Show, Adapt and Tell: Adversarial Training of Cross-domain Image Captioner
Supplementary Material

Tseng-Hung Chen†, Yuan-Hong Liao†, Ching-Yao Chuang†, Wan-Ting Hsu†, Jianlong Fu‡, Min Sun†

†Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan
‡Microsoft Research, Beijing, China

{tsenghung@gapp, andrewliao11@gapp, cychuang@gapp, hsuwanting@gapp, sunmin@ee}.nthu.edu.tw
jianf@microsoft.com

1. Dataset

1.1. Statistics

In Section 4.1 of the main paper, we mention that we use the same vocabulary in all experiments. Therefore, after removing the training sentences that contain out-of-vocabulary words, the numbers of captions for four target domain datasets (i.e., CUB-200, Oxford-102, TGIF [1], Flickr30k) are shown in Table 1.

Table 1: Dataset statistics.

Dataset	Training captions	Training images
CUB-200	25,926	4,000
Oxford-102	22,716	5,823
TGIF	65,526	79,984
Flickr30k	117,664	29,000

1.2. Sentence-level Distribution

In order to compare the difference across datasets in sentence-level, we encode sentences using Skip-Thought Vectors [4] and use Barnes-Hut-SNE [5] to visualize the embeddings given a fixed number of sentences. For MSCOCO, Oxford-102 and CUB-200, sentence representations are similar in a single dataset but different across datasets (see Fig. 1). On the other hand, sentence representations in MSCOCO, Flickr30k and TGIF are more similar and have subtle difference (see Fig. 2).

2. Baseline: Sentence Augmentation (SA)

We re-implement the sentence augmentation method described in [7]. In order to train the captioner with target domain unpaired captions, we follow [7] to replace the image feature with the same all-zero vector. Then we train our captioner on both the source domain image-caption pairs and target domain sentence-only examples. The performance is shown in Table 2. Our method outperforms SA consistently for all four target domain datasets. Note that SA performs better than DCC [2]. We find that DCC emphasizes on image-caption semantic relation, whereas SA forces the sentence style resemble to target domain and does not consider if the semantic meaning is grounded in the image. On the contrary, our method considers both factors important.

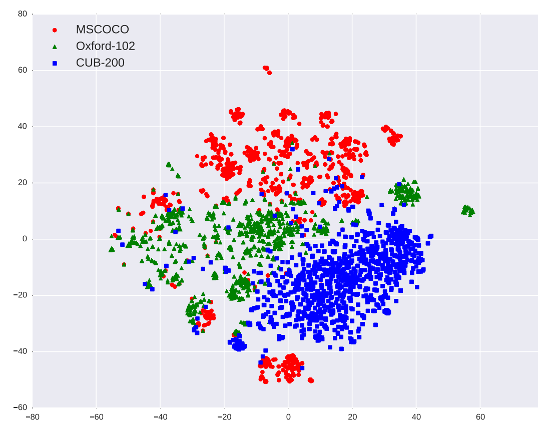

Figure 1: Barnes-Hut-SNE embeddings of skip-thought vectors on MSCOCO (red), Oxford-102 (green) and CUB-200 (blue).

We also provide the in-domain performances when training directly on paired image-caption data from the target domains in comparison with fine-tuning the source (MSCOCO) pre-trained model. We found the in-domain performances lower than fine-tuning (see Table 2) mainly because MSCOCO is a much larger dataset with diverse im-

[1] In TGIF, the visual contents are animated GIFs. To make it compatible with our captioner, we sample the first frame of each animated GIF as input image.

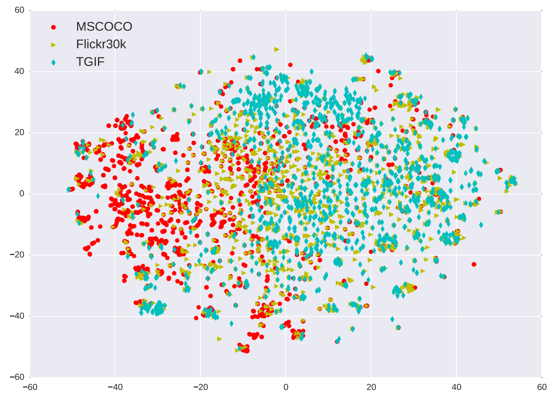

Figure 2: Barnes-Hut-SNE embeddings of skip-thought vectors on MSCOCO (red), Flickr30k (yellow) and TGIF (cyan).

