Week14 Pre-Report: Overview of Conditional Generative Adversarial Nets and Image-to-Image Translation with Conditional Adversarial Networks

Euijin Hong[1]
[1]Department of Electrical and Electronic Engineering, Yonsei University.

1 Introduction

Generative adversarial nets (GAN) [8] were introduced as an original way to train generative models.

The first paper [22] proposes a conditional version of GAN. The model can simply be constructed by feeding the data y, which is what we want to condition on to both the generator and the discriminator. The paper also shows the MNIST digits generated by this model and the usability to multi-modal learning, while giving an example of its application in image tagging where the model generates descriptive tags which are not provided as training labels.

Based on the proposal of the previous paper, the second paper [13] shows conditional GAN being used as a general-purpose solution to image-to-image translation problems. The network can learn a loss function to train the mapping from input image to output image, which is more than just learning the mapping itself. Thus, the same generic approach can be applied to obtaining very different loss functions. Some of those tasks are synthesizing images from label maps, reconstructing objects from edge maps, and adding colors to images. Furthermore, the paper suggests that engineers no longer have to handcraft the mapping functions and obtain reasonable results.

2 Conditional Generative Adversarial Nets

2.1 Theory

Previously, we have seen generative adversarial nets comprised of two non-linear mapping functions (e.g. multi-layered perceptrons): a generative model (G) which captures the data distribution, and a discriminative model (D) which estimates the probability that the sample came from the training set rather than G.

Models G and D can be trained simultaneously, whereas the models are playing a minmax game with value function $V(G,D)$: adjusting parameters of G for minimizing the loss function $log(1-D(G(z)))$ and adjusting parameters of D for minimizing the loss $log(D(x))$, as follows

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

Conditional Adversarial Nets: The basic structure of generative and discriminative model in conditional GAN can be illustrated as in Figure 1.

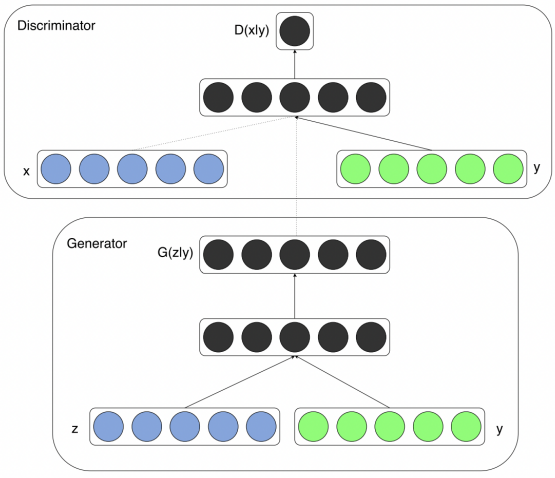

Figure 1: Conditional adversarial net

An expansion of GANs as a conditional model is possible if the generator and discriminator are both conditioned on some extra information y, where y is any kind of auxiliary information including class labels or data from other modalities. Conditioning can be performed by feeding y to both D and G as an additional input layer.

In the generator (G) side, the prior input noise $p_z(z)$ and y are combined by joint hidden representation. The adversarial training framework shows how this hidden representation is constructed, allowing for massive flexibility. In the discriminator (D) side, the input x and y are presented to a discriminative function (configured as a MLP). Hence the objective function of a two-player min-max game can be expressed as the following equation:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x}|\boldsymbol{y})] + \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z}|\boldsymbol{y})))].$$

2.2 Experimental Results of Conditional GAN

Unimodal: The MNIST images [1] were used for training a conditional adversarial net, where the labels are encoded as one-hot vectors.

The generator net was given a noise prior z with dimensionality of 100, which was drawn from a unifor distribution within the unit hypercube. z and y are both mapped to hidden layers with Rectified Linear Unit (ReLU) [6, 14] as an activation function, of their layers sizes as 200 and 1000 respectively, and then mapped into the second, combined hidden ReLU layer of size 1200. Then the final sigmoid unit layer generates the output of 784-dimensional MNIST samples.

The discriminator net maps x to a maxout [7] layer having 240 units and 5 pieces and y to a maxout layer with 50 units and 5 pieces. The hidden layers are both mapped in a joint maxout layer with 240 units and 4 pieces, then being fed to the sigmoid layer. Here, no certain architecture is prominent since the discriminator has sufficient power, although maxout units are typically well suited.

The conditional GAN approach showed comparable results with some of the network based results, while outperforming several other approaches, including plain GANs. The results are thus presented as a proof-of-concept, where showing that fine-tuning the model would exceed the non-conditional results.

