ON CONDITIONING GANS TO HIERARCHICAL ONTOLOGIES *

A PREPRINT

Hamid Eghbal-zadeh†
LIT AI Lab & Institute of Computational Perception
Johannes Kepler University Linz, Austria
hamid.eghbal-zadeh@jku.at

Lukas Fischer†
Software Competence Center Hagenberg GmbH (SCCH)
Softwarepark 21, 4232 Hagenberg, Austria
lukas.fischer@scch.at

Thomas Hoch†
Software Competence Center Hagenberg GmbH (SCCH)
Softwarepark 21, 4232 Hagenberg, Austria
thomas.hoch@scch.at

May 17, 2019

ABSTRACT

The recent success of Generative Adversarial Networks (GAN) is a result of their ability to generate high quality images from a latent vector space. An important application is the generation of images from a text description, where the text description is encoded and further used in the conditioning of the generated image. Thus the generative network has to additionally learn a mapping from the text latent vector space to a highly complex and multi-modal image data distribution, which makes the training of such models challenging. To handle the complexities of fashion image and meta data, we propose Ontology Generative Adversarial Networks (O-GANs) for fashion image synthesis that is conditioned on an hierarchical fashion ontology in order to improve the image generation fidelity. We show that the incorporation of the ontology leads to better image quality as measured by Fréchet Inception Distance and Inception Score. Additionally, we show that the O-GAN achieves better conditioning results evaluated by implicit similarity between the text and the generated image.

***K*eywords** Generative Adversarial Networks · Text-to-image synthesis · Ontology-driven deep learning.

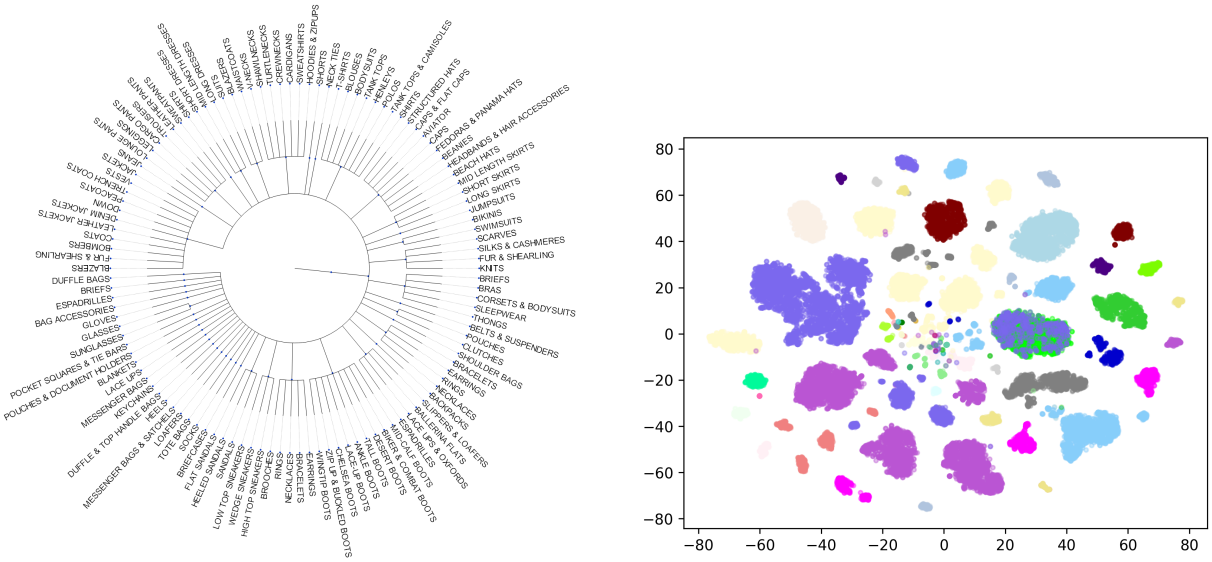
1 Introduction

Text-to-image synthesis is a challenging task where the details about the respective images are provided in a text corpus. The visual image details should best fit to the explanation provided in the text description, while maintaining a high-level of image detail fidelity. Generative adversarial networks (GANs) have proven to be a very powerful method to tackle this task [3, 6]. In [7] it has been shown that hierarchical model training by means of an ontology helps to learn more discriminant high-level features for fashion image representations. We adopt their strategy for generative models and show in this paper how a two layer fashion category ontology can be leveraged to improve GANs training for fashion image generation from text. The recently organized Fashion-Gen challenge [1] [9] provides a perfect test bed for the evaluation of novel methods for text-to-image synthesis. The provided dataset consists of 293.008 images, with 48 main and 132 fine-grained categories as well as a detailed description text. An ontology of sub-categories is visualized in Figure 1a. To handle the complexities of the Fashion-Gen challenge, we propose Ontology Generative Adversarial Networks (O-GANs) for high-resolution text-conditional fashion image synthesis. We detail the O-GAN in the following section.

*This work was supported by Austrian Ministry for Transport, Innovation and Technology, the Ministry of Science, Research and Economy, and the Province of Upper Austria in the frame of the COMET center SCCH.

