5T-09

文字認識モデル訓練のためのスタイル変換を用いた手書き文字生成

北川 智樹†
山梨大学大学院医工農学総合教育部†
レオ チーシャン‡
山梨大学大学院医工農学総合教育部‡
矢島 英明§
山梨大学大学院医工農学総合教育部§
西崎 博光¶
山梨大学大学院医工農学総合教育部¶

1 はじめに

光学文字認識（OCR）技術は, 手書き文書からテキストを抽出する手段であり，深層学習を活用した AI-OCR 技術が進化している. この技術の効果的な実装には大量の手書き文字データが必要であり，特に文字種が多い日本語の場合，手書き文字画像データの収集は困難である.

本研究では，生成モデルを使用して文字認識器の訓練データを生成し，その認識性能の向上を目指す.

生成モデルとしては本研究では Adaptive Instance Normalization (AdaIN) [1] によるスタイル変換モデルに加え Contrastive Language-Image Pre-training (CLIP) [2] によって学習された Text Encoder を組み込み込んだ AdaIN with MLP (CLIP Text Encoder) により既存のスタイル変換よりバリエーションに富んだ文字画像を生成できるようにした. 生成された文字画像を手書き文字識別器の訓練データに加え，手書き文字認識性能向上に対する文字画像生成の有効性を確認した. 実験の結果，文字認識器の訓練データに生成画像を加えることで，加えない場合に比べて認識性能の向上を確認でき，本手法が文字認識器訓練のための手書き文字生成として有効であることが確認できた.

2 AdaIN

AdaIN[1] とは画像のスタイル変換で使われる層である. 式 (1) で定義されるように，本研究の文字生成の文字の形である Content の特徴量 c の平均と分散を文字生成の文字の筆跡である Style の特徴量 s の平均と分散に合わせることにより，学習時に一切のパラメータを持たないでスタイル変換が可能となる.

$$AdaIN(c,s) = \sigma(s)(\frac{c-\mu(c)}{\sigma(c)}) + \mu(s) \quad (1)$$

3 CLIP

CLIP[2] とは対照学習を用いて画像とテキストの関連性を学習する深層学習モデルである. 本研究では図 1 に示すように画像として文字画像，テキストとして漢字の構造情報を分割したテキストを用いて対照学習を行なった.

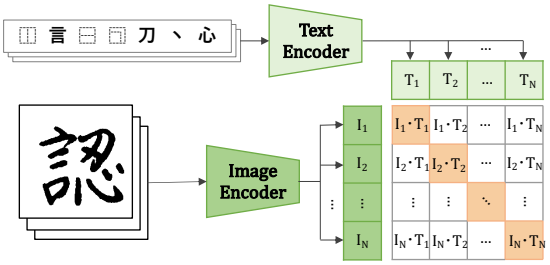

図 1 CLIP の概要

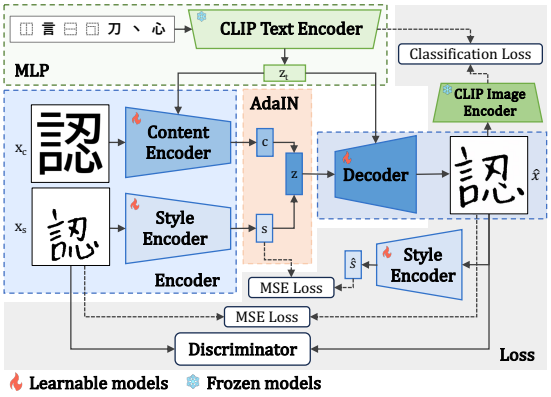

図 2 AdaIN with MLP (CLIP Text Encoder) のモデル構造

4 AdaIN with MLP (CLIP Text Encoder)

4.1 モデルについて

図 2 の提案する生成モデルの構造に示すように Content Encoder が文字の形の特徴 c，Style Encoder が文字の筆跡から特徴 s を抽出する. AdaIN 層によって組み合わせた z を Decoder に入力し最終的な画像 $\hat{x}$ を生成する. また既存の AdaIN によるスタイル変換モデルとは異なり Content Encoder，Decoder に関してダウン，アップサンプリング時に 3 節で説明したように訓練した Text Encoder 用いて得られる特徴 z_t を組み込み，文字の構造情報を考慮した圧縮，復元を促す. 訓練時には 3 つの損失関数を利用する. 1 つ目は，出力画像 $\hat{x}$ に対して Classification Loss として CLIP で用いられる損失関数で，文字の構造の正しさに関しての制約を与える. 2 つ目は MSE Loss で，Style Encoder の出力 s，$\hat{s}$ によるスタイルの特徴量の一致と x_s，$\hat{x}$ による復元性の制約を与える，そして最後は Generative adversarial network (GAN)[3] で用いられる Discriminator Loss を使用して訓練した. これにより，スタイル変換の精

Handwritten Character Image Generation using Style Transformation for Character Recognizer Training
† Tomoki Kitagawa, University of Yamanashi
‡ Chee Siang Leow, University of Yamanashi
§ Hideaki Yajima, University of Yamanashi
¶ Hiromitsu Nishizaki, University of Yamanashi

