Skeleton Key: Image Captioning by Skeleton-Attribute Decomposition (Supplementary Material)

Yufei Wang[1] Zhe Lin[2] Xiaohui Shen[2] Scott Cohen[2] Garrison W. Cottrell[1]
[1]University of California, San Diego [2]Adobe Research
{yuw176, gary}@ucsd.edu {zlin, xshen, scohen}@adobe.com

1. Qualitative Results and Analysis

1.1. Comparison of the baseline and our method

In Figure 1 and Figure 2, we show some randomly sampled images with qualitative results for our proposed method and the baseline method on the MS-COCO and Stock3M datasets respectively (images with very similar captions generated by the baseline and proposed model are eliminated here in order to show the difference between the two models). Captions in green text boxes are generated by the proposed coarse-to-fine method, and captions in red text boxes are generated by the baseline method. We can see that for most images, our proposed method outperforms the baseline method, which is mainly reflected in two different aspects: (1) more accurate number/color attributes (e.g. Figure 1 1st column 2nd, 5th and 6th image; Figure 1 2nd column 1st image); (2) more accurate skeleton captions with better objects (e.g. Figure 1 2st column 3rd image; Figure 2 1st column 1st, 3rd and 6th image; Figure 2 2nd column 4th image and 6th image).

We also explicitly choose 5 failure examples in which coarse-to-fine method performs no better than baseline method in Figure 1 and Figure 2. The examples are shown in the third column, on the right of Figure 1 and Figure 2. We can see that incorrect recognition of objects or missing main objects in the image is still the dominant cause of error.

1.2. The ability to generate variable length captions

In Figure 3 and Figure 4, we demonstrate the ability of our proposed method to generate variable length captions, by showing the four captions generated by coarse-to-fine method (green box in the middle) and baseline method (red box in the right) for a test image. For the captions generated by coarse-to-fine method, four captions are generated with four pairs of (skeleton, attribute) length factor values: $(-1,-1),(-1,1.5),(1.5,-1),(1.5,1.5)$. The four value pairs represent all combinations of encouraging less/more information in skeleton/attributes. Attributes are marked in red in the generated caption. As a comparison, we also show four captions generated by the baseline method for the same images, using four different length factor values: $\gamma \in \{-1,-0.5,0.5,1.5\}$.

We can see that the captions generated by our coarse-to-fine model are much more flexible and useful than the ones generated by the baseline method. And we can also observe the effect of separate control of attribute/skeleton in our coarse-to-fine model.

1.3. Analysis of evaluation metrics

In Figure 5, we demonstrate that different evaluation metrics can yield opposite judgement results for a pair of generated captions. In the main paper, we emphasize our results on SPICE metrics rather than the conventional metrics. Here, in order to validate our claim, we show the comparison between SPICE and METEOR metrics. Similar observations also hold for comparisons of SPICE to other conventional metrics. Similar observations also hold for comparisons of SPICE to other conventional metrics. For each image in Figure 5, there are three text boxes below it: the caption in green text box is generated by our proposed coarse-to-fine model; the caption in red box is generated by baseline model; the caption(s) in black text box is(are) ground-truth caption(s). For Stock3M, there is one ground-truth caption per image; for MS-COCO, there are 5 ground-truth captions per image. For the predicted captions by two models, we also show in parentheses the SPICE score (S) and METEOR score (M) for each of the predictions. We can see that for the first two rows of eight images, our coarse-to-fine model produces qualitatively better results with higher SPICE scores than the baseline but METEOR shows that our method is worse than the baseline.

For example, the first image in the first row is about the tea leaves, and the main object "tea" occurs in the coarse-to-fine model prediction, but not in the baseline prediction. The baseline prediction method incorrectly recognized the main object in the image as coffee beans. However, the baseline prediction has a higher METEOR score, because

Figure 1: Qualitative comparison of our proposed coarse-to-fine algorithm (green text box to the right of each image) and baseline algorithm (red text box) on random samples from the MS-COCO test set. On the left, we can see the coarse-to-fine method outperforms baseline method in most cases. On the right, we explicitly choose to show some examples on which either methods generates captions that have clear flaws.

of the mention of "pile of". However, the correct mention of objects is obviously more important in human judgement.

1.4. More examples of attention refinement process

In Figure 6, we show more examples of the attention mechanism during skeleton sentence generation of our proposed coarse-to-fine model. For each word predicted by Skel-LSTM, the attention map, predicted words for each location, and the refined attention map are shown. When the word is an object, we can see how the refined attention map produces more accurate attention focusing on objects of interest (e.g. First image last word "bottle"). Since attributes are object-specific, improvement in the quality of attention maps for each object appearing in the skeleton is critical for the accuracy of attributes as well as the whole caption.

