Introduction

This essay will review thee papers which present improvements to already existing Generative models. Generative models are models that can generate realistic data and assign likelihoods to data. The characteristics of such models enable us to utilize them for tasks such as image and text synthesis but are also useful for feature extraction of our input-data by mapping input to a learnt latent space [Kingma and Dhariwal, 2018]. A real-world application where generative models can be useful is in the entertainment/video-game industry. Recently revealed next-gen gaming console PS5 show improved disk read speeds [Leadbetter, 2020] which make them more suitable for high-res textures/models compared to before. Video-games that require large crowds, such as football game Fifa, can therefore make use of generated realistic faces for the audience with the features/emotions that are required for a specific scenario, for example cheering or crying audience. This would save a lot of time for the developers since they can generate data instead of having to spend time collecting it, and it would also elevate the gaming experience by making it more realistic.

Previous generative models have shown room for improvements which is why it is worth trying to improve them. Some of the previous successful generative models include Variational Autoencoders (VAEs), Flow-based generative models, and Generative adversarial networks (GANs). The three selected papers for review each present improvements for a model associated with one of the three mentioned model types. The first paper, "Glow: Generative Flow with Invertible 1×1 Convolutions", [Kingma and Dhariwal, 2018] propose Glow, which is an improved model to previous flow-based models. [Karras et al., 2020] is the second selected paper. They present StyleGAN2, which includes improvements to StyleGAN [Karras et al., 2018], which is based on the GAN class. The final paper I will review is "NVAE: A Deep Hierarchical Variational Autoencoder" [Vahdat and Kautz, 2020], where they present a hierarchical VAE called Nouveau VAE (NVAE) which is an extension to the work on hierarchical VAEs presented in [Kingma et al., 2016].

Method

A common goal of the selected papers is to improve the quantitative and qualitative performance of pre-existing generative models. In this section, I will summarize their approaches. An overview of some of the common approaches among the papers make can be seen in table 1. This table also presents some additional features of the models and what the authors aim for.

Glow [Kingma and Dhariwal, 2018] is a model built on flows, which in this paper is a series of invertible transformations from the data used to a latent space. It is likelihood-based and is trained on image data using SGD. [Kingma and Dhariwal, 2018] work on a multi-level architecture that forms a sequential series of multiple flow steps. The invertibility property of each step in the network makes the whole model invertible. These steps consist of an actnorm (activation normalization) layer, 1x1 convolution and a coupling layer.

The actnorm layer is introduced as a replacement to batch normalization (BN), used in previous work, due to the small size of the minibatches when the authors train Glow. They train with 1 data point per GPU in each minibatch due to the memory requirements imposed by large resolution images. Their proposed actnorm layer transform the activation with trainable, channel-wise, mean and variances. The invertible 1x1 convolution layer is introduced to replace channel reversing permutations. The final coupling layer of each step is used similarly to previous work. The number of levels and depth level (number of steps each level) varies in their experiments but the levels range from 3 to 6 and the depth ranges from 32 to 64 [Kingma and Dhariwal, 2018].

[Karras et al., 2020] propose StyleGAN2 by analyzing and improving StyleGAN [Karras et al., 2018]. They iteratively add changes to the original StyleGAN and show how the performance is improved. The first major architectural change they make is to alter the feature map normalization by replacing it with modulation and demodulation operations.

They furthermore introduce path length regularization, which enforces the generator to be trained such that stepping a fixed length in the latent space results in a fixed change in the data-space, independent of the direction of the change. To reduce computational costs, they apply R_1 [Mescheder et al., 2018] and path length regularization in a lazy fashion by only applying them every 16th and 8th training iteration for the discriminator respective generator. [Karras et al., 2020] also introduce residual nets to the discriminator and skip-connections to the generator. At last, they increase the model size by doubling the $64^2 - 1024^2$ res feature maps, since they discovered that the previous setup limited their network for high-res data.

The authors of NVAE [Vahdat and Kautz, 2020] aims at improving VAEs since they're outperformed by other generative models but still have unique characteristics. NVAE uses a deep hierarchal VAE with latent variables $\boldsymbol{z} = \{\boldsymbol{z}_1, ..., \boldsymbol{z}_L\}$ where the prior of the latent space is $p(\boldsymbol{z}) = \prod_{l=1}^{L} p(\boldsymbol{z}_l|\boldsymbol{z}_{<l})$. The hierarchical model that [Vahdat and Kautz, 2020] use samples from $\boldsymbol{z}_1$ first and then conditionally sample the next $\boldsymbol{z}_l$ until $\boldsymbol{z}_L$ is reached. A new data point can be generated when the whole $\boldsymbol{z}$ has been sampled. Each time a new $\boldsymbol{z}_l$ is sampled, they double the dimensions, starting from a relatively small dimensional $\boldsymbol{z}_1$, to combat the difficulties of learning spatially long-range correlations in high-res images. They furthermore model residual cells that are inserted between every $\boldsymbol{z}_l$ level in the hierarchy. These cells have depth-wise convolutions for the generative model, but they omit these from the encoder cells since the convolutions add parameters to the network and should only be added if they boost performance. BN is introduced again (previous work on VAEs [Kingma et al., 2016] omitted BN) but modifies the momentum since the previously used momentum influenced the test phase negatively.

