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
{mxf164419, yuefeng.chenyf, daniel.lyh, weilue.xt, heyuan.hy, hui.xueh}@alibaba-inc.com

ABSTRACT

The task of Language-Based Image Editing (LBIE) aims at generating a target image by editing the source image based on the given language description. The main challenge of LBIE is to disentangle the semantics in image and text and then combine them to generate realistic images. Therefore, the editing performance is heavily dependent on the learned representation. In this work, conditional generative adversarial network (cGAN) is utilized for LBIE. We find that existing conditioning methods in cGAN lack of representation power as they cannot learn the second-order correlation between two conditioning vectors. To solve this problem, we propose an improved conditional layer named Bilinear Residual Layer (BRL) to learning more powerful representations for LBIE task. Qualitative and quantitative comparisons demonstrate that our method can generate images with higher quality when compared to previous LBIE techniques.

***Index Terms*—** Generative adversarial networks, Bilinear, Language-based image editing

1. INTRODUCTION

The task of Language-Based Image Editing (LBIE) aims at manipulating a source image semantically to match the given description well. LBIE has seen applications to domains as diverse as Computer-Aided Design (CAD), Fashion Generation and Virtual Reality (VR) [1]. As illustrated 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.

Nevertheless, LBIE is still challenging due to the following two difficulties: i). the model should find the areas in image which are relevant to the given text description; ii). the relations of disentangled semantics in image and text description should be learned for a better generation of realistic image. To tackle these problems, several methods have been proposed [1, 2, 3, 4, 5], and most of them utilize the generative models, e.g., GANs [6]. [1, 2] divide LBIE into two subtasks: language-based image segmentation and image generation. Specifically, Zhu et al. [2] performes LBIE to "redress" the person with the given outfit description, while at the same time keeping the wearer and his posture or expression unchanged. They use a two stages GAN that outputs a semantic segmentation map as intermediate step, which is further used to render the final image with precise regions and textures at the second step. Some other approaches [3, 4, 5] can achieve LBIE without any segmentation map or explicit spatial constraints by adversarially train a conditional GAN [7]. Among them, [5] is the seminal work and it uses concatenation operation to condition the image generation process with text embeddings. [3, 4] follow up this framework, and replace the concatenation operation with Feature-wise Linear Modulation (FiLM), which is a more efficient and powerful method as a generalization of concatenation.

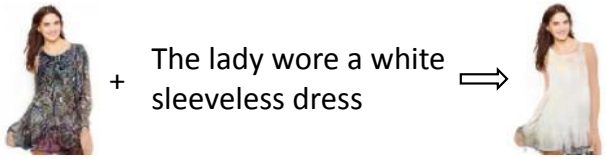

Fig. 1. LBIE for fashion generation.

In this work, we first theoretically analyse these works which edit the image based on fused visual-text representations using different conditioning methods. We found that all these conditioning methods can be modeled by a universal form of bilinear transformation based on [8]. However, all these methods are lack of representation power as they cannot learn the second-order correlation between two conditioning embeddings. To solve this problem, we present an improved conditoning method named Bilinear Residual Layer which can strike a happy compromise between representation effectiveness and model efficiency. We have both theoretically and experimentally proved that the Bilinear Residual Layer can provide richer representations than previous approaches. Quantitative and qualitative results on Caltech-200 bird [9], Oxford-102 flower [10] and Fashion Synthesis datasets [2] suggest that our approach can generate images with higher quality when compared to previous LBIE techniques.

2. METHOD

In this section, we first theoretically analyse existing conditioning methods in cGANs. Then an improved conditional

layer called Bilinear Residual Layer (BRL) is proposed in Sec 2.2. Finally, we introduce overall framework in Sec 2.3.