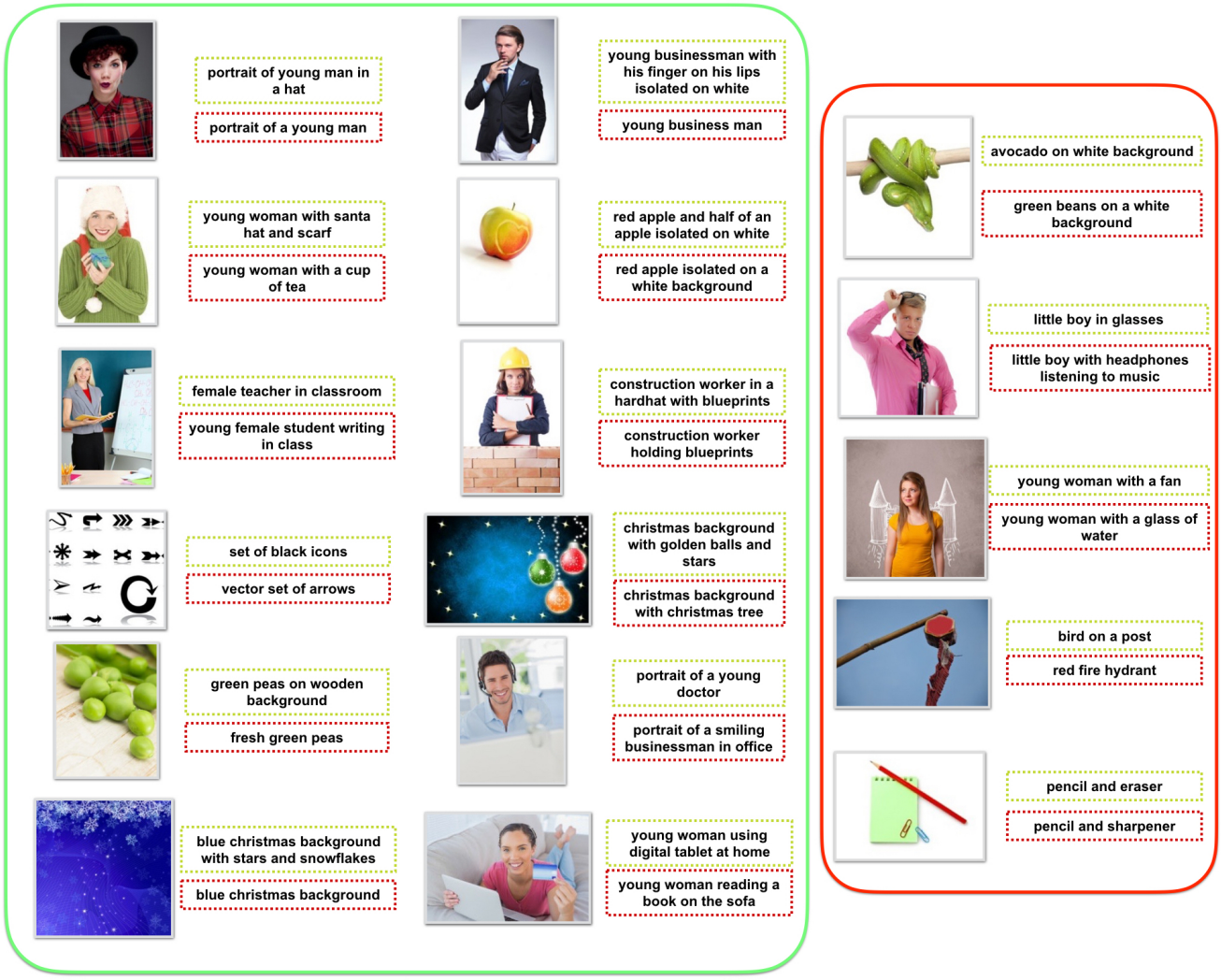

Figure 2: Qualitative comparison of our proposed coarse-to-fine algorithm (green text box to the right of each image) and baseline algorithm (red text box) on random samples from the Stock3M test set. On the left, we can see the coarse-to-fine method outperforms baseline method in most cases. On the right, we explicitly choose to show some examples on which either methods generates captions that have clear flaws.

