can make full use of undescribed images for positive training. But the growth rate slows down with the ratio going up (i.e., after 70%), probably owing to the interference of pseudo-label noise.

1.3. Experiments on FLICKR30K Dataset

We add more experiments on FLICKR30K dataset [Young et al. \(2014\)](#). Unsupervised captioning methods Graph-align and UIC have not provided the source codes or performed experiments on the FLICKR30K dataset, so the results of FLICKR30K of these two methods cannot be provided. From the results in Table 2, we can obtain conclusions similar to the COCO dataset, thus verifying the effectiveness of CPRC in different datasets.

Table 1: Performance with different ratio data from unsupervised data (i.e., the supervised is fixed with 1%) on MS-COCO “Karpathy” test split, where B@N, M, R, C and S are short for BLEU@N, METEOR, ROUGE-L, CIDEr-D and SPICE scores.

| Methods | Cross Entropy Loss | | | | | | | |
|---------|----------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | B@1 | B@2 | B@3 | B@4 | M | R | C | S |
| 10% | 68.3 | 49.5 | 34.9 | 23.3 | 21.4 | 49.6 | 71.7 | 14.6 |
| 40% | 66.9 | 48.7 | 34.2 | 23.4 | 22.9 | 49.6 | 72.9 | 15.6 |
| 70% | 68.4 | 50.6 | 35.6 | 24.4 | 22.9 | 50.5 | 74.4 | 15.9 |
| 100% | 68.8 | 51.1 | 35.7 | 24.9 | 22.9 | 50.4 | 77.9 | 16.2 |
| Methods | CIDEr-D Score Optimization | | | | | | | |
| | B@1 | B@2 | B@3 | B@4 | M | R | C | S |
| 10% | 68.7 | 51.0 | 25.6 | 23.9 | 22.4 | 50.6 | 74.1 | 14.9 |
| 40% | 69.2 | 50.2 | 35.6 | 24.1 | 22.9 | 50.8 | 75.7 | 15.9 |
| 70% | 69.4 | 51.3 | 36.5 | 24.8 | 22.8 | 50.7 | 76.5 | 16.2 |
| 100% | 69.9 | 51.8 | 36.7 | 25.5 | 23.4 | 50.7 | 78.8 | 16.8 |

Table 2: Performance of comparison methods on FLICKR30K dataset, where B@N, M, R, C and S are short for BLEU@N, METEOR, ROUGE-L, CIDEr-D and SPICE scores.

