They're veGAN but they almost taste the same: generating simili-manuscripts with artificial intelligence

Camps, Jean-Baptiste
jean-baptiste.camps@chartes.psl.eu
École nationale des chartes | Université PSL, France

Vidal-Gorène, Chahan
chahan.vidal-gorene@chartes.psl.eu
École nationale des chartes | Université PSL, France

Introduction

The aim of our research is to artificially generate fake historical manuscripts using generative adversarial networks, better known as GANs (Goodfellow et al. 2014). In this neural network architecture, a generator network is put in competition with a discriminator (e.g., the first one generates images that seem authentic, and the second tries to guess if they are or not).

At this stage, these experiments pursue two objectives: evaluating feasibility of generating realistic fake manuscripts under certain conditions of layout, script, or date, and creating artificial data for HTR training, as is done for printed materials with synthetic data (Etter et al. 2019). Ground truth creation is indeed a time consuming task, in particular for ancient languages for which we still lack specialists able to manually annotate documents. GAN appears as a relevant answer to this challenge, as they reach very convincing results in different scenarii. State-of-the-art results rely on a style-transfer approach (Karras, Laine, and Aila 2018), in which a generator try to map on a targeted dataset unsupervised learned features from an other. Such an approach has already been successfully applied to historical manuscripts to create realistic Latin manuscripts (Vögtlin et al. 2021) with very constrained styles in training (page-level), or for contemporary cursive hand-writings (Fogel et al. 2020) (line-level with a semi-supervised approach).

This short paper investigates the feasibility of the line-level approach for historical manuscripts.

Datasets

We focus on two different manuscript traditions. The first dataset is composed of 8 Armenian manuscripts in a very regular *bolorgir* script (Stone, Kouymjian, and Lehmann 2002; Vidal-Gorène and Decours-Perez 2021), from the 14th to the 18th century. Dataset composition is described in table 1. The Old French dataset is composed of data extracted from the cremma-medieval dataset (Pinche 2022), from which 6324 lines of a 13th century manuscript in *textualis* Latin script were selected (ms. BnF, fr. 412).

Table 1. Composition of the Armenian dataset. The manuscript identifiers follow the ID system of the Index of Digitized Armenian Manuscripts (Vidal-Gorène, Sargsyan, and Van Elverdinghe 2022)

Manuscript*	Language	Date	Script	Lines
M1	Classical Armenian	1632-1633	bolorgir	512
M6	Classical Armenian	16th	bolorgir	819
M982	Classical Armenian	1460	bolorgir	3.035
MAF52	Classical Armenian	15th-16th	bolorgir	954
MAF54	Classical Armenian	16th	bolorgir	1.253
MAF62	Classical Armenian	18th	bolorgir	522
TBI122	Classical Armenian	17th	bolorgir	1.209
	8.304			

Data annotation has been made on eScriptorium (Kiessling et al. 2019) for Old French manuscripts and Calfa Vision (Vidal-Gorène et al. 2021) for Armenian manuscripts.

Method

We mainly follow the ScrabbleGAN approach (Fogel et al. 2020) and code [1], using a generator-discriminator combined with a HTR architecture to assist the discrimination process. Instead of training the GAN to generate characters that are joined to create a sentence (the “scrabble approach”), we investigate the relevance to focus on strokes, that are then combined to create characters and words. The choice of the stroke level is motivated by the nature of the scripts envisioned. Indeed, working at the character level, which might be very suitable for printed materials, could result in generated text with too neatly separated, or even artificially juxtaposed, letters, without the ligatures and stroke fusions that are encountered in manuscripts. The recogniser architecture has been reduced compared to the original paper, as we need less abstraction to distinguish the constituent strokes of the letters.