Figure 3: More examples of predicted titles for image examples from MS-COCO. Four titles are generated from our coarse-to-fine model (middle, in green box) and baseline model (right, in red box) respectively. For coarse-to-fine model, four pairs of length factor value γ for skeletal title and attributes are (-1, -1), (-1, 1.5), (1.5, -1), (1.5, 1.5) respectively. For the baseline method, the γ's are -1, -0.5, 0.5, 1.5 respectively. The captions generated by our coarse-to-fine model are much more flexible and useful than the ones generated by the baseline method.

Figure 4: More examples of predicted titles for image examples from Stock3M. Four titles are generated from our coarse-to-fine model (middle, in green box) and baseline model (right, in red box) respectively. For coarse-to-fine model, four pairs of length factor value γ for skeletal title and attributes are (-1, -1), (-1, 1.5), (1.5, -1), (1.5, 1.5) respectively. For the baseline method, the γ's are -1, -0.5, 0.5, 1.5 respectively. The captions generated by our coarse-to-fine model are much more flexible and useful than the ones generated by the baseline method.

Figure 5: Examples of some generated captions that get high score on SPICE but get low score on METEOR. In the first row of four images from Stock3M and second row of four images from MS-COCO, the captions generated from coarse-to-fine method (green text box) and baseline method (red text box) are shown. Moreover, the SPICE score (S) and METEOR score (M) are also shown for both generated captions. In the black text box, the ground-truth captions given by human annotators are also provided. We can see that for the first two rows, although captions generated from baseline method have higher METEOR score, they have lower SPICE score than coarse-to-fine result, and the SPICE score is closer to human judgement.

