BILINEAR REPRESENTATION FOR LANGUAGE-BASED IMAGE EDITING USING CONDITIONAL GENERATIVE ADVERSARIAL NETWORKS

Xiaofeng Mao; Yuefeng Chen; Yuhong Li; Tao Xiong; Yuan He; Hui Xue
Alibaba Group, China

Introduction

What is language-based image editing?

What if you could tell an AI to edit an image just by describing what the new one should look like? Language-based image editing, which edits images using human linguistic input and AI processing, is already starting to see application in fashion, VR, and CAD.

Like in **Fig 1**, using LBIE technique, one can automatically modify the color, texture or style for a given design drawing by language instructions instead of the traditional complex processes.

Existing literature on LBIE using cGAN

The cGAN [1] approach edits the image based on fused visual-text representations using one of two conditioning methods. The first is concatenation. The second improved approach is Feature-wise Linear Modulation (FiLM) [2], which seeks to mimic the human attention mechanism.

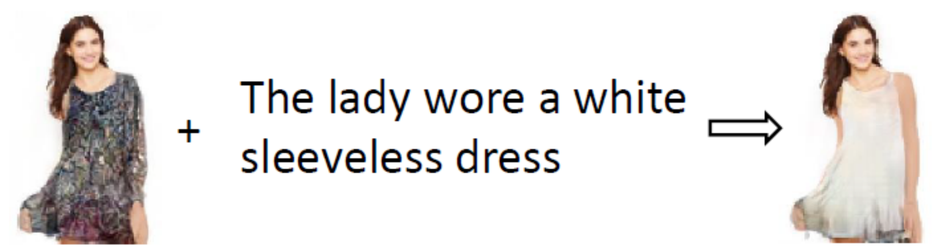
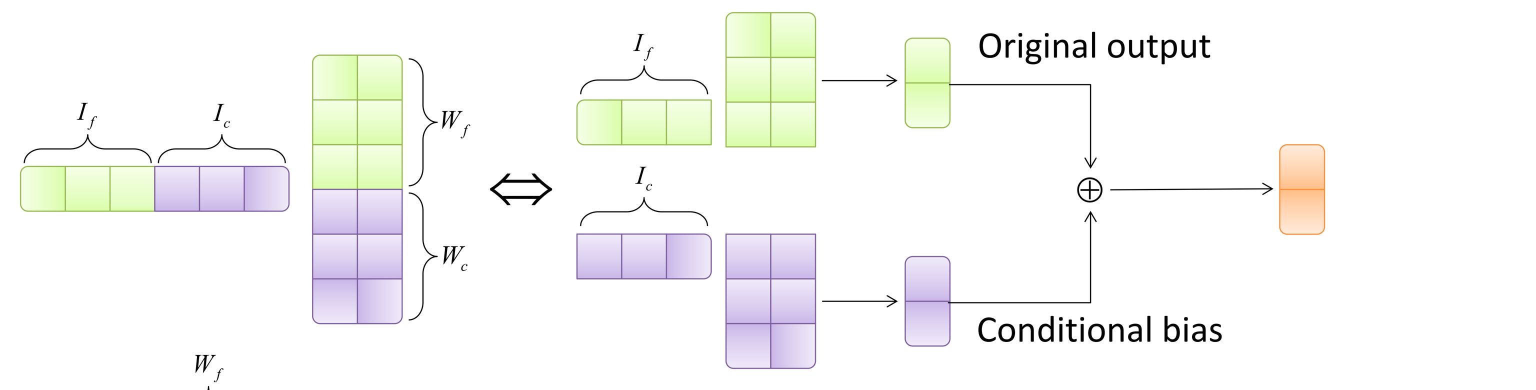