Table 2: Results of our method compared with SA and DCC. DCC and SA are baseline methods. Target denotes the model directly trained on all paired image-caption data from the target domains. Fine-tuning with paired data in target domain serves as the upper bound performance of our CNN-RNN captioner.

	Target (test)	Bleu-4	Meteor	ROUGE	CIDEr
DCC	CUB-200	21.4	23.8	46.4	11.9
SA	CUB-200	30.7	22.6	51.1	21.7
Ours	CUB-200	**32.8**	**27.6**	**58.6**	**24.8**
Target	CUB-200	42.1	30.3	59.6	18.8
Fine-tuning	CUB-200	59	36.1	69.7	61.1
DCC	Oxford-102	16.7	21.5	38.3	6
SA	Oxford-102	20.3	19.4	40.7	15.2
Ours	Oxford-102	**60.5**	**36.4**	**72.1**	**29.3**
Target	Oxford-102	61.6	37	74.4	29.3
Fine-tuning	Oxford-102	66.3	40	75.6	36.3
DCC	TGIF	4.1	11.8	29.5	7.1
SA	TGIF	8.2	13.6	34.9	18
Ours	TGIF	**10.3**	**14.5**	**37**	**22.2**
Target	TGIF	8.1	14.2	34.2	23.2
Fine-tuning	TGIF	11.8	16.2	39.2	29.8
DCC	Flickr30k	13.8	16.1	38.8	27.7
SA	Flickr30k	14.7	15.7	39.6	27.2
Ours	Flickr30k	**17.9**	**16.7**	**42.1**	**32.6**
Target	Flickr30k	16.9	17.7	41.7	33.2
Fine-tuning	Flickr30k	18.3	18	42.9	35.9

ages and sentences.

3. Design Choices for Critics

For domain critic, we choose CNN as sentence encoding since [3] has shown its success on the task of text classification and SeqGAN [6] adopt CNN as their discriminative model. For multi-modal critic, we are inspired by the structure from VQA [1]. We also have explored a combination of design choices for sentence encoding. Performance comparison is shown in Table 3. We find that the design choice reported in the paper achieves the best performance in TGIF and Flickr30k and is comparable with "Both LSTM" (using LSTM as sentence encoding in both DC and MC) in CUB-200 and Oxford-102.

4. Additional Qualitative Results

We show more qualitative results in Fig. 3. For each example, we show two captions: before adaptation (source pre-trained) and after adaptation (using our method).

For critic-based planning, we show qualitative results on TGIF and Flickr30k in Fig. 4. For each example, we list three captions from top to bottom in the order of greedy search, beam search (with beam size 2) and our proposed critic-based planning.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 2

[2] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, S. Kate, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016. 1

[3] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush. Character-aware neural language models. *AAAI*, 2016. 2

[4] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015. 1

[5] L. Van Der Maaten. Barnes-hut-sne. *arXiv preprint arXiv:1301.3342*, 2013. 1

[6] L. Yu, W. Zhang, J. Wang, and Y. Yu. Seqgan: sequence generative adversarial nets with policy gradient. *AAAI*, 2017. 2

[7] K.-H. Zeng, T.-H. Chen, J. C. Niebles, and M. Sun. Title generation for user generated videos. In *ECCV*, 2016. 1

Table 3: Results of different design choices for sentence encoding in Domain Critic (DC) and Multi-modal Critic (MC). Note that "Both CNN" stands for using CNN as sentence encoding in both DC and MC.