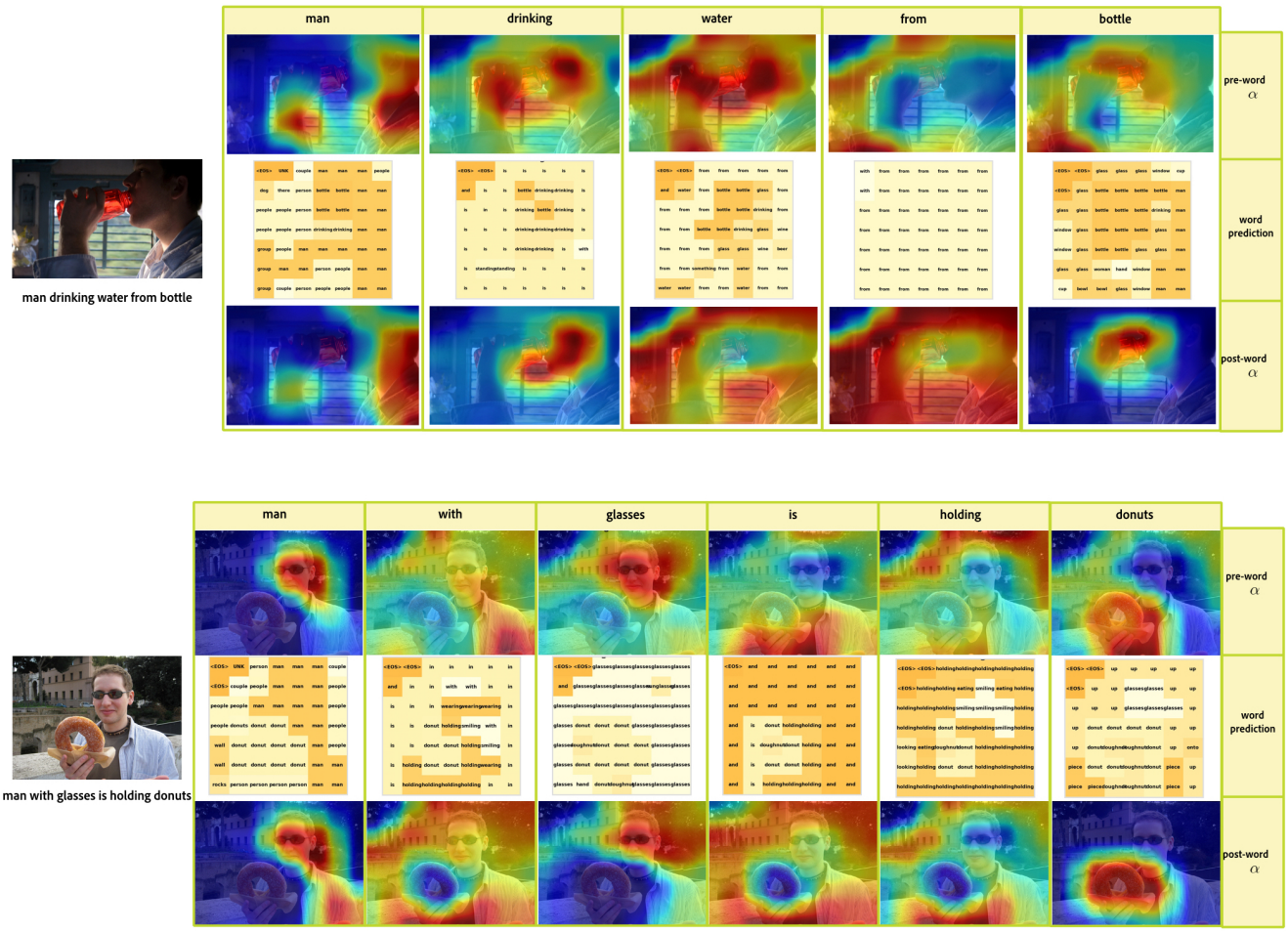

Figure 6: Illustration of attention refinement process during inference stage. All the skeleton words in generated skeleton sentence are shown. For each word, the attention map, predicted words for each location, and refined attention map are shown.

Table 1: Choice of beam size and length factor γ for both baseline model and our proposed coarse-to-fine model. The values are decided on validation set.

	Model	Beam size	length factor γ	Attribute beam size	attribute length factor γ
MS-COCO	Baseline	3	0.1	-	-
	Coarse-to-fine	5	0.1	2	0.9
Stock3M	Baseline	2	1.0	-	-
	Coarse-to-fine	2	1.2	2	0.3

Table 2: Performance of our method on online MS-COCO testing server (https://competitions.codalab.org/competitions/3221). We also show the results of other published state-of-the-art results.

Models	**BLEU-1**		**BLEU-2**		**BLEU-3**		**BLEU-4**		**METEOR**		**ROUGE-L**		**CIDEr**	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Ours	**0.734**	**0.912**	0.564	**0.829**	**0.425**	**0.724**	**0.320**	**0.612**	**0.262**	**0.356**	**0.542**	**0.698**	**1.011**	**1.026**
ATT_VC [4]	0.731	0.900	**0.565**	0.815	0.424	0.709	0.316	0.599	0.250	0.335	0.535	0.682	0.943	0.958
OriolVinyals [2]	0.713	0.895	0.542	0.802	0.407	0.694	0.309	0.587	0.254	0.346	0.530	0.682	0.943	0.946

Table 3: Performance of baseline model, our model and previous state-of-the-art models on Stock3M.

Models	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**ROUGE-L**	**CIDEr**
NIC[1]	0.141	0.074	0.041	0.023	0.069	0.165	0.447
Toronto[3]	0.165	0.091	0.054	0.032	0.083	0.195	0.569
Baseline	0.236	0.133	0.079	0.050	0.108	0.233	0.720
Ours	**0.245**	**0.138**	**0.083**	**0.052**	**0.110**	**0.239**	**0.724**

Table 4: Performance of baseline model, our model and previous state-of-the-art models for MS-COCO on SPICE measurement.

Models	**NIC[1]**	**Toronto[3]**	**Baseline**	**Ours**
SPICE	0.157	0.174	0.188	**0.196**

2. Choice of Hyper-parameters

In Table 1, we list the hyper-parameters of the models we use for all the results we show in the main paper and here. The beam size and length factor γ are chosen on a validation set for both Stock3M and MS-COCO.

3. Performance on MS-COCO Testing Server

In Table 2, we show our submission to the MS-COCO online testing server. The server evaluates models with 40,775 test images with ground-truth captions that competitors do not have access to. We also show the other published state-of-the-art results in Table 2. Note that we do not use any commonly used augmentation tricks such as model ensembling.

4. Performance of Previous State-of-the-art Models on Stock3M

Apart from the comparison between baseline model and our model on Stock3M in the main paper, we re-implement two previous state-of-the-art models using the same parameters in the papers. In Table 3, we provide the comparison of the results. We can see that our baseline method is very strong.

5. Performance of Two Previous State-of-the-art Models for MS-COCO on SPICE Measurement

We provide comparison on SPICE measurement between our model and two previous state-of-the-art models (our reimplementation) in Table 4. Our reimplementation achieves better results than its original implementation in [1] and [3] (for BLEU-4: the difference between our reimplementation and original model for [1] is 0.284 v.s. 0.277; for [3] 0.312 v.s. 0.250). In Table 4, we can see that our baseline and our model outperform the two previous state-of-the-art models by a large margin.

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014. 8, 9

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *CoRR*, abs/1609.06647, 2016. 8

[3] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In D. Blei and F. Bach, editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2048–2057. JMLR Workshop and Conference Proceedings, 2015. 8, 9

[4] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. *CoRR*, abs/1603.03925, 2016. 8