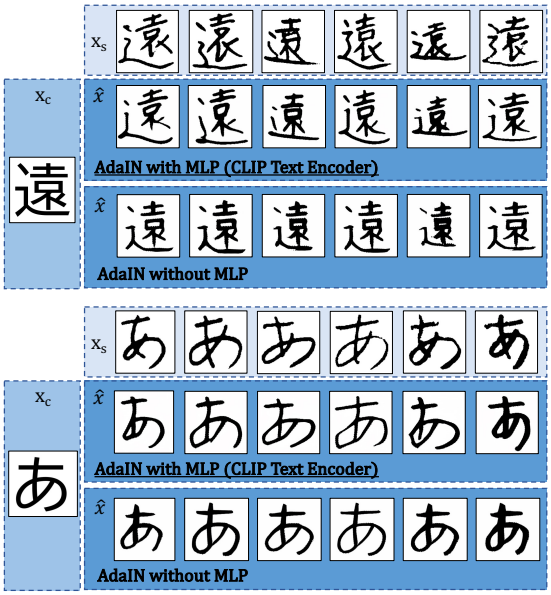

図 3　生成画像の例

度が向上し， Content 画像の特徴を保持しつつ， Style 画像に応じた異なる画像が生成できる．

4.2　生成モデルの訓練

生成モデルの訓練に使用したデータは ETL 文字データベース [4] の 9 つの手書き文字コーパスからひらがなと漢字の各 200 枚を収録する ETL9，カタカナ各 208 枚を収録する ETL5 と，様々な種類の日本語のフォント画像を使用して訓練した．また訓練する際には最初は Content，Style 画像ともにフォント画像から入力して訓練し，徐々に手書き画像を訓練データに加えていくように訓練した．

4.3　文字画像生成

本研究では，Content 画像 x_c として 1 種類のフォント画像を使用し，Style 画像 x_s には訓練に利用したひらがな，カタカナ，漢字の画像を適用した．この手法により，約 61 万枚の文字画像を生成した．生成画像 $\hat{x}$ の例を図 3 に示す．さらに，従来の AdaIN スタイル変換モデル (AdaIN without MLP) の生成画像と比較した結果，学習済みの CLIP Text Encoder を組み込んだモデル (AdaIN with MLP) は， Content 画像の特徴を保持しつつ， Style 画像に応じてより顕著に変化することを確認した．

5　手書き文字認識実験

生成された文字画像を文字認識器の訓練に使用し，手書き文字認識率の評価を行う．

5.1　実験条件

カタカナ 46 種類，ひらがな 46 種類，漢字 2,965 種類，計 3,056 種類の文字認識器を訓練した．ベースラインは手書き文字画像だけとして，手書き文字にデータ拡張の適応，手書き文字と AdaIN without MLP，提案手法の AdaIN with MLP (CLIP Text Encoder) による生成画像から訓練したモデルで，ETL の検証データに対する macro-F1 で有効性を調べた．文字認識器のモデルは 152 層からなる ResNet[5] とし，評価データに対する最大精度のモデルで比較した．また，他のハイパパラメータも全て一致させた．手書き文字画像は生成モデルの訓練・生成に使用した ETL9 と ETL5 とし，生成画像は 4.3 節の文字画像生成の全ての画像とした．検証・評価データは，ETL1・7・8 からひらがなとカタカナについては検証と評価データそれぞれで各文字あたり 200 枚を，漢字については検証のために ETL1・7・8 から各文字 60 枚，評価ために各文字 100 枚の画像を使用した．

表 1　文字認識器の精度

	macro-F1
ベースライン（ETL5，9）	0.9733
データ拡張（ETL5，9）	0.9797
AdaIN with MLP（ETL5，9 + 生成画像）	**0.9832**
AdaIN without MLP（ETL5，9 + 生成画像）	0.9750

5.2　実験結果

表 1 に，文字認識器の認識結果を示す．ベースラインの macro-F1 スコア 0.9733 と比較して，データ拡張を施した場合と AdaIN without MLP の両方で若干の精度向上が見られ，それぞれのスコアは 0.9797 と 0.9750 となった．しかし，最も顕著な改善は AdaIN with MLP を適用した場合で，最高のスコア 0.9832 を達成した．これにより，MLP(CLIP Text Encoder) の適用が生成画像のバリエーションを増加に寄与し，生成された文字画像を加えて訓練することで，手書き文字認識性能が向上することが確認できた．

6　おわりに

本研究は，文字認識の精度向上を目的とし，生成モデルによって手書き文字認識の訓練に必要な文字画像を生成することで文字認識器の訓練データを増やす手法を提案した．手書き文字認識実験の結果，生成された文字画像を文字認識器の訓練データに加えることで文字認識器の認識性能が向上したことから，生成モデルによる文字生成が手書き文字認識性能の向上に有効であることが確認できた．

今後の展望として，さらなる認識器の精度向上につながる画像生成や複数文字の生成による複数文字，複数行の認識性能向上を目指す．

参考文献

[1] X. Huang, S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” Proc. of ICCV, pp. 1510–1519, 2017.
[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, “Learning transferable visual models from natural language supervision,” Proc. of ICML, pp. 8748–8763, 2021.
[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville , Y. Bengio, “Generative Adversarial Nets,” Proc. of NeurIPS, vol. 2, pp. 2672–2680, 2014.
[4] 独立行政法人産業技術総合研究所, “ETL 文字データベース” http://etlcdb.db.aist.go.jp, (2024.1.9 参照)
[5] K. He, X. Zhang, S. Ren, J. Sun, “Deep residual learning for image recognition,” Proc. of CVPR, pp. 770–778, 2016.