The idea of Generative Adversarial Networks (GAN) was first proposed by Goodfellow et al[1] in 2014. Their key contribution to the field of generative models was the idea of simultaneously training two networks, a generator (G) and a discriminator (D) pitted against each other in a minimax two player game. G is tasked with trying to generate outputs that would fool D into falsely classifying it's output as real, while D tries to optimise it's performance at discriminating output from G against real images.

Initially, the image quality of the output was not significantly better than alternative approaches and were low resolution. GANs were initially very difficult to train, being very sensitive to hyperparameters and training instability. Training instability causes two problems, low image quality and mode collapse – in which the model captures one or a few larger modes and produces images of low diversity. Additionally, GANs suffered from vanishing graidients, where G is not given enough information to improve. The main trends since have been to try to improve the architecture, loss and objective functions and regularisation of the networks.

The original GAN generated the image in a single step, at the final output resolution, which was very low. LAPGAN[2] introduced the idea of using a Laplacian Pyramid of several GANs, starting at a low resolution with each level progressively upscaling the image. As each step upscaled the image and filled in detail, it was able to produce substantially higher quality images at a much higher resolution, being mistaken for real images 40% of the time by human evaluators, compared to just 10% for the original GAN. This, however, compares poorly to the 90% rate for real images, suggesting the quality was still very low.

This was followed by PROGAN[3], for which the key idea was to train a single network first at low resolutions, and progressively increase the resolution during training. The network faced simpler learning tasks and was significantly faster to train, much more stable and produced images of much higher quality (with a record 8.8 inception score). As early training was with low resolution images, it was better able to discover and represent the large scale structure before adding in finer detail at much higher resolutions, although still limited.

Another key trend was the move away from fully connected layers towards. convolutional layers that had shown much promise in other areas of computer vision. DCGAN[4] was the first network to successfully apply deconvolutional architectures to GANs. This, in addition to their usage of Batch Normalisation, made the network much more stable. Their usage of these deconvolutional layers made it possible to visualise the filters learnt by the network for the first time, enabling researchers to understand which parts of the network had learnt to draw specific objects. Many subsequent models adapted and improved their deconvolutional architecture, with significant performance gains.

However, one problem with convolutional GANs is that they failed to accurately capture structural or geometric patterns, performing better on images with textural detail. Zhang et al[5] hypothesised that this was because they failed to capture long range dependences across different regions of the image. Attention and self-attention models had shown ability to use longer range and global dependencies in other fields. With their SAGAN network, they showed that self attention was able to significantly improve image generation quality (with a record inception score of 52.52). They showed that this resulted in images with clearly defined larger structures like a dogs legs. However, the resultant images often had jarring disunities and were not particularly convincing.

This was improved on by Daras et al[6] with Your Local GAN (YLG) which contributed a new local sparse attention layer, introduced ideas from Information Flow Graphs and made the inversion process of gradient descent work on bigger models with Enumerate, Shift, Apply (ESA). They reported significant improvements in FID, inception score and human judgement, as well as a 40% reduction in training time from their architecture over SAGAN.

Gong et al[7] advanced the field by taking the Neural Architecture Search (NAS) algorithm that has shown success in image classification and used it to automatically discover a network architecture. Their AutoGAN achieved performance comparable to other state of the art handcrafted GANs. Furthermore, the generated architecture was significantly more efficient than other models with comparative performance (2.16 GFLOPs to PROGAN's 6.39 GFLOPs). This suggests that further work on automatic approaches to discovering

architecture may produce models that have significant performance or speed improvements – making it easier to develop, train and deploy models. While making it easier to explore architectures for more efficient and powerful networks is useful in itself, the work raises a whole new class of unsolved problems in efficiently finding these networks.

Another trend in architectural improvements is work on the ability of the network to produce images in a large number of output categories. Conditional GAN[8] (CGAN) introduced the idea of feeding the class label to the generator and discriminator, resulting in a network whose data generation process can be directed, opening up the ability to generate images of a different classes at will, which wasn't previously possible. This idea was built on by InfoGAN[9] and AC-GAN[10]. However, one common problem to these approaches is the reliance on well labelled images.

Self-Supervised GAN[11] (SS-GAN) followed on from other advances in representation learning and was inspired by previous conditional models. The key idea was to train the network on an image rotation task and transfer the image representations to image classification - allowing the generator and discriminator to adopt a collaborative approach to the rotation task while maintaining an adversarial approach to image generation. This approach performed well and addressed the problem of catastrophic forgetting that was a problem for GAN instability.