†Equal contribution.

[1] https://fashion-gen.com/



(a) (b)

Figure 1: Fashion-Gen dataset: a) Ontology of sub-categories. b) Sub-category glovector embedding.

2 Ontology Generative Adversarial Networks

Recently a new training methodology for GANs, namely progressive growing of GANs (PGAN [6]) was proposed to improve the variation, stability and quality for image generation. We introduce a modified and improved version of PGAN to cope with the challenges of high-resolution text-conditional image synthesis, called Ontology Generative Adversarial Networks (O-GANs). A block-diagram of our proposed method is provided in Figure 2. We incorporate a progressive training based on PGAN and use a Wasserstein objective with gradient penalty [3] during training. For modelling the text, we use a word-level vector representation known as glovectors [8]. An embedding of the created sub-category glovectors can be found in Figure 1b. The utilized Discriminator and Generator are described in the following paragraphs:

Our model consists of 4 modules: 1) the Generator network G to generate images, 2) the Discriminator network D to distinguish between real and fake images, 3) the label predictor network L to classify images and 4) the Regressor R that regresses an image to its text embedding. The discriminator D, label predictor L and the Regressor R share most of their model parameters and only have different output layers.

The Discriminator D was constructed having three outputs each minimizing one of three respective objectives. This setting allows for parameter sharing between the D, L and R networks used for different objectives. The first objective, is to minimize the Wasserstein distance between real and fake (generated) images in the D network. The second objective, is a *classification* objective that minimizes the categorical cross entropy between labels (sub-categories) provided with the fashion images and the classification output in the L network. Finally the third objective, is a *regression* objective that minimizes an L_2 loss between the regression output of the R network and the average of the glovectors of the words in the description provided with the image.

The Generator G uses three concatenated vectors as the input: first the average glovectors of the text, second the labels as one-hot encoding and third a uniform random noise. During generation time, a bidirectional LSTM [5] trained on the sequences of glovectors is used to predict the labels from text.

3 Evaluation Measures

We use three evaluation measures to compare the proposed method with the PGAN baseline. We evaluate the quality of the generated images from different models based on the Fréchet Inception Distance (FID) [4] and Inception Score (IS) [10] (computed on 1k generated and 1k real images). The Inception model was trained on ImageNet [2].

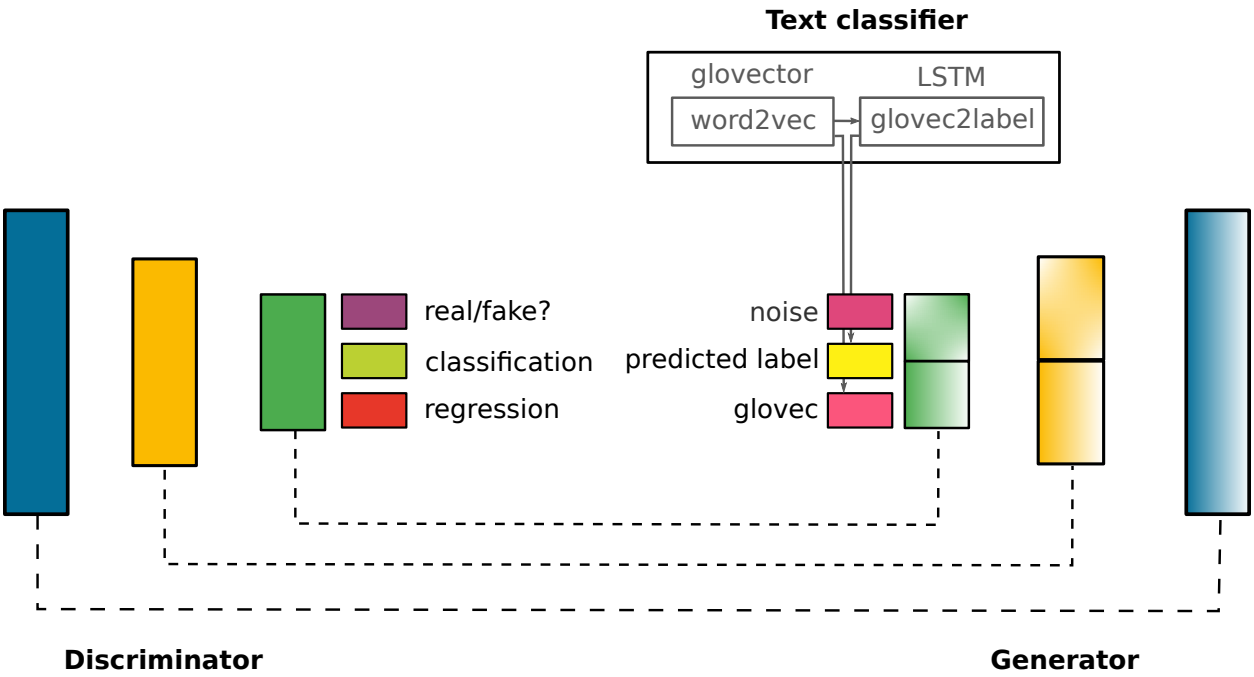


Figure 2: Diagram of the proposed O-GAN.