| Methods | Cross Entropy Loss | | | | | | | | CIDEr-D Score Optimization | | | | | | | |
|----------------|--------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|----------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | B@1 | B@2 | B@3 | B@4 | M | R | C | S | B@1 | B@2 | B@3 | B@4 | M | R | C | S |
| SCST | 35.5 | 21.0 | 12.5 | 7.7 | 11.3 | 31.7 | 7.1 | 7.1 | 38.2 | 22.9 | 13.8 | 8.6 | 11.7 | 32.8 | 8.3 | 7.4 |
| AoANet | 55.2 | 35.8 | 22.7 | 14.2 | 15.7 | 39.4 | 24.5 | 10.1 | 58.9 | 38.5 | 24.3 | 15.1 | 15.0 | 39.9 | 23.9 | 9.2 |
| AAT | 53.9 | 34.6 | 21.0 | 13.0 | 15.0 | 38.6 | 19.5 | 9.3 | 52.5 | 33.1 | 19.7 | 11.8 | 14.0 | 35.4 | 18.5 | 8.9 |
| ORT | 54.3 | 34.9 | 21.5 | 13.5 | 15.2 | 38.9 | 23.1 | 9.4 | 56.9 | 37.3 | 22.5 | 14.2 | 14.8 | 38.6 | 22.4 | 9.1 |
| GIC | 34.7 | 20.5 | 12.0 | 7.3 | 10.8 | 30.5 | 7.0 | 6.8 | 37.6 | 22.1 | 13.6 | 8.4 | 11.4 | 31.6 | 8.3 | 7.5 |
| Anchor | 35.2 | 20.8 | 12.1 | 7.5 | 11.0 | 30.8 | 7.1 | 6.8 | 38.0 | 22.6 | 13.6 | 8.4 | 11.2 | 32.6 | 8.1 | 7.3 |
| RSTNet | 55.6 | 35.8 | 22.9 | 14.6 | 15.8 | 39.7 | 24.8 | 10.2 | 55.4 | 35.3 | 22.5 | 14.5 | 15.6 | 39.5 | 24.2 | 9.5 |
| Graph-align | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| UIC | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| A3VSE | 56.6 | 37.1 | 23.7 | 15.0 | 15.7 | 39.7 | 25.4 | 10.1 | 56.3 | 36.8 | 23.5 | 14.8 | 15.9 | 39.6 | 25.4 | 10.2 |
| AoANet+P | 55.7 | 36.4 | 22.6 | 14.3 | 16.0 | 39.5 | 25.2 | 10.2 | 59.3 | 39.0 | 24.7 | 15.4 | 15.2 | 40.3 | 24.5 | 9.6 |
| AoANet+C | 55.5 | 36.0 | 22.5 | 14.2 | 15.8 | 39.5 | 24.9 | 10.2 | 59.0 | 38.6 | 24.4 | 15.1 | 15.0 | 40.3 | 24.1 | 9.5 |
| PL | 56.1 | 36.5 | 23.1 | 14.4 | 16.2 | 39.5 | 25.5 | 10.2 | 59.4 | 39.0 | 24.7 | 15.5 | 15.4 | 40.4 | 24.6 | 9.7 |
| AC | 54.2 | 35.1 | 22.1 | 12.4 | 15.0 | 38.5 | 23.2 | 9.4 | 57.0 | 37.5 | 22.9 | 14.5 | 14.5 | 38.4 | 22.5 | 9.1 |
| Embedding+ | 53.6 | 34.7 | 22.0 | 13.3 | 14.5 | 39.0 | 23.1 | 9.1 | 55.7 | 37.2 | 23.5 | 14.6 | 14.6 | 39.3 | 23.4 | 9.3 |
| Semantic+ | 55.9 | 37.0 | 23.6 | 14.1 | 15.8 | 39.9 | 25.0 | 10.2 | 59.4 | 39.0 | 25.1 | 15.4 | 15.9 | 40.7 | 25.4 | 10.3 |
| Strong+ | 57.1 | 37.6 | 24.2 | 15.2 | 16.0 | 40.3 | 26.3 | 10.4 | 59.2 | 38.5 | 25.7 | 15.4 | 16.4 | 41.0 | 27.4 | 10.5 |
| w/o Prediction | 57.0 | 37.4 | 22.0 | 15.1 | 15.7 | 40.1 | 25.5 | 10.3 | 59.0 | 38.2 | 25.3 | 15.2 | 16.1 | 40.5 | 25.9 | 10.3 |
| w/o Relation | 57.2 | 37.6 | 22.9 | 15.3 | 15.9 | 40.2 | 26.0 | 10.3 | 59.2 | 38.7 | 25.8 | 15.5 | 16.5 | 40.7 | 26.3 | 10.4 |
| w/o τ | 56.8 | 36.7 | 22.4 | 15.1 | 15.4 | 40.0 | 25.6 | 10.3 | 58.7 | 38.2 | 25.0 | 15.1 | 16.3 | 40.3 | 26.0 | 10.4 |
| CPRC | 57.6 | 37.9 | 24.5 | 15.6 | 16.3 | 40.4 | 26.9 | 10.6 | 59.8 | 39.2 | 26.1 | 15.9 | 16.7 | 41.2 | 27.6 | 10.8 |