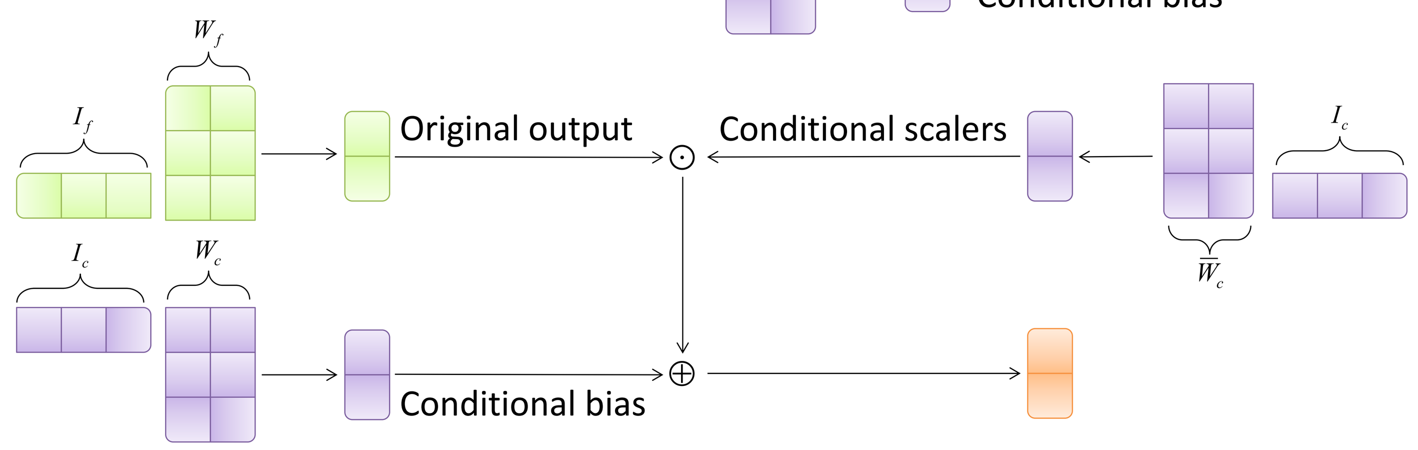
Fig 1. LBIE for fashion generation.

Motivation

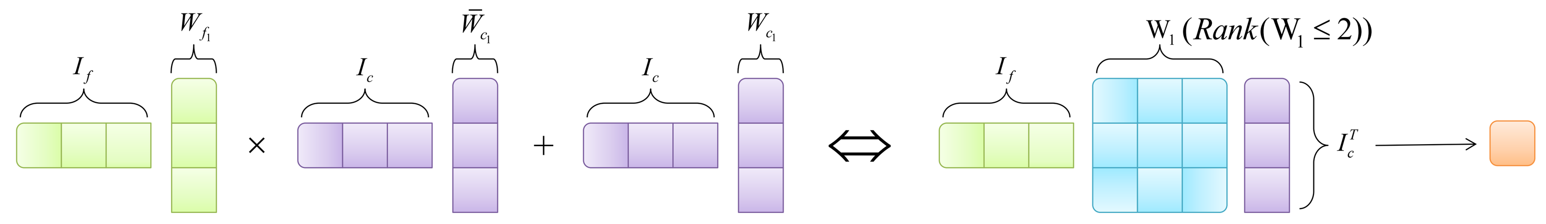
Traditional cGAN:



Feature-wise linear modulation (FiLM) :



Feature-wise bilinear modulation:



Concatenation and FiLM only apply a linear transformation between the input and conditional features. In this work, we go a step further and generalize these linear methods to the more powerful bilinear version, which can provide richer representations than linear models by learning the second-order interaction.

Method

Model details

The network architecture is shown in **Fig 2**. The network consists of a generator *G* and a discriminator *D*. The text and image features are fed in the fusing module, which consists of N Bilinear Residual Layer (BRL). The decoding module ϕ_{up} upsamples the fused feature to a high-resolution images. We propose Bilinear Residual Layer for learning conditional bilinear representations. We add some shortcuts to guarantee model's capability to learn identical mapping, and adopts a low-rank bilinear method [3] to simplified the calculation of bilinear transformation.