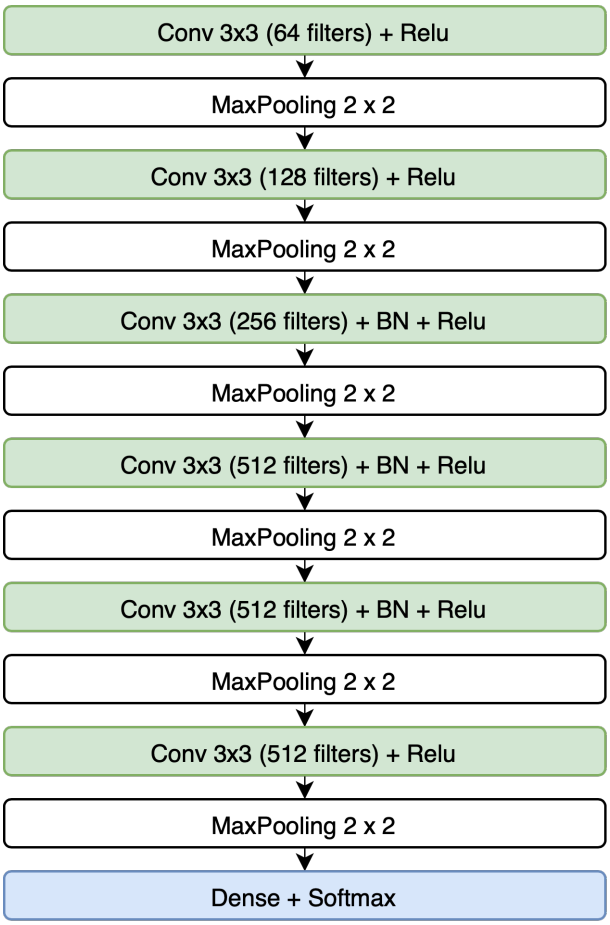

Figure 1: Recogniser architecture, with 3 CNN layers less compared to the original paper. State-of-the-art recogniser generally include LSTM layers, but too high accuracy would lead here to a less realistic generation, as the recogniser would recognise the text even if image is heavily noisy.

Data originally consist in image and ALTO XML pairs. For these experiments, we have applied two *modi operandi*:

1. **with preprocessing**:
   - binarisation of images;
   - extraction of lines patch, basically generating a line image - txt pair, cropping the polygon of the line in a white bbox, with baseline adjustment;
   - Chocomufin of text for Old French (Clérice and Pinche 2021)
2. **without preprocessing**: we only perform the line extraction step with baseline adjustment, keeping the mask of the line.

Results and discussion

Generated images are described below (fig. 2 and 3).

For Armenian, the artificial images, both binarized or non-preprocessed, are very convincing examples of *bolorgir* script, and are fully legible for an expert (and hard to differentiate from an actual manuscript image).

For Old French, the results for the binarized version are less enticing, but nonetheless interesting. In particular, the model seems to insert a spurious character before or after each letter: e.g. the last word of the last line in fig. 2 – that should be *mengier* – actually reads something as *mienigierie*. This might be a way to simplify the work of the recogniser by signalling the end of individual characters.

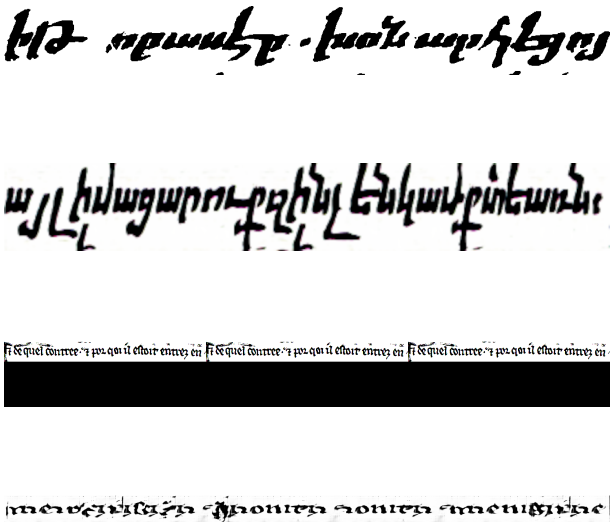

Figure 2: Examples of binarized GT and generated images for Classical Armenian and Old French