The results of some generated samples are shown in Figure 2.

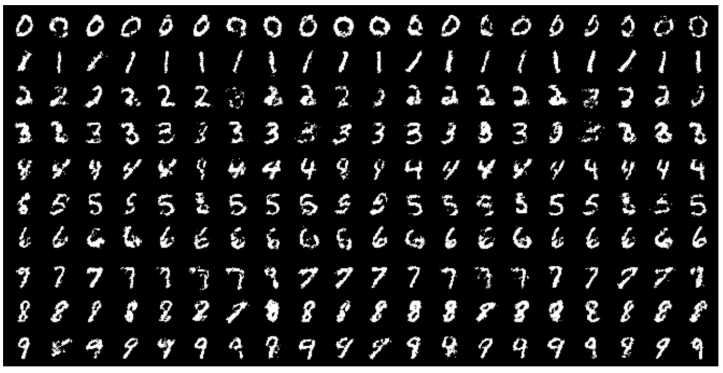

Figure 2: Generated MNIST digits, each row conditioned on one label

Multimodal: User-generated metadata (UGM) is more descriptive and aligned with how humans naturally describe images using language, as opposed to simply identifying objects. UGM also incorporates synonymous terms, requiring label normalization. Conceptual word embeddings [21] prove useful in representing related concepts through similar vectors.

To automate image tagging with multi-label predictions, researchers utilize conditional adversarial nets. These nets generate tag vectors based on image features. The approach involves pre-training a convolutional model on the ImageNet dataset and using a skip-gram model for text representation. They conduct experiments using the MIR Flickr 25,000 dataset [11], omitting images without tags and treating annotations as additional tags. The best model's generator maps noise prior and image features to ReLU

This is an extended abstract. The full paper is available at the Conditional Generative Adversarial Nets, Image-to-Image Translation with Conditional Adversarial Networks.

hidden layer, in order to generate word vectors. The discriminator distinguishes between real and generated tags. The approach involves cross-validation, random grid search, and manual selection to determine hyperparameters and architecture choices.

2.3 Future Work

The preliminary results in this paper highlight the potential of conditional adversarial networks and their promising applications. Future work includes presenting more advanced models, conducting a detailed analysis of their performance, exploring the use of multiple tags simultaneously, and developing a joint training scheme for the language model.

3 Image-to-Image Translation with Conditional Adversarial Networks

3.1 Objective

As shown previously, conditional GANs map an observed image x and random noise vector z to y; such as $G : x, z \rightarrow y$. The generator G learns how to produce outputs that are difficult to distinguish from "real" images, while the discriminator learns to detect the generator-derived "fake" images and works adversarially.

While previous approaches have discovered that mixing the objective function with a traditional loss, such as L2 distance, can be more beneficial [23], this paper utilizes L1 distance loss since L1 distance encourages less blurring.

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1].$$

Thus, the final objective of the model becomes:

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1}(G).$$

The absence of variable z in the network leads to deterministic outputs, limiting its ability to represent diverse distributions. Previous approaches incorporating Gaussian noise [26] was proved ineffective as the generator learned to ignore it in the early experiments [20]. Instead, the final models used dropout noise applied to multiple layers during training and testing, resulting in minimal stochasticity. Designing conditional GANs that produce highly stochastic output remains an open question.

3.2 Network Architecture

Researchers modified the generator and discriminator architectures based on those used in a previous work [24]. Both the generator and discriminator employ modules consisting of convolution-BatchNorm-ReLu [12].

Generator with skips: Image-to-image translation problems involve mapping a high-resolution input grid to a high-resolution output grid. The input and output, while differing in surface appearance, are renderings of the same underlying structure, making their structures roughly aligned. The generator architecture is designed while considering these factors.

Previous approaches [15, 23, 26, 27, 28] to similar problems have often utilized encoder-decoder networks[10], where information flows through all layers, including a bottleneck layer where the flowing process is reversed. However, many image translation problems involve shared low-level information between the input and output, such as the location of prominent edges in image colorization. To address this, researchers incorporate skip connections as in Figure 3, inspired by the "U-Net" structure [25]. These connections link each layer i to layer $n-i$, with the channels at each layer being concatenated. This allows the generator to bypass the bottleneck and preserve important information.