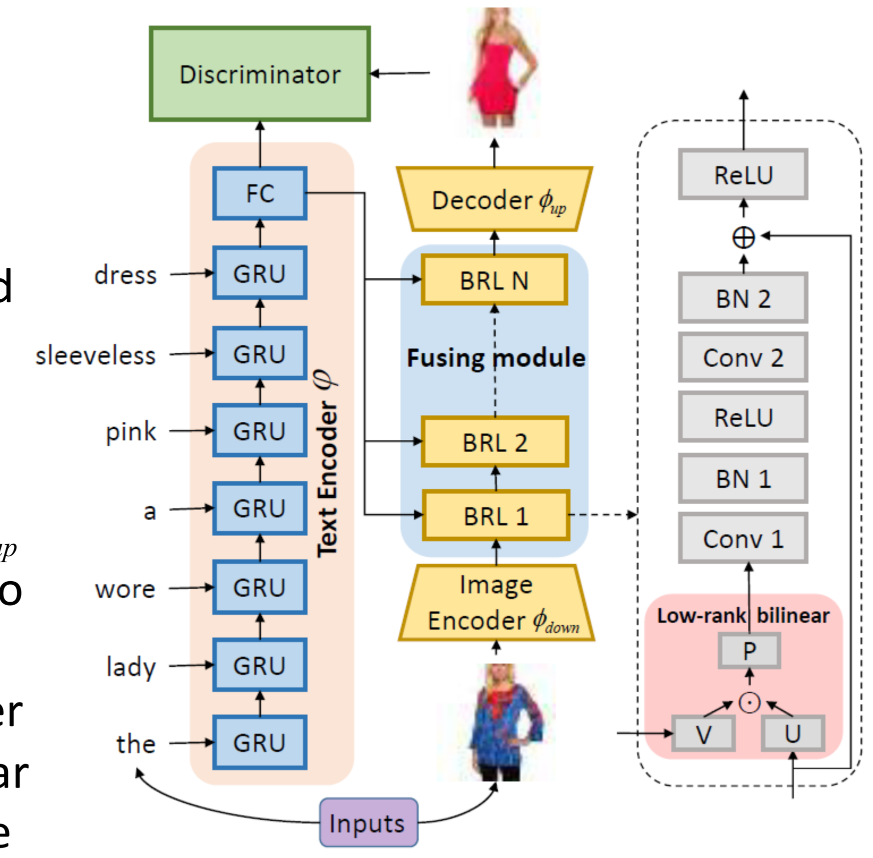


Fig 2. Network overview

Adversarial training objective

For $\bar{t}$ → mismatching text, t → matching text, $\hat{t}$ → manipulating text. The discriminator *D* is trained distinguish semantically differentiated image-text pairs:

$$L_D = \mathrm{E}_{(x,\bar{t})\sim p_{data}}\left[D(x,\varphi(\bar{t}))^2\right] + \mathrm{E}_{(x,t)\sim p_{data}}\left[(D(x,\varphi(t))-1)^2\right] + \mathrm{E}_{(x,\hat{t})\sim p_{data}}\left[D(G(x,\varphi(\hat{t})),\varphi(\hat{t}))^2\right]$$

The generator *G* is trained to generate more semantically similar images with the editing text $\hat{t}$:

$$L_G = \mathrm{E}_{(x,\hat{t})\sim p_{data}}\left[(D(G(x,\varphi(\hat{t})),\varphi(\hat{t}))-1)^2\right]$$

Results

Qualitative evaluation:

Fig 3 shows the performance of traditional cGAN, FiLM and our method on Caltech-200 bird dataset [4], Oxford-102 flower dataset [5] and Fashion Synthesis dataset [6].