2.1. Overview of conditioning methods

Conditioning is a general-purpose operation and can be used for different tasks, e.g., conditional image generation [11, 12] and cross-modality distillation [13]. The most commonly used approach in conditional GANs is concatenation. Formally, denote $I_f \in \mathbb{R}^D$ and $I_c \in \mathbb{R}^{D'}$ as the output of previous layer and conditioning feature respectively, where D and D' are the dimensionality of features. The concated representation $[I_f\ I_c] \in \mathbb{R}^{D+D'}$ can be further encoded by a matrix $W = [W_f; W_c]$, $W_f \in \mathbb{R}^{D\times O}$ and $W_c \in \mathbb{R}^{D'\times O}$ are the corresponding weights for I_f and I_c. O is the output dimension. Formally, we can get the following transformation:

$$I_o = [I_f\ I_c]\,[W_f; W_c] = I_f W_f + I_c W_c \tag{1}$$

where I_o is the output tensor. Equation 1 suggests that concatenation based conditioning method amounts to adding a feature-wise bias on the unconditional output $I_f W_f$. Therefore, some other approaches [14, 15] suggest to add conditional bias directly instead of concatenation.

Recently, some works [16] have validated that deep models could mimic the human attention mechanism by gating each feature using a value between 0 and 1. Inspired by this, Perez et al. [17] proposes a more general conditioning method named feature-wise linear modulation (FiLM), which rescales the features by adding multiplicative interactions:

$$I_o = (I_f W_f) \odot (I_c \overline{W}_c) + I_c W_c \tag{2}$$

$\overline{W}_c \in \mathbb{R}^{D'\times O}$ is the weight for learning rescaling coefficients. From this formulation, we can conclude that concatenation is a special case of FiLM when $I_c\overline{W}_c = \mathbb{1}$, where $\mathbb{1}$ is a matrix of ones. FiLM has shown its superiority over conventional concatenation method and has been widely applied to the multimodal interaction.

However, 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. In bilinear model, the ith feature in output I_o can be calculated as

$$I_{o_i} = I_f \mathbf{W_i} I_c^T \tag{3}$$

$\mathbf{W_i} \in \mathbb{R}^{D\times D'}$ is a weight matrix for the output feature I_{o_i}. Interestingly, we have found FiLM can be presented by bilinear transformations. Denote the weights corresponding to the ith output feature in W_f, $\overline{W}_c$ and W_c as w_{f_i}, $\overline{w}_{c_i}$ and w_{c_i}. The FiLM transformation for $I_{o_i} = (I_f w_{f_i})(I_c \overline{w}_{c_i}) + I_c w_{c_i}$ can be represented by

$$I_f \underbrace{(w_{f_i}\overline{w}_{c_i}^T + \mathbf{W_i}')}_{\mathbf{W_i}} I_c^T \tag{4}$$

where $I_f \mathbf{W_i}' = w_{c_i}^T$, $\mathbf{W_i}'$ can be constructed by randomly choosing a nonzero element I_{f_k} in I_f, we have

$$\mathbf{W_i}' = \begin{bmatrix} 0 & 0 & 0 & \cdots \\ \vdots & \vdots & & \\ \frac{w_{c_{i1}}}{I_{f_k}} & \frac{w_{c_{i2}}}{I_{f_k}} & \cdots & \frac{w_{c_{iD'}}}{I_{f_k}} \\ \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots \end{bmatrix} \tag{5}$$

where elements in the $\mathbf{W_i}'$ are 0 except for the kth row. Obviously, the rank of matrix $w_{f_i}\overline{w}_{c_i}^T$ and $\mathbf{W_i}'$ are both 1. So we have $Rank(\mathbf{W_i}) \leq 2^*$. The constructed formulation indicates that FiLM is equivalent to bilinear transformation with transformation matrix $\mathbf{W_i}$ is sparse and has the rank no greater than 2. From a theoretical perspective, it illustrates that bilinear transformations can provide more fine-grained conditioning representations than the concatenation and FiLM.

2.2. Bilinear Residual Layer

We propose Bilinear Residual Layer (BRL) for learning conditional bilinear representations as illustrated in dashed box of Fig 2. Similar to FiLM, we add shortcuts to guarantee the model's capability to learn identical mapping. As a consequence, our bilinear residual layer can automaticly decide whether or not the model needs to incorporate the conditioning information in the later layers.

