Semantically Consistent Hierarchical Text to Fashion Image Synthesis with an enhanced-Attentional Generative Adversarial Network

Kenan E. Ak[1,2] Joo Hwee Lim [2] Jo Yew Tham[3] Ashraf A. Kassim[1]
[1]National University of Singapore, Singapore
[2]Institute for Infocomm Research, A*STAR, Singapore
[3]ESP xMedia Pte. Ltd., Singapore
emir.ak@u.nus.edu, joohwee@i2r.a-star.edu.sg, thamjy@espxmedia.com, ashraf@nus.edu.sg

Abstract

In this paper, we present the enhanced Attentional Generative Adversarial Network (e-AttnGAN) with improved training stability for text-to-image synthesis. e-AttnGAN's integrated attention module utilizes both sentence and word context features and performs feature-wise linear modulation (FiLM) to fuse visual and natural language representations. In addition to multimodal similarity learning for text and image features of AttnGAN [28], cosine and feature matching losses of real and generated images are included while employing a classification loss for "significant attributes". In order to improve the stability of the training and solve the issue of model collapse, spectral normalization and two-time scale update for the discriminator are used together with instance noise. Our experiments show that e-AttnGAN outperforms state-of-the-art methods on the FashionGen and DeepFashion-Synthesis datasets.

1. Introduction

The focus of the work presented in this paper is the task of text-to-image generation which aims to produce realistic images that match text descriptions. Recently introduced Attentional Generative Adversarial Network (AttnGAN) [28] has improved both image quality and text-image similarity compared to the previous methods [20, 31]. AttnGAN is comprised of hierarchically connected attentional generative networks to gradually generate multi-scale images with an attention model over word embeddings in the network.

In this paper, we present an enhanced version of AttnGAN called e-AttnGAN. The e-AttnGAN incorporates an integrated attention module which includes both word and sentence context features in the image generation process with **F**eature-w**i**se **L**inear **M**odulation (FiLM) [18] layers which have the ability to manipulate the visual features without extra supervision.

Figure 1. Text-to-image synthesis examples of e-AttnGAN using the FashionGen dataset [21]

In addition to the integrated attention module, e-AttGAN includes the following enhancements for improving text/image similarity and for stabilizing training (1) cosine similarity and feature matching [22] learning between the generated and real samples to guide the generator network on the expected data representations, (2) classification losses to ensure that the generated image consists of important attributes such as clothing category and color, (3) spectral normalization [17] to address the instabilities in the training of AttnGAN, (4) two-time-scale update rule [12] for the discriminator network that is affected by the spectral normalization, and (5) instance noise to inputs of the discriminator network to evade "mode collapse".

Figure 1 presents some images generated by our proposed e-AttnGAN based on text descriptions from the FashionGen dataset [21]. It is evident that e-AttnGAN is able to generate high-quality precise images that are semantically consistent with the desired clothing attributes such as clothing category, color, sleeve length and pattern. An earlier version of e-AttnGAN which lacked the integrated attention module and stabilized training attained the second rank at the FashionGen Challenge [21] held at ECCV2018 workshop Computer Vision for Fashion, Art and Design.

2. Related Work

Generative Adversarial Networks (GANs) introduced by Goodfellow et. al [9] have demonstrated remarkable suc-

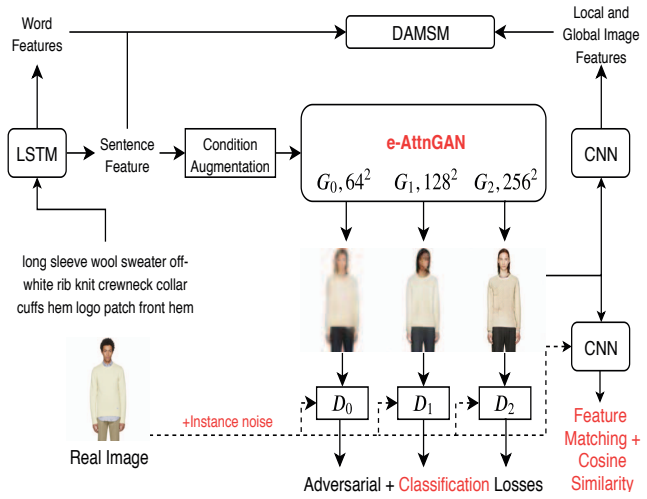

Figure 2. e-AttnGAN architecture for text-to-image synthesis.

cess in many computer vision problems including image generation [22, 19], image-to-image translation [13, 7]. GANs consist of a generator and a discriminator networks where they compete with each other in a minmax game.