Quantitative evaluation:

Inception score (IS) is used for quantitative evaluation. Diverse and meaningful images can get larger inception score. **Table 1** shows IS for traditional cGAN model, FiLM and three variants of our method (Bil-R2, Bil-R64 and Bil-R256 for rank constraint d=2,64,256)

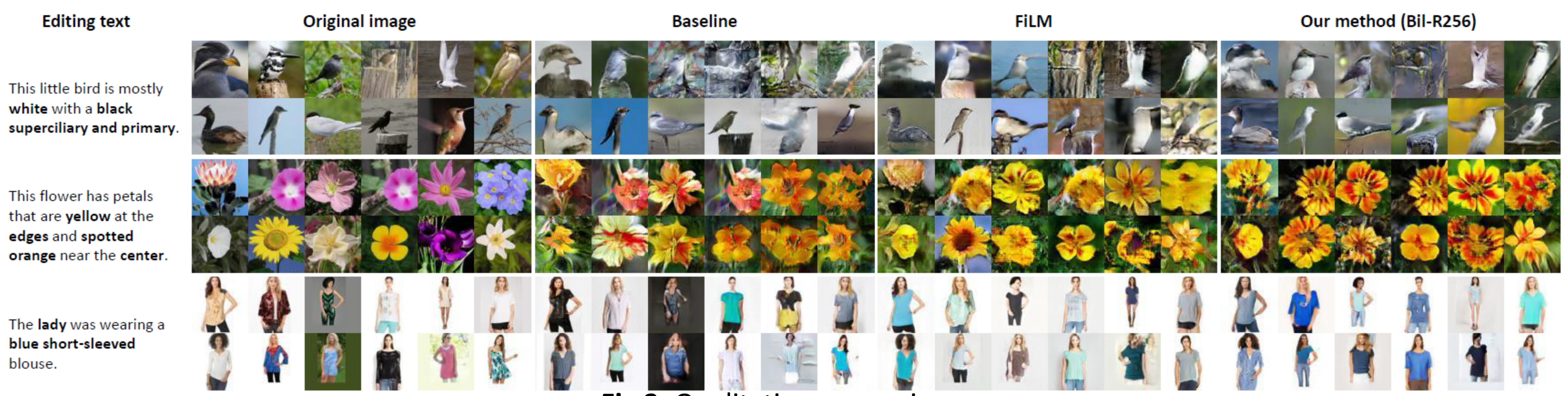


Fig 3. Qualitative comparisons

Methods	Caltech bird	Oxford flower	Fashion
Baseline	1.92±0.05	5.03±0.62	8.65±1.33
FiLM	2.59±0.11	4.83±0.48	8.78±1.43
Bil-R2	2.60±0.11	4.93±0.39	9.30±1.48
Bil-R64	2.63±0.17	5.40±0.62	10.94±2.28
Bil-R256	**2.76±0.08**	**6.26±0.44**	**11.63±2.15**

Table 1. The comparison of IS score of methods

Conclusions

In this work, we propose a conditional GAN based encoderdecoder architecture to semantically manipulate images by text descriptions. A general condition layer called Bilinear Residual Layer (BRL) is proposed to learn more powerful bilinear representations for LBIE. BRL is also applicable for other common conditional tasks. Our evaluation results on Caltech-200 bird dataset, Oxford-102 flower dataset and Fashion Synthesis dataset achieve plausible effects and outperform the state-of-art methods on LBIE.

References

[1] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo, "Semantic image synthesis via adversarial learning," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5706–5714.
[2] Mehmet G¨unel, Erkut Erdem, and Aykut Erdem, "Language guided fashion image manipulation with feature-wise transformations," arXiv preprint arXiv:1808.04000, 2018.
[3] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang, "Hadamard product for low-rank bilinear pooling," arXiv preprint arXiv:1610.04325, 2016.
[4] Wah C., Branson S., Welinder P., Perona P., Belongie S. "The Caltech-UCSD Birds-200-2011 Dataset." Computation & Neural Systems Technical Report, CNS-TR-2011-001.
[5] M-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing, Dec 2008.
[6] Shizhan Zhu, Sanja Fidler, Raquel Urtasun, Dahua Lin, and Chen Change Loy, "Be your own prada: Fashion synthesis with structural coherence," in Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017, pp. 1689–1697.

Contact

Name: Xiaofeng Mao
Department: Turing Laboratory of Alibaba Security Department, Alibaba Group
Address: No. 969 Wenyi West Road, Yuhang District, Hangzhou City, Zhejiang Province
Email: mxf164419@alibaba-inc.com
Phone: +86 18668411821
Github: https://github.com/vtddggg