[Vahdat and Kautz, 2020] moreover parameterizes the distribution $q(\boldsymbol{z}|\boldsymbol{x})$ to make it easier to minimize the KL from $q(\boldsymbol{z}_l|\boldsymbol{x}, \boldsymbol{z}_{<l})$ to $q(\boldsymbol{z}_l|\boldsymbol{z}_{<l})$. In addition to this, they add a regularizing term to the lower bound to ensure that the KL is bounded. Lastly, the authors incorporate normalizing flows to the encoder part of the model.

This summarizes some of the major changes each method used, see table 1 for some additional features and goals that the authors present in their papers.

Table 1: Features of the models and the changes that the authors made to these models in relation to previous work.

Features, changes or goals\Model	Glow	StyleGAN2	NVAE
Likelihood-based	✓		✓
Modifies narmalization operations	✓	✓	✓
Explicitly features Flow components	✓		✓
Aims for higher resolution	✓	✓	✓
Takes memory requirements into consideration	✓	✓	✓
Max image resolution generated	256^2	1024^2	256^2

Comparison

The major difference between StyleGAN2 and the two other models is that Glow and NVAE are likelihood-based while StyleGAN2 is not. [Vahdat and Kautz, 2020] focus their comparison to other likelihood-based methods, but is still interesting to compare NVAE and Glow to StyleGAN2 since not all all real-world applications require likelihood-based methods.

[Vahdat and Kautz, 2020] compare the performance of NVAE to Glow and NVAE shows better performance when measuring in bits/dimension on CIFAR-10, ImageNet and the CelebA HQ dataset. The qualitative performance of NVAE on CelebA is also better than those generated by Glow in my opinion. In figure 1, I have tried to handpick representative generated images from the three models and, as seen, the leftmost image (generated by Glow) looks more artificial than the one furthest to the right (generated by NVAE). The middle image in figure 1 shows how much better/realistic the images generated by StyleGAN2 are compared to the other two. The middle image is more detailed, especially in the hair and skin textures. The different face components also fit better in StyleGAN2 generated images as well and are furthermore generating images with 16 times higher resolution. Despite the larger resolution, StyleGAN2 still manages to make spatially far away details fit well together. Note that the middle image is generated by a model trained on FFHQ data, instead of CelebA HQ like the two others. I would refrain from using pictures of celebrities when training generative models with the goal of trying to achieve realistic images of humans. Celebrities often use makeup when photographed at public settings which increases the bias for smooth skin for example.

A common approach that the authors make when trying to improve their generative models is to alter normalization operations. Glow and NVAE modify the use of BN compared to previous work and the authors of StyleGAN2 identify problems with and alter the feature map normalization operation. This shows that revisiting normalization operations in a model may be a good idea if one tries to improve the performance of a generative model.

A common obstacle for the authors is memory limitations. They all mention that they in some way have done something to combat this obstacle which shows how memory intensive generative models are. One should consider this when working with generative models so that it doesn't unexpectedly bottleneck the progress.

Going back to the potential real-world application of using generated faces in the crowd of video-games, Fifa for example, all three models generate images of sufficient resolution for background characters that are small enough that 2D faces suffices. The smudgy, and sometimes inconsistent, backgrounds that we see on generated images of faces do not matter much, since they can be cropped out. An important aspect to consider when incorporating crowds is their emotions. This puts a requirement of being able to interpolate between emotions while preserving the other features. [Kingma and Dhariwal, 2018] show that this is possible in Glow, but, while their interpolation is smooth, the expression that they interpolate to (smiling) is subtle, awkward, and looks non-authentic. StyleGAN2 interpolates smoothly as well and generates faces with authentic-looking emotions, as seen in the video the authors refer to on their Github repository. However, they do not show interpolation between emotions, while keeping the other features fixed. I believe interpolation between emotions is an important aspect of generative models used for face-image synthesis and is something that is useful in human-AI interaction. This could be an area of future study and a potential improvement area for all three models presented in this essay. Their capability of interpolating well between emotions could be the deciding factor in some real-world applications in the future.

Figure 1: Generated images taken from the featured papers. From left to right: Glow on CelebA HQ data, StyleGAN2 on FFHQ, NVAE on CelebA HQ.

References

[Karras et al., 2018] Karras, T., Laine, S., and Aila, T. (2018). A style-based generator architecture for generative adversarial networks. *arXiv*, pages arXiv–1812.

[Karras et al., 2020] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. *arXiv preprint arXiv:1912.04958v2*.

[Kingma and Dhariwal, 2018] Kingma, D. P. and Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*.

[Kingma et al., 2016] Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751.

[Leadbetter, 2020] Leadbetter, R. (2020). Inside playstation 5: the specs and the tech that deliver sony's next-gen vision. https://www.eurogamer.net/articles/digitalfoundry-2020-playstation-5-specs-and-tech-that-deliver-sonys-next-gen-vision.

[Mescheder et al., 2018] Mescheder, L., Geiger, A., and Nowozin, S. (2018). Which training methods for gans do actually converge? *arXiv preprint arXiv:1801.04406*.

[Vahdat and Kautz, 2020] Vahdat, A. and Kautz, J. (2020). Nvae: A deep hierarchical variational autoencoder. *arXiv preprint arXiv:2007.03898*.