1.4. Supervised and Unsupervised Image Captioning

We also evaluate our proposed method under the supervised and unsupervised scenarios. In detail, we compared two types of methods: 1) State-of-the-art supervised captioning approaches: GIC [Zhou et al. \(2020\)](#), Anchor [Xu et al. \(2021\)](#) and RSTNet [Zhang et al. \(2021\)](#). 2) State-of-the-art unsupervised captioning methods: Graph-align [Gu et al. \(2019\)](#) and UIC [Feng et al. \(2019\)](#). Considering the performance improvements of different methods in supervised scenarios, we do not compare other traditional supervised comparison methods introduced in the main text. Meanwhile, in the unsupervised scenario, we train the CPRC in two ways: 1) CPRC (Pre-train) fine-tunes the generator G with only label prediction module, by using the pre-trained model from FLICKR30K dataset [Young et al. \(2014\)](#). 2) CPRC trains the generator G from scratch with only a label prediction module. Table 3 and Table 4 record the results of supervised and unsupervised settings. The results indicate that: 1) CPRC performs better than the state-of-the-art supervised captioning approaches with only AoANet structure for the generator G (note that AoANet performs worse than other methods under full supervision), which verifies that the task of multi-label prediction can facilitate the task of text generation; 2) To explore the generality of CPRC, we conduct more experiments by incorporating CPRC with the supervised captioning approaches, i.e., GIC, Anchor, and RSTNet, for the supervised image captioning. We find that GIC+CPRC, Anchor+CPRC, and RSTNet+CPRC have further improved performance, which validates that CPRC can well combine the label prediction module for existing supervised captioning models; 3) CPRC suffers performance degradation under the unsupervised scenario, for the reason that Graph-align and UIC additionally use pre-trained models obtained from large-scale data to calculate the scene graphs or constrain the sentence generation. On the other hand, CPRC (Pre-train) improves the performance on all criteria, which shows that CPRC can effectively transfer the pre-trained generator.

Table 3: Performance of comparison methods on supervised setting, where B@N, M, R, C and S are short for BLEU@N, METEOR, ROUGE-L, CIDEr-D and SPICE scores.

| Methods | Cross Entropy Loss | | | | | | | | CIDEr-D Score Optimization | | | | | | | |
|-------------|--------------------|-------------|-------------|-------------|-------------|-------------|--------------|-------------|----------------------------|-------------|-------------|-------------|-------------|-------------|--------------|-------------|
| | B@1 | B@2 | B@3 | B@4 | M | R | C | S | B@1 | B@2 | B@3 | B@4 | M | R | C | S |
| GIC | 75.6 | 63.1 | 48.5 | 36.3 | 27.9 | 53.9 | 114.2 | 20.4 | 80.0 | 63.8 | 48.9 | 37.5 | 28.3 | 55.7 | 125.4 | 22.0 |
| Anchor | 72.5 | 61.8 | 47.4 | 35.6 | 26.7 | 52.4 | 105.7 | 19.6 | 74.9 | 63.2 | 48.8 | 34.4 | 27.0 | 56.0 | 110.1 | 20.2 |
| RSTNet | 78.0 | 65.2 | 51.0 | 36.5 | 28.3 | 57.4 | 119.8 | 21.4 | 81.1 | 65.8 | 51.3 | 39.1 | 29.2 | 58.8 | 132.7 | 22.8 |
| GIC+CPRC | 76.4 | 63.6 | 49.1 | 36.8 | 28.4 | 54.3 | 116.5 | 20.9 | 80.8 | 64.7 | 49.5 | 38.5 | 29.0 | 55.9 | 126.2 | 22.4 |
| Anchor+CPRC | 73.8 | 62.7 | 48.3 | 36.1 | 27.2 | 52.9 | 107.0 | 20.1 | 75.6 | 64.2 | 49.5 | 35.1 | 27.7 | 56.5 | 126.3 | 20.6 |
| RSTNet+CPRC | 78.1 | 65.9 | 51.3 | 37.6 | 28.7 | 57.6 | 120.1 | 21.6 | 81.6 | 66.3 | 51.7 | 39.4 | 29.5 | 59.2 | 133.7 | 23.2 |
| CPRC | 77.9 | 65.7 | 51.2 | 37.4 | 28.6 | 57.6 | 120.0 | 21.5 | 80.8 | 65.6 | 51.0 | 39.2 | 29.3 | 59.1 | 129.4 | 22.9 |

1.5. Computation Costs

We record the time of supervised and semi-supervised comparison methods considering the availability of source code. We are unable to get the running time of unsupervised methods because there is no source code. The experimental results in Table 5 reveal that: 1) CPRC costs a longer training time than supervised comparison methods. For the reason

Table 4: Performance of comparison methods on unsupervised setting, where B@N, M, R, C and S are short for BLEU@N, METEOR, ROUGE-L, CIDEr-D and SPICE scores.