However, the representational power of bilinear features comes with the cost of very high-dimensional model parameters , which require substantial computing and large quantities of training data to fit [18]. For example, the dimensionality of $\mathbf{W}$ is $|D \times D' \times O|$ which is cubical expansion. To reduce the dimensionality of model parameters, our approach adopts a low-rank bilinear method [19] to reduce the rank of $\mathbf{W_i}$. Based on this idea, I_{o_i} can be rewritten as follows:

$$I_{o_i} = I_f \mathbf{W_i} I_c^T = I_f \mathbf{U_i} \mathbf{V_i}^T I_c^T = I_f \mathbf{U_i} \odot I_c \mathbf{V_i} \tag{6}$$

where $\mathbf{U_i} \in \mathbb{R}^{D\times d}$ and $\mathbf{V_i} \in \mathbb{R}^{D'\times d}$ are the decomposed submatrices and they restrict the rank of $\mathbf{W_i}$ to be at most $d \leq \min(D, D')$. Then the final feature vector I_o can be projected by $\mathbf{P} \in \mathbb{R}^{O\times d}$ as follows:

$$I_o = \mathbf{P}(I_f \mathbf{U} \odot I_c \mathbf{V}) \tag{7}$$

Moreover, Our bilinear residual layer is a general condition layer, and it is applicable not only for LBIE, but also for other conditional models or applications, e.g., text-to-image generation [20]. In following sections, we will present the overall framework of our work and we denote the bilinear residual layer as $\mathcal{F}$ for convenience.

*Properties of rank: https://en.wikipedia.org/wiki/Rank_(linear_algebra)

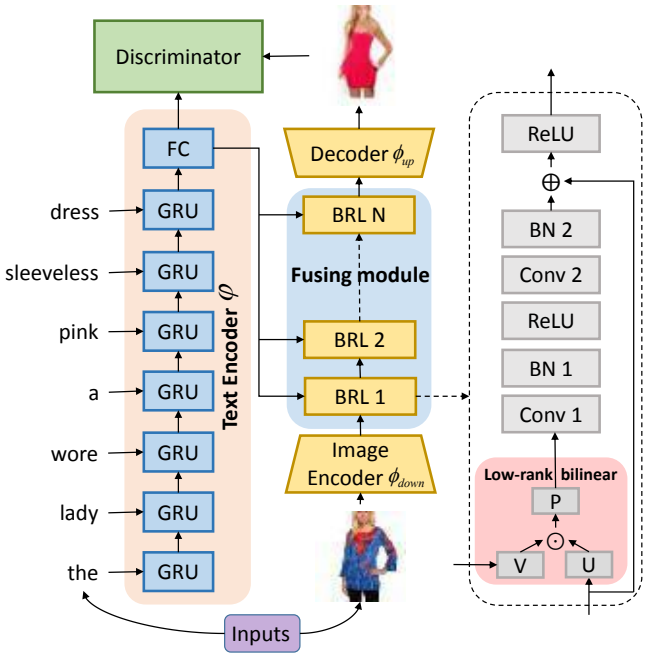

Fig. 2. Overview of our network architecture. Detail of our bilinear residual layer is presented in the dashed box.

2.3. Overall framework

We follow the work of Dong et al. [5] which utilizes the cGAN to learn the target mapping conditioning based on image and text description. As shown in Fig 2, the network consists of a generator G and a discriminator D. The generator has three modules: encoding module, fusing module and decoding module. Encoding module contains pre-trained encoders φ and ϕ_{down}, and they are used to extract text and image features respectively. We adopt the procedure in [21] to pre-train the text encoder φ and use parameters of *conv*1-4 layers in VGG16 as the feature extractor ϕ_{down} for image. The text and image features are then fed in the following fusing module, which can be seen as a conditioning layer to compromise the semantics of multiple modalities. The final decoding module ϕ_{up} upsamples the fused feature to a high-resolution images. Finally, the discriminator is a classifier which takes the generated image and text embeddings as input and output the probability whether the description matches the image.