A recent trend has been the application of masked language models to GANs in order to condition images on natural language captions. One notable contribution in X-LXMERT[12] improved on other models and was the to make it possible to generate images based on natural language input, allowing fine grained control of the desired image. While the images are high resolution, they are not convincing. Further work to improve the quality of the images, while allowing fine grained semantic control over the output is needed and this unsolved problem would be very interesting to work on.

Parallel to advances in architecture, there was a great deal of work on loss functions. The motivation for this was that an insufficient loss function would reward a model for generating "safe" samples but not for exploring ways of learning ways of generating more diverse images – leading to mode collapse.

One approach was the Wasserstein Loss[13], which used the Earth Mover (EM) distance and the fact that the EM distance is continuous and differentiable and thus not only were they were able to train D to optimality at each step, but it resulted in a more reliable gradient for G. Thus their network (WGAN) was much more stable and resistant to mode collapse, vanishing gradients and meant that G and D no longer had to be trained in near lock-step. Their network, however, did not converge easily and performed poorly with deeper models or with momentum based optimisers. This was improved on by WGAN-GP[14].

Che et al[15] introduced a mode regularizer and the idea of explicitly penalising missing modes and methods to measure the number of missing modes and showed their approach led to a large reduction. As they discussed, this is important to produce more diverse images and prevent mode collapse, showing that their model was better able to represent rare classes of training data which is key for generating images from imbalanced data sets.

One persistent problem in the development of GANs is the difficulty of making objective functions that can match human judgement, making it hard to compare approaches. Initially this had been done by human judgement, which can be unreliable and time consuming. The Inception score[16] was introduced to make it easier to compare models. The key idea was to use a semi supervised model that would classify whether or not an image was in a "generated" class or in one of K other classes corresponding to supervised image classes. This was improved on by the Fréchet Inception Distance (FID) [17], which the authors showed was able to better score images disturbed by noise, blur, swirl and other contamination and more closely matching human evaluation.

As GANs have improved, measuring improvements has become increasingly difficult and images that are still visually unconvincing can acquire high scores. Making an objective function that can effectively match human level judgements is a challenging unsolved problem that I would be in working on.

Bibliography

1: , Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial networks. arXiv preprint arXiv:1406.2661.,
2: , Denton, E., Chintala, S., Szlam, A. and Fergus, R., 2015. Deep generative image models using a laplacian pyramid of adversarial networks. arXiv preprint arXiv:1506.05751.,
3: , Karras, T., Aila, T., Laine, S. and Lehtinen, J., 2017. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196.,
4: , Radford, A., Metz, L. and Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.,
5: , Zhang, H., Goodfellow, I., Metaxas, D. and Odena, A., 2019, May. Self-attention generative adversarial networks. In International conference on machine learning (pp. 7354-7363). PMLR.,
6: , Daras, G., Odena, A., Zhang, H. and Dimakis, A.G., 2020. Your local GAN: Designing two dimensional local attention mechanisms for generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14531-14539).,
7: , Gong, X., Chang, S., Jiang, Y. and Wang, Z., 2019. Autogan: Neural architecture search for generative adversarial networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3224-3234).,
8: , Mirza, M. and Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.,
9: , Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I. and Abbeel, P., 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. arXiv preprint arXiv:1606.03657.,
10: , Odena, A., Olah, C. and Shlens, J., 2017, July. Conditional image synthesis with auxiliary classifier gans. In International conference on machine learning (pp. 2642-2651). PMLR.,
11: , Chen, T., Zhai, X., Ritter, M., Lucic, M. and Houlsby, N., 2019. Self-supervised gans via auxiliary rotation loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12154-12163).,
12: , Cho, J., Lu, J., Schwenk, D., Hajishirzi, H. and Kembhavi, A., 2020. X-LXMERT: Paint, Caption and Answer Questions with Multi-Modal Transformers. arXiv preprint arXiv:2009.11278.,
13: , Arjovsky, M., Chintala, S. and Bottou, L., 2017, July. Wasserstein generative adversarial networks. In International conference on machine learning (pp. 214-223). PMLR.,
14: , Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. and Courville, A., 2017. Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028.,
15: , Che, T., Li, Y., Jacob, A.P., Bengio, Y. and Li, W., 2016. Mode regularized generative adversarial networks. arXiv preprint arXiv:1612.02136.,
16: , Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. and Chen, X., 2016. Improved techniques for training gans. arXiv preprint arXiv:1606.03498.,
17: , Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. and Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. arXiv preprint arXiv:1706.08500.,