Sentence Encoding	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor	ROUGE	CIDEr	SPICE
MSCOCO $\rightarrow$ CUB-200								
Both CNN	79.1	64.9	47.7	33.4	24.1	52.3	22.1	11
Both LSTM	87.6	**73.7**	**56.9**	**41.6**	26	57.1	**27.6**	**14.1**
DC: LSTM, MC: CNN	78.1	64.9	48.3	34.6	24	52.9	23.3	11.9
DC: CNN, MC: LSTM	**91.4**	73.1	51.9	32.8	**27.6**	**58.6**	24.8	13.2
MSCOCO $\rightarrow$ Oxford-102								
Both CNN	56.7	34	17.4	7.9	14.4	32.9	7.3	6.3
Both LSTM	**86.3**	**77.4**	**67.6**	**60.6**	**36.5**	71.9	**32.1**	**18**
DC: LSTM, MC: CNN	58.5	35.3	18.7	8.5	14.7	33.7	7.2	6.6
DC: CNN, MC: LSTM	85.6	76.9	67.4	60.5	36.4	**72.1**	29.3	17.9
MSCOCO $\rightarrow$ TGIF								
Both CNN	47.7	28.7	17.4	9.8	14.2	36.6	20.3	10.2
Both LSTM	47.5	28.1	16.8	9.6	14.2	36.3	21	10.1
DC: LSTM, MC: CNN	47.5	28	16.6	9.2	14.2	36.4	19.2	10
DC: CNN, MC: LSTM	47.5	**29.2**	**17.9**	**10.3**	**14.5**	**37**	**22.2**	**10.6**
MSCOCO $\rightarrow$ Flickr30k								
Both CNN	62	41.6	26.8	16.8	16.5	41.5	32.3	10
Both LSTM	61.6	41.1	26.8	17.1	16.4	41.8	32.3	**10.1**
DC: LSTM, MC: CNN	61.3	41	26.7	17.1	16.2	41.3	31.8	9.9
DC: CNN, MC: LSTM	**62.1**	**41.7**	**27.6**	**17.9**	**16.7**	**42.1**	**32.6**	9.9