Reed et. al [20], the first to use GANs for text-to-image synthesis, is able to generate low-resolution images (64^2). StackGAN++ [31] aim to generate realistic images using tree-like structures with multiple generators and discriminators. AttnGAN [29] has a similar structure with StackGAN++ but additionally consists of attention modules and a deep similarity model.

Fashion related research include [15, 24], attribute discovery [11, 26], recommendation [5, 23], retrieval [8, 2, 1, 3] , fashion parsing [14, 30] as well as GANs [32, 27, 4].

3. e-AttnGAN

In this section, following a brief description of the AttnGAN architecture [28], we present details of enhancements made to realize the e-AttnGAN architecture which is summarized in the block diagram of Figure 2.

Attentional adversarial generative network (AttnGAN) uses attention over word embeddings within an input sequence to generate images guided by the deep attentional multimodal similarity model (DAMSM). The training of DAMSM [29] is based on the multimodal similarity between text and image which is made possible with joint cooperation of a Convolutional Neural Network (CNN) and a bi-directional Long Short-Term Memory (LSTM). Word representations are extracted by concatenating two hidden states of a word. A global vector which represents the sentence is created by concatenating the last hidden state of the bi-directional LSTM while local and global image features are extracted from the Inception-v3 network [25]. AttnGAN consists of hierarchically connected discriminators D_i, and generators G_i for different resolutions and training is based on multimodal similarity losses between "word features vs local image features" and "sentence features vs global image features" in addition to adversarial loss.

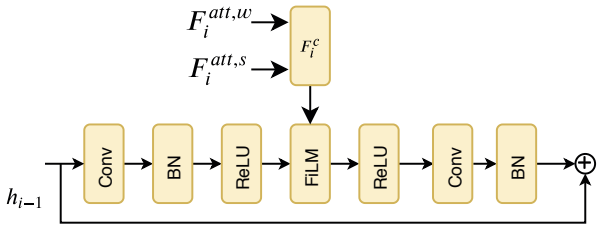

Figure 3. FiLM-ed resblock structure.

3.1. Improved Conditioning with FiLM Layers.

In addition to word-context features (attention model) $F_i^{att,w}$ proposed in AttnGAN, we include sentence-context features $F_i^{att,s}$ to the generation process of e-AttnGAN. Rather than simply concatenating visual and language descriptions to calculate the next hidden feature h_i, we follow an advanced procedure to fuse language and visual representations. FiLM [18] applies a feature-wise affine transformation to the output of a convolutional block, where the transformation weights are conditioned on language. This process enables language features to modify visual features defined as follows:

$$\gamma_i = \mathbf{W}_{\gamma,i} F_i^c \qquad \beta_i = \mathbf{W}_{\beta,i} F_i^c$$
$$\mathbf{Z}'_i = \gamma_i \circ h_{i-1} c_i + \beta_i \tag{1}$$

where β_i, γ_i are modeled as convolutional layers and the output is defined as $\mathbf{Z}_i$, F_i is concatenated form of $F_i^{att,w}$ and $F_i^{att,s}$. As can be seen from Figure 3 and Eq. 1, the output is estimated as a combination of visual and language descriptions. Figure 3 shows the combined form of FiLM and residual layers which is called FilM-ed resblock [10, 16]. We replace res-blocks in AttnGAN with FilM-ed resblocks.

3.2. Deep Similarity & Attribute Learning

The aim of employing similarity learning is to establish a new semantic feature space where image and language features are similar to each other. In addition to the DAMSM used in AttnGAN [28], we include feature matching, cosine similarity and classification losses between real and generated images to further improve text/image correlation and image quality.

Feature matching & cosine similarity losses A trick to avoid the instabilities in the training which is also discussed in Salimans et. al [22] is called feature matching. It can be applied to the generator and the aim is to generate data that is close to the real ones in the feature space. We define the feature matching loss for the generator as:

$$L_{feat}^G = \lambda_{feat} ||f(x^*) - f(x)||_2^2 \tag{2}$$

where f represents the final layer from CNN in DAMSM. Similarly, we define a cosine similarity loss using x^* and x and denote it L^G_{sim} with λ_{sim} weight.