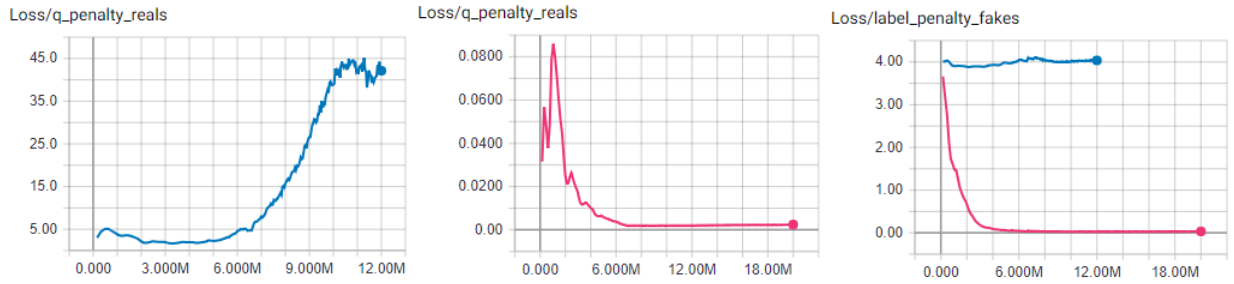
method \measure	IS (1k)↑	FID (1k)↓
Real Images	5.03 ± 0.36	0
Vanilla PGAN	4.54 ± 0.28	33.80
Proposed O-GAN	$\mathbf{4.81 \pm 0.61}$	**31.14**

Table 1: Image quality evaluation results. For Inception Score (IS) higher values and for Fréchet Inception Distance (FID) lower values are better.

To evaluate the quality of the conditioning, we report the cross-entropy between the conditioning labels and the probability of the labels for generated images, estimated via the label predictor L. We train the label predictor L to minimize the cross-entropy between the labels and the generated images through the probabilities produced in its output. In addition, we report the L_2 distance between the glovector extracted from the conditioning text, and an estimated glovector via the Regressor model R. The Regressor model R also shares its weights with the discriminator, but has a separate output layer trained to minimize the L_2 distance between images and their corresponding glovector. Hence, the Regressor model R implicitly estimates the similarity between the generated image and the glovector used for the conditioning [1].

4 Results and Discussion

In Table 1, we provide evaluation metrics on the Fashion-Gen dataset. Generated images of the implemented O-GAN are shown in Figure 4, where 4a depicts random generated examples and 4b-4c show generated images with their corresponding textual description. We used a modified PGAN as a baseline, in which a label predictor is added in addition to PGAN's discriminator. To demonstrate the importance of the ontologies, we only trained this baseline with category labels which has no information about the ontology of the finer-level classes. Additionally the generator was modified to use the label as conditioning vector. As shown in Figure 3, we can see that the O-GAN achieves lower L_2 distance to the conditioning glovector, compared to PGAN. This suggests that the images generated by O-GAN have higher similarity to the conditioning text compared to PGAN, as estimated by the Regressor model R. It can also be seen that a lower cross-entropy between the labels and the output of the label Regressor R can be achieved which demonstrates better ability in label conditioning in O-GAN compared to PGAN. As can be seen in Table 1 and Figure 3, the proposed method outperforms the PGAN in all cases of the reported evaluations. Training and evaluation was done on a NVidia DGX Station with 4 Tesla V100 GPUs using the Tensorflow framework.



(a) (b) (c)

Figure 3: Comparison of discriminator losses for the modified PGAN (blue) and the proposed O-GAN (magenta). a) L_2 distance between the condition glovector and the estimated glovector from images in PGAN by the Regressor model R in different epochs. b) L_2 distance between the condition glovector and the estimated glovector from images in O-GAN by the Regressor model R in different epochs. c) Cross-Entropy between label condition and the probabilities of the label predictor L from the generated images in PGAN and O-GAN in different epochs.

5 Conclusion

In this paper, we proposed a novel Text-Conditional GAN, the O-GAN, capable of generating realistic high-resolution images from a text describing the characteristic of the target image. We showed that our GAN can be used for generating fashion images from text description provided by the experts in fashion industry. We demonstrated the ability of O-GAN in incorporating ontologies in the generative process and showed how it improves the performance of both conditioning and quality of the generated images. We also showed that the proposed model outperforms the PGAN model in terms of text-conditioning evaluation measures.

References

[1] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[3] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *arXiv preprint arXiv:1706.08500*, 12(1), 2017.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *ICLR*, 2018. arXiv: 1710.10196.

[7] Jun Yu Zhou Yu Kuang, Zhenzhong and Jianping Fan. Ontology-driven hierarchical deep learning for fashion recognition. *In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2018.

[8] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[9] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-Gen: The Generative Fashion Dataset and Challenge. *arXiv:1806.08317 [cs, stat]*, June 2018. arXiv: 1806.08317.

[10] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 2234–2242, USA, 2016. Curran Associates Inc.

Figure 4: a) Random generated example images of our model. b-c) Generated images from our model and their conditioned text.