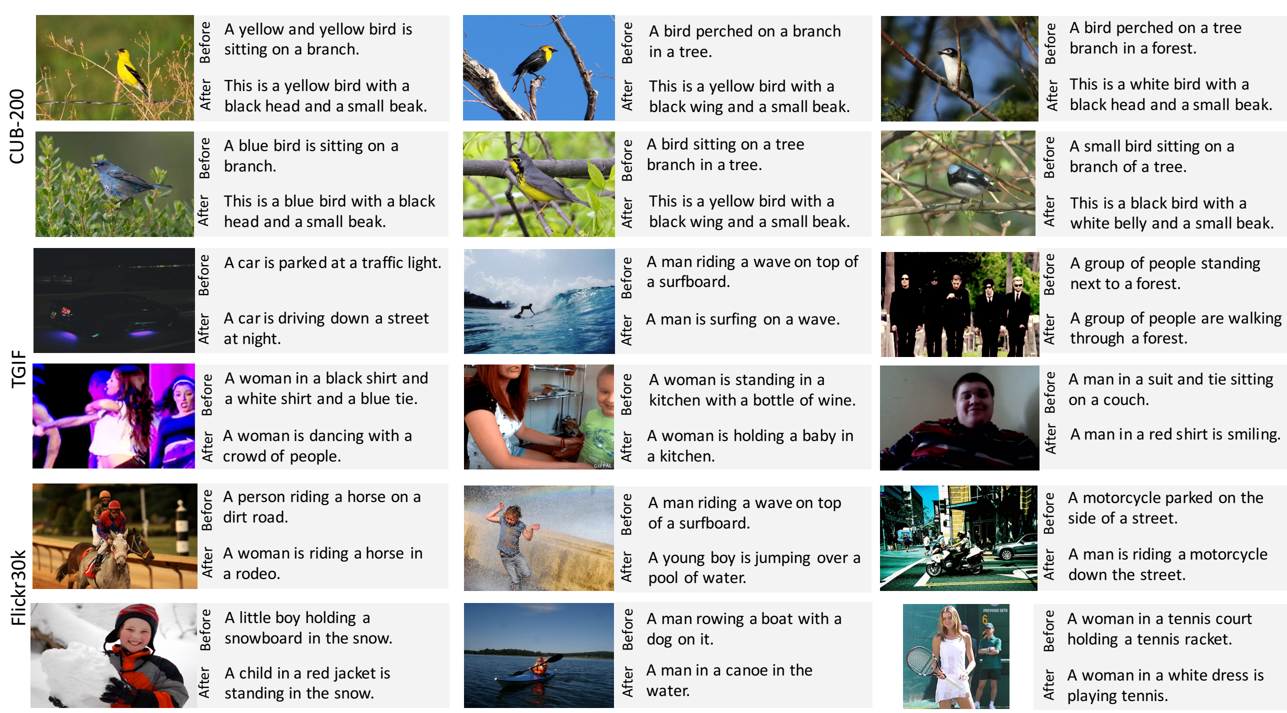

Figure 3: Additional examples of captions before and after domain adaptation.

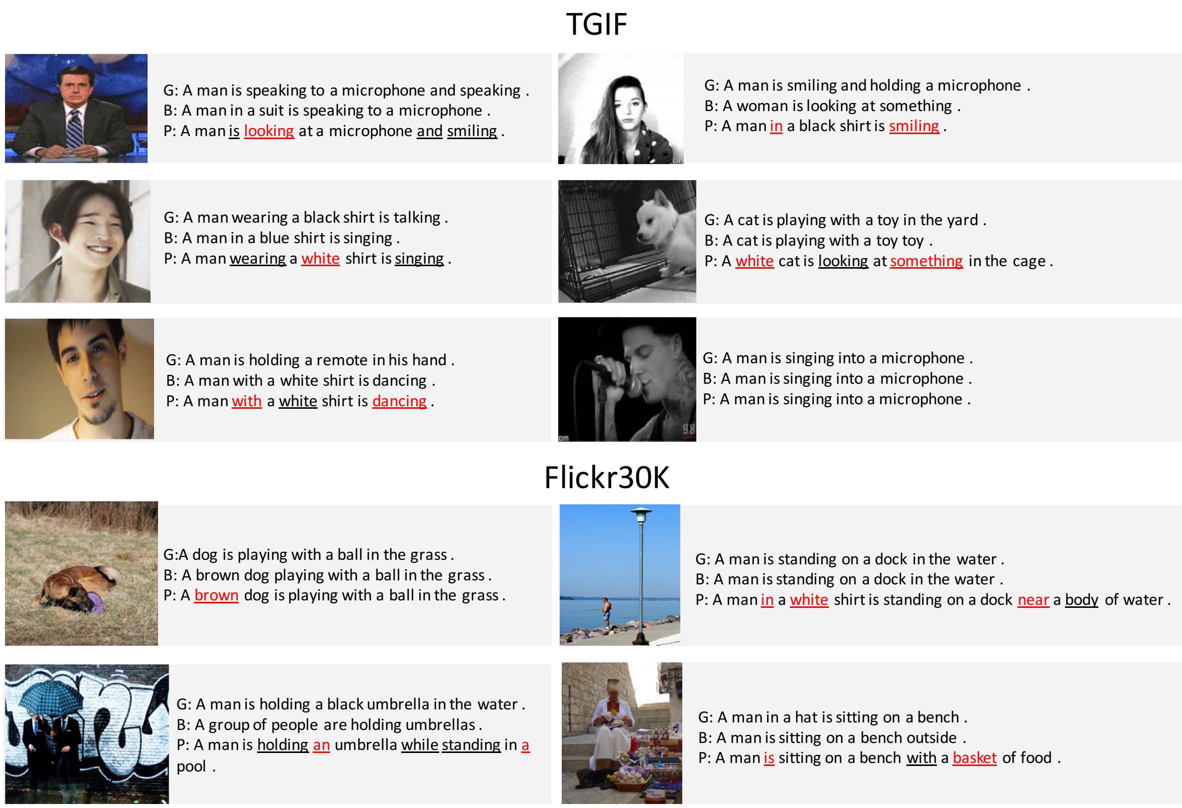

Figure 4: Additional examples of critic-based planning. G stands for greedy search, B for beam search, and P for critic-based planning. The underlined words denote that the difference between the maximum probability and the second largest probability of π is lower than Γ (selected by critic). When critic-based planning does not choose the word with maximum probability of π, the word is colored in red.