Comic Art Generation using GANs

ir. Marnix Verduyn[1]
Thesis supervisor: prof. dr. ir. Luc De Raedt[1 2]
Mentor: ir. Thomas Winters[1]

[1] Department of Computer Science, KU Leuven, Belgium
[2] AASS, Örebro University, Sweden

There is a practical problem that often arises in the comic book industry. If a well-known author is no longer able or willing to continue working on his oeuvre, who will take up the torch? Would a machine be up to this challenge?

In practice, young talented authors will meticulously study and rehearse the drawing style until they can produce new work in the style of the old master. For this training, they rely on the existing work and that is often a limited number of comic books. Moreover, these are diverse stories with different characters in all kinds of settings and portrayed from a great many different camera angles. In the realm of AI, we then speak of a limited data set, which is also very diverse. This is not a problem for humans because we understand the semantics of drawings. Characters, objects, and backgrounds may be represented distorted in a caricatured style as is usual in comics, we all capture the meaning of the drawings. Content and form are separated in our brains. For machines, this is less obvious. Moreover, style imitation is only part of the real challenge. The moonshot project in this field, of course, is a machine that can generate comic strips autonomously, from start to finish.

This thesis describes an empirical study of the technical difficulties of comic art generation using GANs. As the dataset, I use the Kinky & Cosy series, an internationally published three-panel comic series that I created. Starting from unlabeled images, I perform a comparative test with three types of GANs: DCGAN, WGAN, and StyleGAN2-ADA.

Similary [Morris, 2021] and [Proven-Bessel et al., 2021] performed an experimental analysis of different GANs on single comic panels of the newspaper series Dilbert. Both use less recent GANs and a dataset with less diversity. The Dilbert series is characterized by a great degree of monotony in the images and, as such, is not representative of the average newspaper comic.

For the experiments, two datasets were constructed from 500 30-second animated cartoons. The cartoons were based on the newspaper strip and had the advantage that there were no speech bubbles in the picture. The images of the cartoons were first selected so that they were sufficiently different and then cropped left and right into squares. This produced an initial dataset of 56.234 very diverse images. A new dataset was then created from this set, by focusing on the main characters Kinky and Cosy. This second dataset consists of 8.543 images and is clearly less diverse, although still much more diverse than, for instance, a dataset of human faces.

The experiments show interesting differences at two levels. First, there is the clearly different performance of DCGAN, WGAN, and StyleGAN2-ADA. And second, the diversity of the dataset also has an important impact on the results. The former is a consequence of the phenomenal developments in GAN algorithms in recent years. And the latter illustrates the limitations GANs have for the problem at hand.

The conclusion of this study is that only StyleGAN2-ADA is eligible for style imitation and only after training on the dataset with the smallest diversity (see Figure 1). Under these conditions, the lowest FID score is obtained and the generated images received the best rating by a small test audience. There are limitations, of course. Although the style is well-captured, there are still imperfections in the image. This is best seen in walks in latent space. Between two nearly perfect images, a whole series of nonsensical images are often generated. The second dataset scores significantly better with respect to this limitation. A dataset that is denser and has fewer "holes" of missing information ultimately leads to fewer nonsensical generations. Zooming in on Kinky and Cosy resulted in almost error-free images. However, this meant that all the information about the background and side characters was not included during training. This clearly indicates an upper limit to the capabilities of current state-of-the-art GAN models on the problem at hand.

Another interesting discovery is the existence of semantically meaningful directions in our trained GANs. Random explorations of latent space delivered walks in which only meaningful changes occur, for example a character in a standing position who only turns his head. Furthermore, by projecting images into latent space, it was possible to generate new images of Kinky & Cosy that did not appear in the dataset. Although they were not flawless, this proved that the model had learned the style.

Future work will have to show whether these results are repeatable on other datasets. Conditional generation will also need to be looked at. The text needed for the labels can be extracted from the speech bubbles or the scenarios.

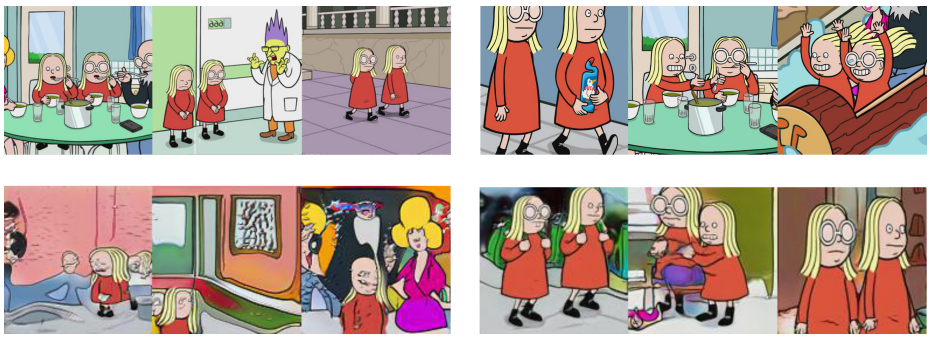

Fig. 1. StyleGAN2-ADA samples after training on two datasets. Top: reals, bottom: fakes. Left: dataset 1, right: dataset 2. © 2014 Nix/Ellipsanime Productions/Belvision