| Methods | Cross Entropy Loss | | | | | | | | CIDEr-D Score Optimization | | | | | | | |
|------------------|--------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|----------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | B@1 | B@2 | B@3 | B@4 | M | R | C | S | B@1 | B@2 | B@3 | B@4 | M | R | C | S |
| Graph-align | - | - | - | - | - | - | - | - | 67.1 | 47.8 | 32.3 | 21.5 | 20.9 | 47.2 | 69.5 | 15.0 |
| UIC | - | - | - | - | - | - | - | - | 41.0 | 22.5 | 11.2 | 5.6 | 12.4 | 28.7 | 28.6 | 8.1 |
| CPRC (Pre-train) | 65.4 | 46.7 | 31.5 | 21.2 | 20.5 | 47.0 | 68.6 | 15.2 | 68.8 | 48.6 | 32.6 | 22.1 | 21.6 | 47.8 | 71.2 | 15.6 |
| CPRC | 37.3 | 18.1 | 6.8 | 2.6 | 10.0 | 28.9 | 16.1 | 6.6 | 37.9 | 18.4 | 7.0 | 3.5 | 10.2 | 29.4 | 17.2 | 7.3 |

Table 5: Computation time of comparison methods. The unit is the hour.

| Methods | AoANet | AAT | ORT | SCST | GIC | Anchor | RSTNet | A3VSE | CPRC |
|---------|--------|------|------|------|------|--------|--------|-------|-------|
| Times | 1.23 | 1.50 | 1.75 | 0.87 | 2.00 | 1.38 | 1.17 | 32.80 | 30.35 |

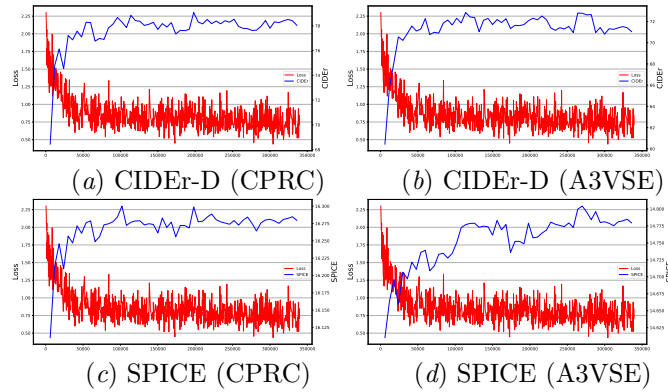


Figure 1: The illustration of convergence vs performance (i.e., CIDEr-D and SPICE) of CPRC and semi-supervised method A3VSE.

that supervised methods cannot use large amounts of undescribed images, and only train with described images, so requires shorter training time. 2) CPRC trains faster than the semi-supervised method, i.e., A3VSE, under the premise of data augmentation. The phenomenon indicates that CPRC converges fast. Figure 1 further exhibits the convergence vs performance (i.e., CIDEr-D and SPICE) of CPRC and A3VSE. The left vertical axis represents the loss function value, and the right vertical axis is the performance of the indicator. The horizontal axis represents the number of iterations. We find that A3VSE converges more slowly and performance is unstable.

References

Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *CVPR*, pages 4125–4134, Long Beach, CA, 2019.

- 70 Jiuxiang Gu, Shafiq R. Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Un-
71 paired image captioning via scene graph alignments. In *ICCV*, pages 10322–10331, Seoul,
72 Korea, 2019.
- 73 Guanghui Xu, Shuaicheng Niu, Mingkui Tan, Yucheng Luo, Qing Du, and Qi Wu. Towards
74 accurate text-based image captioning with content diversity exploration. In *CVPR*, pages
75 12637–12646, virtual, 2021.
- 76 Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to
77 visual denotations: New similarity metrics for semantic inference over event descriptions.
78 *Trans. Assoc. Comput. Linguistics*, 2:67–78, 2014.
- 79 Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue
80 Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and
81 non-visual words. In *CVPR*, pages 15465–15474, virtual, 2021.
- 82 Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded
83 image captioning by distilling image-text matching model. In *CVPR*, pages 4776–4785,
84 2020.