Classification loss. We impose classification losses to make sure the network pays more attention to "significant attributes" such as clothing category, color, sleeve, etc. In order to ensure that these attributes are present in the generated image, we add an additional convolutional layer at the end of each discriminator and impose classification losses on both D and G. The classification loss for D is defined as:

$$L^D_{cls} = \lambda_{cls} \sum_{i=1}^{n} \Big[E_x[-logD_{i,cls}(c|x_i)] \Big] \quad (3)$$

where x_i represents real images in different resolutions and $D_{i,cls}(c|x_i)$ is the probability distribution of classifying x_i as the corresponding domain label c.

While D is trained to classify the real image with its corresponding label, we enforce a similar loss function for the generator network to ensure that the generated images x^*_i consist of desired attributes defined as:

$$L^G_{cls} = \lambda_{cls} \sum_{i=1}^{n} \Big[E_{x^*}[-logD_{i,cls}(c|x^*_i)] \Big] \quad (4)$$

where the term $D_{i,cls}(c|x^*_i)$ is calculated using the additional classification layers of D_i's. Note that, we do not include an additional input to signal desired attributes as that information is already available in language descriptions.

e-AttnGAN uses three generators to generate 256x256 images as shown in Figure 2. We use the same DAMSM structure from AttnGAN for feature extraction and limit number of words to 15. The loss term can be jointly optimized with $\lambda_{cls} = 10$, $\lambda_{feat} = 10$, $\lambda_{sim} = 50$, in addition to adversarial and DAMSM losses of AttnGAN.

4. Stabilizing the Training of e-AttnGAN

Several methods were adopted to improve the training and address instabilities have proven to be effective. The first is Spectral Normalization by Miyato et. al [17], a computationally inexpensive weight normalization technique that can be used to address the instabilities in training GANs without requiring any parameter tuning. We also adopt the two time-scale update rule (TTUR) introduced by Heusal et. al [12] which proposes the use of individual learning rates for the discriminator and generator networks to accelerate the training. An issue when training GANs is "mode collapse" where the generator network collapses and starts to generate a limited range of samples. To address this, we add instance noise to the inputs of the discriminator as in [6].

5. Experiments

In this section, we compare e-AttnGAN with state-of-the-art methods and also present an ablation study. We choose to use FashionGen [21] and DeepFashion-Synthesis [32] datasets as they include text descriptions to describe each image. For comparison, the inception score [22], R-precision and classification accuracy are used.

Table 1. Quantitative results on FashionGen dataset.

	Inception Score	R-precision (%)
StackGAN++	5.46 ± 0.13	17.5
AttnGAN	7.94 ± 0.13	68.28
e-AttnGAN	**8.97 ± 0.15**	**72.00**
w/ integ. att.	8.87 ± 0.12	67.79
w/ integ. att. (no sent.)	8.27 ± 0.12	65.73
w/ L^D_{cls}, L^G_{cls}	10.41 ± 0.18	71.71
w/ $L^G_{feat} + L^G_{sim}$	9.27 ± 0.22	67.79

Table 2. Quantitative results on DeepFashion dataset.

	Inception Score	R-precision (%)	Avg. Cls. Acc. (%)
StackGAN++	1.74 ± 0.02	12.3	37.08
AttnGAN	4.12 ± 0.06	70.73	56.18
e-AttnGAN	**4.77 ± 0.10**	**76.21**	**58.39**
w/ integ. att.	4.11 ± 0.10	67.99	53.3
w/ integ. att. (no sent.)	3.72 ± 0.04	62.85	54.87
w/ L^D_{cls}, L^G_{cls}	4.75 ± 0.14	71.29	59.02
w/ $L^G_{feat} + L^G_{sim}$	4.28 ± 0.11	71.53	56.55

5.1. Results

The quantitative results reported in Tables 1 and 2 clearly show that e-AttnGAN outperforms both AttnGAN and StackGAN++ for all three evaluation metrics in both datasets. These results show that the enhancements enable e-AttnGAN to generate more realistic images (Inception score) while improving text/image correlation (R-precision). For classification accuracy, e-AttnGAN outperforms AttnGAN by 2.21% .