Generated color lines, both in Armenian and Old French, are even more convincing for the expert. The presence of the background and all variations of intensity in the ink reinforce the credibility of the images generated. This is an initially counter-intuitive result. The very wide variety of backgrounds (papyrus, parchment, paper ; colorimetry, etc.) nevertheless leads to a defect in the GAN which produces a result with a somewhat blurred appearance, with a slightly dripping color see fig. 3, images 2 and 5). Some of the very rare letters are also less well generated, and produce noise in the colorimetry of the line (see fig. 3, image 6). From a qualitative point of view, it is not possible to identify whether an Armenian line is true or not, except on the question of the abbreviative signs (*badiw*, horizontal line above the lines) drawn randomly (above unabbreviated words). This abbreviation sign is naturally considered as background and therefore generated randomly, because it is not present in the transcription provided.

It is interesting to note that the GANs also reproduce the presence of a black random mask around the text, mask initially generated during the extraction of the lines from the ALTO files. This may appear as noise, however it therefore multiplies the possible scenarios in the composition of the masks and may also be a mean of reinforcing the robustness of the recognition.

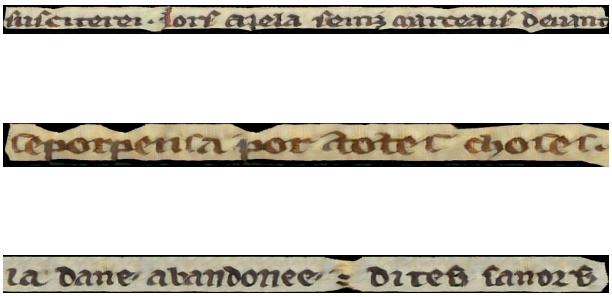

Figure 3: Examples of colour generated images for Classical Armenian and Old French

For now, the recogniser proves too efficient, recognising even images with no human-readable text, and preventing the generator from producing fully historically plausible images. As the recogniser is able to recognise the text even if lines are not realistic, the generator seems to be penalised and stops its attempts to create more realistic results.

At this stage, generated images with their corresponding transcription can already be used as a relevant data augmentation. We have led a very short experiment with the Armenian dataset. We have divided our dataset in 4 subsets : train (827 real lines), val (1 703 real lines), test in-domain (3 772 real lines) and test out-of-domain (2 002 lines) and have performed 4 trainings:

1. default: train set only
2. default+gan250: train set + 250 generated lines
3. default+gan500: train set + 500 generated lines
4. default+gan1000: train set + 1.000 generated lines

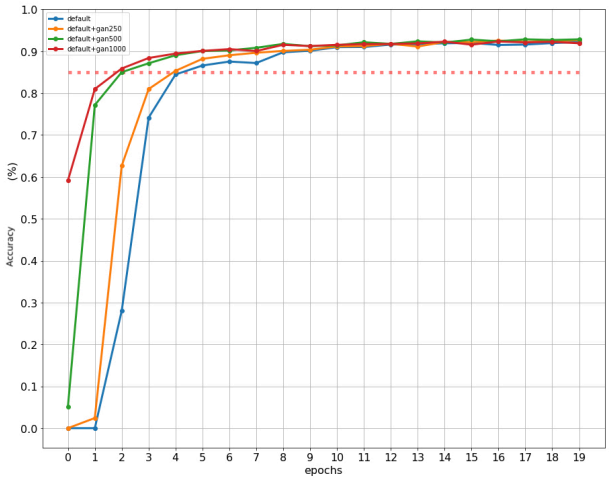

Figure 4: Val accuracy during training

For artificial data generation in Armenian, we have used text from chronicles really far from training real manuscripts (mainly Gospels). This addition of a new vocabulary and sentences is maybe the reason of the increase of HTR performance on out-of-domain tasks.

Training curves (Fig. 4) show that we significantly speed-up training with fake images, as we could have with a bigger training dataset. CER achieved by default model and default+gan500 model are respectively 6.13% and 5.63% on in-domain test, and 22.72% and 10.79% on out-of-domain test.