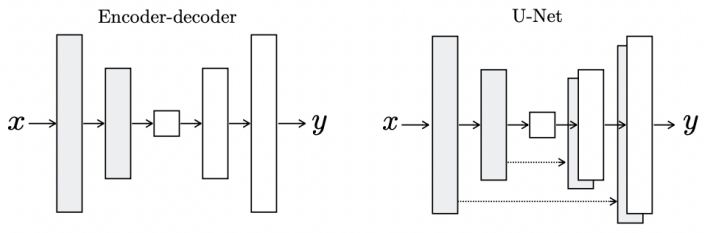

Figure 3: The "U-Net" is an encoder-decoder with skip connections between mirrored layers in the encoder and decoder stacks.

Markovian discriminator (PatchGAN): L2 and L1 losses in image generation produce blurry results [17] as in Figure 4. While they capture low frequencies well, they lack high-frequency crispness. Hence, correctness in low frequency is already guaranteed by L1 loss. To address high-frequency crispness problem, the researchers propose a *Patch*GAN discriminator that focuses on high-frequency structure within local image patches. This discriminator classifies $N \times N$ patches as real or fake, allowing for fewer parameters, efficient computation, and application to large images. The PatchGAN can be understood as a texture [2, 4] and style loss [3, 5, 9, 18], modeling the image as a Markov random field [19].

Figure 4: Various loss functions lead to varying levels of result quality. Each column in the display showcases outcomes achieved with different loss functions during training.

Optimization and inference: During optimization, researchers followed the standard approach of alternating gradient descent steps between the discriminator (D) and generator (G) [8]. We maximize log likelihood for $D(x, G(x,z))$ instead of minimizing $log(1 - D(x, G(x,z)))$ for G training. The objective for D is divided by 2, in order to slow its learning rate compared to G. We use minibatch SGD with the Adam solver [16], employing a learning rate of 0.0002 and momentum parameters $\beta 1 = 0.5$, $\beta 2 = 0.999$.

During inference, the generator is used similarly to training, but with dropout applied and instance normalization using test batch statistics [12]. The batch size varies between 1 and 10 in the experiments.

3.3 Experiments

To test the versatility of conditional GANs, various tasks and datasets were explored. The tasks included graphics tasks like photo generation and vision tasks like semantic segmentation. The experiments covered a range of datasets and tasks, such as semantic labels to photos, architectural labels to photos, map to aerial photo translation, black and white to color photo conversion, edges to photo generation, sketch to photo translation, day to night conversion, thermal to color photo conversion, and photo inpainting. The training sets varied in size, with some achieving good results even with small datasets. The evaluation of synthesized images involved perceptual studies and the use of recognition systems to assess their realism and interpretability. Ablation studies were conducted to analyze the impact of different components and variations in the objective function, generator architecture, and discriminator receptive fields. The results showed that conditional GANs produce sharp images and can improve colorfulness. The U-Net architecture and patch-based discriminators were found to be effective. The study also demonstrated the successful application of conditional GANs in generating labels for vision tasks, although reconstruction losses like L1 were generally sufficient for such problems. Furthermore, the paper highlighted the community-driven research and creative projects that have expanded the application of the pix2pix framework.

3.4 Conclusion

The findings presented in this study indicate that conditional adversarial networks show great potential for a range of image-to-image translation tasks, particularly those that require generating well-structured graphical outputs. These networks have the ability to learn a customized loss function specific to the task and data, allowing them to be applied in diverse scenarios.

[1] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

[2] AA Efros and TK Leung. Texture synthesis by non-parametric sampling ieee international conference on computer vision (iccv'99). *Corfu, Greece, September*, 17, 1999.

[3] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346, 2001.

[4] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28, 2015.

[5] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[6] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[7] Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. In *International conference on machine learning*, pages 1319–1327. PMLR, 2013.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets ian. *Mining of Massive Datasets; Cambridge University Press: Cambridge, UK*, 2014.

[9] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, and Brian Curless. Dh, salesin. *Image analogies. SIGGRAPH*, 2001.

[10] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[11] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43, 2008.

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[14] Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th international conference on computer vision*, pages 2146–2153. IEEE, 2009.

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR, 2016.

[18] Chuan Li and Michael Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2479–2486, 2016.

[19] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 702–716. Springer, 2016.

[20] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[21] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

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

[23] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

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

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[26] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 318–335. Springer, 2016.

[27] Donggeun Yoo, Namil Kim, Sunggyun Park, Anthony S Paek, and In So Kweon. Pixel-level domain transfer. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 517–532. Springer, 2016.

[28] Yipin Zhou and Tamara L Berg. Learning temporal transformations from time-lapse videos. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 262–277. Springer, 2016.