Formally, given an original image-text pair $<x, t>$, t is the text matching with the image x. Suppose that we use description text $\hat{t}$ to manipulate the image x, typically $\hat{t}$ is a text relevant to x. The generator can transform the image according to text embedding $\varphi(\hat{t})$ and output

$$G(x, \varphi(\hat{t})) = \phi_{up}(\mathcal{F}(\phi_{down}(x), \varphi(\hat{t}))) \tag{8}$$

the discriminator D is trained to distinguish semantically differentiated image-text pairs. To this end, we need to take a mismatching text $\bar{t}$ as negative sample. Original pair $<x, t>$, current editing pair $<x, \hat{t}>$ and negative pair $<x, \bar{t}>$ are fed into discriminator D to minimizing

$$\begin{aligned}\mathcal{L}_D = &\mathbb{E}_{(x,\bar{t})\sim p_{data}}\left[D(x, \varphi(\bar{t}))^2\right] \\ &+ \mathbb{E}_{(x,t)\sim p_{data}}\left[(D(x, \varphi(t)) - 1)^2\right] \\ &+ \mathbb{E}_{(x,\hat{t})\sim p_{data}}\left[D(G(x, \varphi(\hat{t})), \varphi(\hat{t}))^2\right]\end{aligned} \tag{9}$$

here the objective of the first and second terms is to classify negative and original real-world image-text pairs. The third term makes D to identify the synthesized image with its editing text as mismatching as possible. Alternately to the training of D, the generator G is trained to generate more semantically similar images with the editing text $\hat{t}$:

$$\mathcal{L}_G = \mathbb{E}_{(x,\hat{t})\sim p_{data}}\left[(D(G(x, \varphi(\hat{t})), \varphi(\hat{t})) - 1)^2\right] \tag{10}$$

In this work, $\bar{t}$ and $\hat{t}$ are selected from the text descriptions of other images in the dataset.

3. EXPERIMENTS

We conduct experiments on Caltech-200 bird dataset [9], Oxford-102 flower dataset [10] and Fashion Synthesis dataset [2]. The bird dataset has 11,788 images with 200 classes of birds. We split it to 160 training classes and 40 testing classes. The flower dataset has 8,189 images with 102 classes of flowers, and we split it to 82 training classes and 20 testing classes. The fashion dataset has much more classes with 78,979 images totally. We choose 3200 classes from 4119 for training and the rest for testing.

3.1. Implementation details

The source code has been released[†]. Our encoder φ for text descriptions is a recurrent network. Given the pair of image and text $<x, t>$, the method in [21] is used to pre-train the text encoder to minimize the pair-wise ranking loss. This pre-trained text encoder encodes the text description t into visual-semantic text representation $\varphi(t)$, which will be further used in the adversarial training process as detailed in Sec 2.3.

For image encoder, it receives images with size of 64×64 as input and output features with dimension of 16×16×512. Text encoder encodes descriptions to the text embeddings with dimensionality of 128. Our fusing module consists of 4 (i.e., N in the Fig. 2) bilinear residual layers. To implement the low-rank bilinear method, we duplicate the text embeddings to be of dimension 16×16×128, so as to keep the same spatial size with image features. Then the dimensions of both text and image features are reduced to d (cf. Section 2.2) by using 1×1 convolutions. The decoding module consists of several upsampling layers that transform the learned representations into 64×64 images. For the discriminator, we first apply convolutional layers to encode the images

[†]https://github.com/vtddggg/BilinearGAN_for_LBIE

Fig. 3. Qualitative comparisons.

into feature representations. We then concatenate the image representation with text embeddings, then apply two convolutional layers to produce final probabilities. Note that we use concatenation to conditioning for limiting the discriminator capability to prevent the mode collapse effect.

To train the generator and discriminator, we adopt the Adam optimizer with momentum of 0.5. The learning rate is 0.0002. We set batch size to 64 for all three experiments and number of iterative epochs to 600 for birds and flowers synthesis, 200 for fashion synthesis. The parameters of VGG part were fixed during training the generator. The training takes about 1 day to converge on a single Tesla P100 GPU.