In this scenario of an under-resourced script, the data augmentation, in particular with out-of-domain words, therefore appears relevant. We have not been able to conduct similar experimentations at this stage for Old French, for which we have a large number of abbreviative signs that are very poorly covered, and therefore poorly generated by the GAN. The main advantage of the method seems to be the qualitative dimension of the data augmentation carried out, which is not just a simple and random transformation of the image. If the final accuracy reached is the same, the model with GAN converges faster – in terms of number of epochs – and demonstrates its relevance in a complete out-of-domain situation. A 100% GAN model, on the other hand, does not yet seem relevant.

Future research

Current experiments highlight the feasibility and the relevance of using GAN to create fake historical manuscripts. We show that state-of-the-art systems can be transferred to historical data, and intend in the future to reduce the recogniser abilities to increase image generation. We also plan to generate complete pages instead of single lines. The approach presented has the disadvantage of requiring an already well-balanced dataset to be able to generate all the characters in a relevant way. A complementary style transfer for poorly endowed characters might have its relevance.

For now, we do not, strictly speaking, offer a evaluation of the (pseudo-)authenticity of generated lines, which is by any means a question difficult to approach. Yet, the verisimilitude of the generated lines for the expert upon direct inspection, on one hand, and the improve in HTR performance (meaning that the HTR models found some measure of information of historical relevance in the generated lines), on the other, offer two complementary approaches to this evaluation. In the future, both approaches could be extended.

Notes

1. https://github.com/amzn/convolutional-handwriting-gan

Bibliography

Clérice, Thibault, and Ariane Pinche. 2021. “Choco-Mufin, a tool for controlling characters used in OCR and HTR projects.” https://doi.org/10.5281/zenodo.5356154.

Etter, David, Stephen Rawls, Cameron Carpenter, and Gregory Sell. 2019. “A Synthetic Recipe for Ocr.” In *2019 International Conference on Document Analysis and Recognition (Icdar)*, 864–69. IEEE.

Fogel, Sharon, Hadar Averbuch-Elor, Sarel Cohen, Shai Mazor, and Roee Litman. 2020. “ScrabbleGAN: Semi-Supervised Varying Length Handwritten Text Generation.” arXiv. https://doi.org/10.48550/ARXIV.2003.10557.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. “Generative Adversarial Networks.” In *Proceedings of Advances in Neural Information Processing Systems 27 (Nips 2014)*.

Karras, Tero, Samuli Laine, and Timo Aila. 2018. “A Style-Based Generator Architecture for Generative Adversarial Networks.” arXiv. https://doi.org/10.48550/ARXIV.1812.04948.

Kiessling, Benjamin, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. “EScriptorium: An Open Source Platform for Historical Document Analysis.” In *2019 International Conference on Document Analysis and Recognition Workshops (Icdarw)*, 2:19–19. IEEE.

Pinche, Ariane. 2022. “CREMMA Medieval, an Old French dataset for HTR and segmentation.” https://doi.org/10.5281/zenodo.5617782.

Stone, Michael E, Dickran Kouymjian, and Henning Lehmann. 2002. *Album of Armenian Paleography*. Aarhus University Press Aarhus.

Vidal-Gorène, Chahan, and Aliénor Decours-Perez. 2021. “A Computational Approach of Armenian Paleography.” In *International Conference on Document Analysis and Recognition*, 295–305. Springer.

Vidal-Gorène, Chahan, Boris Dupin, Aliénor Decours-Perez, and Thomas Riccioli. 2021. “A Modular and Automated Annotation Platform for Handwritings: Evaluation on Under-Resourced Languages.” In *International Conference on Document Analysis and Recognition*, 507–22. Springer.

Vidal-Gorène, Chahan, Anush Sargsyan, and Emmanuel Van Elverdinghe. 2022. “Index of Digitized Armenian Manuscripts.” Zenodo. https://doi.org/10.5281/zenodo.6894290.

Vögtlin, Lars, Manuel Drazyk, Vinaychandran Pondenkandath, Michele Alberti, and Rolf Ingold. 2021. “Generating Synthetic Handwritten Historical Documents with OCR Constrained GANs.” In *2021 16th Iapr International Conference on Document Analysis and Recognition (Icdar)*. Lausanne, Switzerland.