Ablation studies are reported at the last four rows of Tables 1 and 2 to examine the effects of enhancements compared to AttnGAN. The proposed integrated attention module "$w/integ.\ att.$" increases the inception score by 0.93 for the FashionGen dataset. In terms of R-precision and classification accuracy, a performance drop is observed due to the increased complexity and unstabilized training. The inclusion of sentence-context features "$w/integ.\ att.\ (nosent.)$" improves the performance for both datasets. The most successful enhancement is the inclusion of classification losses "$w/L^D_{cls}, L^G_{cls}$" which results on improvements for all metrics. Lastly, including feature similarity and matching objectives "$w/L^G_{feat} + L^G_{sim}$" improves the inception score but decreases R-precision in FashionGen dataset. It should be noted that some of the proposed enhancements may increase certain metrics while decreasing others which we believe is due to the increased complexity of appended enhancements. This problem was overcome by using the proposed training stabilization techniques which enables e-AttnGAN to achieve state-of-the-art results.

6. Conclusion

The e-AttnGAN proposed in this paper for synthesizing fashion images from the text descriptions has been shown to have major performance improvements over state-of-the-art methods. A possible future direction would be solving problems when generating humans parts as e-AttnGAN tends to focus more on text/image correlation.

References

[1] Kenan E. Ak, Ashraf A. Kassim, Joo Hwee Lim, and Jo Yew Tham. Learning attribute representations with localization for flexible fashion search. In *CVPR*, June 2018.

[2] Kenan E. Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A. Kassim. Efficient multi-attribute similarity learning towards attribute-based fashion search. In *WACV*, pages 1671–1679. IEEE, 2018.

[3] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Which shirt for my first date? towards a flexible attribute-based fashion query system. *Pattern Recognition Letters*, 112:212–218, 2018.

[4] Kenan E. Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A. Kassim. Attribute manipulation generative adversarial networks for fashion images. In *ICCV*. IEEE, 2019.

[5] Ziad Al-Halah, Rainer Stiefelhagen, and Kristen Grauman. Fashion forward: Forecasting visual style in fashion. *arXiv:1705.06394*, 2017.

[6] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, June 2018.

[8] Charles Corbiere, Hedi Ben-Younes, Alexandre Ramé, and Charles Ollion. Leveraging weakly annotated data for fashion image retrieval and label prediction. *arXiv:1709.09426*, 2017.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014.

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

[11] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *CVPR*, pages 1463–1471, 2017.

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pages 6626–6637, 2017.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[14] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized convolutional neural network. In *ICCV*, pages 1386–1394, 2015.

[15] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, pages 1096–1104, 2016.

[16] Varun Manjunatha, Mohit Iyyer, Jordan Boyd-Graber, and Larry Davis. Learning to color from language. *arXiv preprint arXiv:1804.06026*, 2018.

[17] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[18] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

[19] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*, 2015.

[20] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv:1605.05396*, 2016.

[21] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, and C. Pal. Fashion-Gen: The Generative Fashion Dataset and Challenge. *ArXiv e-prints*, June 2018.

[22] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, pages 2234–2242, 2016.

[23] Hosnieh Sattar, Gerard Pons-Moll, and Mario Fritz. Fashion is taking shape: Understanding clothing preference based on body shape from online sources. In *WACV*, pages 968–977. IEEE, 2019.

[24] Edgar Simo-Serra and Hiroshi Ishikawa. Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In *CVPR*, pages 298–307, 2016.

[25] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[26] Sirion Vittayakorn, Takayuki Umeda, Kazuhiko Murasaki, Kyoko Sudo, Takayuki Okatani, and Kota Yamaguchi. Automatic attribute discovery with neural activations. In *ECCV*, pages 252–268. Springer, 2016.

[27] Wenqi Xian, Patsorn Sangkloy, Varun Agrawal, Amit Raj, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Texturegan: Controlling deep image synthesis with texture patches. In *CVPR*, pages 8456–8465, 2018.

[28] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018.

[29] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, pages 1316–1324, 2018.

[30] Kota Yamaguchi, M Hadi Kiapour, and Tamara L Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, pages 3519–3526, 2013.

[31] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1710.10916*, 2017.

[32] Shizhan Zhu, Sanja Fidler, Raquel Urtasun, Dahua Lin, and Chen Change Loy. Be your own prada: Fashion synthesis with structural coherence. In *ICCV*, 2017.