3.2. Qualitative comparison

We compare our proposed model with the baseline [5] (i.e., concatenation) and FiLM on three datasets. The results are shown in Fig 3. Baseline method fails to transform the detail attributes based on the given description because the learned representations are not powerful as it does not contain enough detail information. For example, the generated images by baseline method in editing flowers demonstrate the model has learned the colors of yellow and orange, but it is unaware of the location of these colors. Meanwhile, when original image has a complex background (e.g. 3th and 4th samples in first row), the model will fall into mode collapse and output the same meaningless image. On the contrary, our method can capture the specific semantic changes in detail, which is attribute to our richer bilinear representations. It correctly disentangles semantically related objects from some messy images and prevent the occurrence of mode collapse. As a consequence, our approach can successfully generate meaningful images subject to the text description.

3.3. Quantitative comparison

We choose inception score (IS) for quantitative evaluation. Inception score is a well-known metric for evaluating GANs. IS can be computed by $IS = \exp(\mathbb{E}_x D_{KL}(p(y|x)\|p(y)))$, where x denotes one generated sample, and y is the label predicted by the inception model. The better models which generate diverse and meaningful images can get larger inception score. In this experiment, we use the test dataset for evaluation. We first finetune the inception model with test images for classification. Then, for every test class, we randomly choose an image and text description (e.g. if test dataset has 40 classes, 40 images and 40 descriptions are selected). The images are generated by inputting every pair of images and descriptions (e.g. 40 images and 40 descriptions can generate 40×40 edited images).

Methods	Caltech bird	Oxford flower	Fashion
Baseline	1.92±0.05	5.03±0.62	8.65±1.33
FiLM [4]	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

The results are shown in Table 1. To explore the influence of rank constraint d, we set $d = 2, 64, 256$ and get three variants: Bil-R2, Bil-R64 and Bil-R256. The Bil-R256 gets the highest IS in all three tasks. Interestingly, the baseline method has higher IS than FiLM on Oxford flower dataset because flower editing is simple and is not very dependent on the power of learned representation. For more complicated bird and fashion editings, our method gets the highest IS and achieves better performance with the increasing of d. Experimental result suggests that the learned bilinear representation is more powerful and do help to generate images with higher quality.

4. CONCLUSION

In this work, we propose a conditional GAN based encoder-decoder 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.

5. REFERENCES

[1] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu, "Language-based image editing with recurrent attentive models," *arXiv preprint arXiv:1711.06288*, 2017.

[2] 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.

[3] Varun Manjunatha, Mohit Iyyer, Jordan Boyd-Graber, and Larry Davis, "Learning to color from language," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, vol. 2, pp. 764–769.

[4] Mehmet Günel, Erkut Erdem, and Aykut Erdem, "Language guided fashion image manipulation with feature-wise transformations," *arXiv preprint arXiv:1808.04000*, 2018.

[5] 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.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[7] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[8] Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio, "Feature-wise transformations," *Distill*, 2018, https://distill.pub/2018/feature-wise-transformations.

[9] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," Tech. Rep., 2011.

[10] 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.

[11] Qiang Huang, Philip Jackson, Mark D Plumbley, and Wenwu Wang, "Synthesis of images by two-stage generative adversarial networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2018.

[12] Zhu-Liang Chen, Qian-Hua He, Wen-Feng Pang, and Yan-Xiong Li, "Frontal face generation from multiple pose-variant faces with cgan in real-world surveillance scene," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1308–1312.

[13] Siddharth Roheda, Benjamin S Riggan, Hamid Krim, and Liyi Dai, "Cross-modality distillation: A case for conditional generative adversarial networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2926–2930.

[14] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, pp. 125–125.

[15] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al., "Conditional image generation with pixelcnn decoders," in *Advances in Neural Information Processing Systems*, 2016, pp. 4790–4798.

[16] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[17] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville, "Film: Visual reasoning with a general conditioning layer," in *http://www. aaai. org/Conferences/AAAI/aaai. php*, 2018.

[18] Shu Kong and Charless Fowlkes, "Low-rank bilinear pooling for fine-grained classification," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 7025–7034.

[19] 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.

[20] Han Zhang, Tao Xu, and Hongsheng Li, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 5908–5916.

[21] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.