Generative Adversarial Stacked Autoencoders

Ariel Ruiz-Garcia[*]
LatinX in AI and SeeChange.ai
Manchester, United Kingdom

Ibrahim Almakky,[†] **Vasile Palade,**[‡] **Luke Hicks**[§]
Coventry University
Coventry, United Kingdom

1 Introduction

Generative Adversarial Networks (GANs) [1], [2] have become predominant in image generation tasks. Their success is attributed to the training regime which employs two models: a generator G and discriminator D that compete in a minimax zero sum game. Nonetheless, GANs are difficult to train due to their sensitivity to hyperparameter and parameter initialisation, which often leads to vanishing gradients, non-convergence, or mode collapse, where the generator is unable to create samples with different variations.

In this work, we propose a novel Generative Adversarial Stacked Convolutional Autoencoder (GASCA) model and a generative adversarial gradual greedy layer-wise learning algorithm designed to train Adversarial Autoencoders in an efficient and incremental manner. Our training approach produces images with significantly lower reconstruction error than *vanilla* joint training.

2 Generative Adversarial Stacked Convolutional Autoencoders

Let x_φ be a sample from the data distribution $p_d(x_\varphi)$ and x_μ the sample from the data distribution $p_d(x_\mu)$ used as the desired target reconstruction. The autoencoder G model in a GASCA learns to map x_φ to a latent space z and back to a reconstruction y that resembles x_μ and lies in the distribution $q(y)$. The discriminator D attempts to differentiate between y and x_μ.

In the conventional adversarial autoencoder framework [2], a distribution $p(z)$ —often a Gaussian distribution—is imposed on $q(z)$ by estimating the divergence between q and p. This imposition can be used to produce reconstructions with specific features. However, in order to produce reconstructions that are as close as possible to the desired target image x_μ, instead of imposing random noise on the hidden representation vector, the GASCA model imposes $p_d(x_\mu)$ on $q(y)$ in the following manner: $q(y) = \mathbb{R}_{x_\mu} q(y|x_\mu) p_d(x_\varphi) dx_\mu$. With this formulation, the discriminator model D is optimised to rate samples from $p_d(x_\mu)$ with a higher probability, and samples from $q(y)$ with a low probability. Formally, this is defined as: $\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D(x_\mu^{(i)}) + \log \left(1 - D(G(y^{(i)}))\right) \right]$ where x_μ is an input image and x_φ is the target reconstruction image and m is the number of samples. Note that since $(x_\varphi = x_\mu)$ is not necessarily true, the discriminator D is not guaranteed to see the input to G.

The objective of the autoencoder model G is to convince the discriminator model D that a sample reconstruction y was drawn from the data distribution $p_d(x_\mu)$ and not from $q(y)$. This optimisation is done according to: $\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log \left(1 - D(G(y^{(i)}))\right)$. Furthermore, since GANs are known to be difficult to train due to their sensitivity to hyper-parameters and parameters initialisation, which often leads to mode collapse, we propose training the GASCA model in a Greedy Layer Wise (GLW) [3] fashion. However, since the greedy nature of GLW leads to error accumulation as individual layers

---

[*]ariel.9arcia@gmail.com. Most of this work was done while at Coventry University.

[†]ibrahimalmakky@gmail.com

[‡]vasile.palade@convetry.ac.uk

[§]hicksl@convetry.ac.uk

are trained and stacked [4], we build on the gradual greedy layer-wise training algorithm from [4] and adapt it for adversarial autoencoders. We introduce the GAN gradual greedy layer-wise (GANGGLW) training framework and formally define it in Algorithm 1.

Algorithm 1 Given a training set X and validation set $\tilde{X}$ each containing input images x_φ and target images x_μ, m shallow autoencoders, an unsupervised feature learning algorithm $\mathcal{L}$ which returns a trained shallow autoencoder and a discriminator model, and a fine-tuning algorithm $\mathcal{T}$: train D^1 and G^1 jointly with raw data and add them to their corresponding stacks G and D. For the remaining autoencoders and generator models: encode X and $\tilde{X}$ using the encoder layers ξ from the stack G. Create a new discriminator D^k and train together with the new autoencoder G^k and add them to their corresponding stacks. Fine-tune G on raw pixel data. Forward propagate $x_\varphi \subset X$ through G and use the resulting features, along with x_μ, to fine-tune D for binary classification.

1: $[G^1, D^1] \leftarrow \mathcal{L}(G^1, D^1, X, \tilde{X})$
2: $G \leftarrow G \circ G^1$
3: $D \leftarrow D \circ D^1$
4: **for** $k = 2, \ldots, m$ **do**
5:     $[\xi, \delta] \leftarrow D$
6:     $[X_g, \tilde{X}_g] \leftarrow \xi(X, \tilde{X})$
7:     $[G^k, D^k] \leftarrow \mathcal{L}(G^k, D^k, X_d, \tilde{X}_d)$
8:     $G \leftarrow G^{(k)} \circ G$
9:     $D \leftarrow D^{(k)} \circ D$
10:     $G \leftarrow \mathcal{T}(G, X, \tilde{X})$
11:     $X_\varphi \leftarrow G(x_\varphi)$
12:     $D \leftarrow \mathcal{T}(D, \{X_\varphi, x_\mu \subset X\})$
13: **end for**
14: **return** G, D

By fine-tuning G and D in Algorithm 1 we avoid error accumulation from one layer to the next and reduce the required number of fine-tuning steps for deeper layers.

The key difference in GANGGLW as opposed to conventional GLW is that each time a new layer or shallow autoencoder is trained, the entire model is fine-tuned on raw data. This allows for stronger inter-layer connections and significantly decreased error accumulation. Moreover, both the discriminator and generator are trained in an incremental manner where layers are added one by one.

This formulation of GAN autoencoder also allows for the target reconstructions to be different than the input image as in conventional autoencoders. Figure 1 shows a sample reconstruction of facial images with different poses (up to 60 degrees), where the same image (at 0 degrees, top right) is used as target reconstruction. Our GANGGLW training approach has also proven to be effective when used a as a pre-training approach for supervised tasks such as classification, where it produces state-of-the-art accuracy. An extended version of this work has been published in [5].

Figure 1: Sample reconstructions produced by a GASCA model on face images from [6].

List of References

[1] J. Schmidhuber, "Generative adversarial networks are special cases of artificial curiosity (1990) and also closely related to predictability minimization (1991)," *Neural Networks*, vol. 127, pp. 58 – 66, 2020.

[2] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial Autoencoders," 2015.

[3] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, (Cambridge, MA, USA), p. 153–160, MIT Press, 2006.

[4] A. Ruiz-Garcia, V. Palade, M. Elshaw, and I. Almakky, "Deep Learning for Illumination Invariant Facial Expression Recognition," in *Proceedings of the International Joint Conference on Neural Networks*, (Rio de Janeiro), IEEE, 2018.

[5] A. Ruiz-Garcia, V. Palade, M. Elshaw, and M. Awad, "Generative adversarial stacked autoencoders for facial pose normalization and emotion recognition," in *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2020.

[6] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," in *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pp. 1–8, 2008.