Autoregressive Styled Text Image Generation, but Make it Reliable

Supplementary Material

In this document, we report additional analyses on style text independence and the effect of the training strategies adopted in the second phase of training. Moreover, we report results that include color correction of the output.

1. Style Text Reliance Analysis

When compared to Emuru [50], Eruku does not need style text input. A possible workaround to use Eruku with a style sample with no known ground-truth textual transcription is to use an OCR model to obtain it. We test both Emuru and Eruku with style text input obtained from running TrOCR-Base [31] as an OCR model and comparing them against each other when using the T_s^* text generated by TrOCR, against Eruku ran with no style text input and against a version of Eruku which has never been trained with style text dropout. The results are displayed in Table 6. Eruku is (except for FID) better than Emuru even when using the ground truth text T_s. When using T_s^*, Eruku is able to maintain very low ΔCER, whereas Emuru tends to collapse and/or generate incorrect text more often. Both manage to maintain style consistency. Eruku with no style text gets even better ΔCER scores, but compromises in a significant way on style adherence, as indicated by the high HWD score. The version of Eruku trained with no style text dropout and style text from OCR suffers, just like Emuru, from significantly increased ΔCER from the reliance on this noisy style text. Emuru is incapable of running with no style text input.

2. Ablation on Second Stage Training

In the second stage of pretraining, as described in Section 4, two variations are made to the way the model trains: it is trained on the dataset of images with longer context described in Section 4.2, and it is trained to randomly drop style text conditioning with a probability of $p_{drop} = 0.1$. We investigate the effects of each of those by running training for the same amount of iterations as the full Eruku second stage of training, but with just one strategy or the other. We then compare those runs on IAM lines to the full second stage of training and to the result of just the first stage of training. The results, shown in Table 7, highlight how long-context training improves ΔCER significantly. Style text dropout instead, in addition to allowing the model to generate unconditionally as shown in Section 5, also improves style image adherence, as indicated by the improvement in HWD. The model using both strategies (Eruku) combines the advantages of both and reaches the best HWD values and much-improved ΔCER values when compared to the model resulting from the first stage of training.

	HWD↓	ΔCER↓	FID↓	BFID↓
Eruku w/ T_s	**1.70**	**0.06**	16.40	**4.88**
Emuru w/ T_s	1.87	0.14	**13.89**	6.19
Eruku w/ T_s^*	1.73	**0.06**	16.59	**5.07**
Eruku $p_{drop} = 0$ w/ T_s^*	**1.72**	0.53	15.81	7.68
Emuru w/ T_s^*	1.79	0.42	**14.09**	6.23
Eruku w/o T_s	**2.51**	**0.04**	**20.44**	**9.63**
Emuru w/o T_s	-	-	-	-

Table 6. Emuru and Eruku results on IAM lines when fed with the actual T_s or a T_s obtained by running TrOCR on the style image I_s (dubbed T_s^*). As a reference, we report the results of the generation without T_s.

longer input	$p_{drop} = 0.1$	HWD↓	ΔCER↓	FID↓	BFID↓
✗	✗	1.81	0.40	14.20	**3.38**
✓	✗	1.92	**0.04**	19.45	5.50
✗	✓	1.75	0.40	**13.49**	4.45
✓	✓	**1.70**	0.06	16.40	4.88

Table 7. Ablation analysis on the effect of the second training phase inputs and strategy in terms of performance on IAM Lines.

	HWD↓	ΔCER↓	FID↓	BFID↓
Eruku	1.70	0.06	16.40	4.88
Emuru	1.87	0.14	13.89	6.19
Eruku w/ c.c.	**1.68**	**0.04**	12.21	**4.54**
Emuru w/ c.c.	1.85	0.14	**11.40**	6.20

Table 8. Emuru and Eruku results on IAM lines in the standard setting and when the color correction strategy (c.c.) is applied as post-processing.

3. Results Including Color Correction

Since it relies on the same VAE as Emuru, Eruku generates images with a white background and usually very dark text strokes. This allows the simple color correction strategy proposed in [50] for Emuru to be applicable also for Eruku. The strategy uses the VAE's background removal abilities to isolate the mask containing the text within the style image, then computes the average of the color values among the foreground ink pixels and applies that to those of the generated image. The effect of such color correction post-processing can be observed quantitatively in Table 8 (especially in terms of FID).