文字単位の解釈可能な潜在表現の data augmentation

青木 匠 † 北田 俊輔 ‡ 彌冨 仁 †‡

† 法政大学 理工学部 応用情報工学科
‡ 法政大学 理工学研究科 応用情報科学専攻
{takumi.aoki.4g@stu., shunsuke.kitada.8y@stu., iyatomi@}hosei.ac.jp

概 要

深層学習ベースのモデルにおいて，日本語や中国語などのアジア圏の言語の解析は単語単位よりも文字単位での処理が効果を上げている．しかし，過学習が起きやすいため，過学習抑制手法を適用する必要がある．本研究では β-variational auto-encoder (β-VAE) が各次元独立の低次元確率分布を獲得することを活用し，解釈可能な data augmentation である interpretable wildcard training (IWT) を提案する．IWT は β-VAE により得られた文字の低次元表現に対して， ガウス分布に従ったノイズを付加させることで，異なる文字の表現生成が可能であり，従来の wildcard training よりも解釈性が高い．新聞記事の分類タスクによる評価実験において，IWT による解釈可能な文字表現の獲得ならびに，2%程度の分類精度向上から，解釈性のある data augmentation の効果を確認した．

1 はじめに

文書解析において英語などの言語では，一般的に単語単位での処理が行われるが，日本語や中国語などのアジア圏の言語は単語ごとに明確な区切りがないため，事前の単語分割が必要であるが容易ではない．この問題に対して，文字単位で処理する手法が複数提案されている [1, 2, 3]．特に画像処理分野で成果を上げている convolutional neural networks (CNN) を用いた character-level CNN (CLCNN) は，one-hot 表現とした文字ごとに一次元方向に畳み込み処理することで単語分割の困難さを回避し，優れた文書分類能を実現している [1]．しかしながら日本語や中国語など，文字種の多い言語では入力が高次元になるため，過学習を引き起こしてしまう問題がある．

文字種の多い言語の表意性に着目し，各文字を画像として扱い convolutional auto-encoder (CAE) [4] で文字形状を考慮した文字表現を獲得する手法が提案されている．また，こうした文字表現を CLCNN と組み合わせることで，優れた日本語の文書分類 [2] や中国語の言語モデリングと単語分割 [5] といったタスクで高い精度が確認されている．さらに，こうした文字形状に基づく文字表現の獲得から文書分類までを end-to-end で行うモデルでは高い精度が確認されると共に，k-最近傍により似た文字形状の文字表現が近いと報告されている [3, 6]．しかし、CAE で得られた低次元表現の解釈性は低く，また意味のある他の表現に変換することは困難である．解釈可能な潜在表現の獲得する 1 つの手法として，β-variational auto-encoder (β-VAE) [7] が提案されている．β-VAE は，入力データの低次元表現を直接獲得する通常の CAE と異なり，入力データの低次元表現を構成する確率パラメータを学習する．この確率パラメータは正規分布 N(0,I) にな

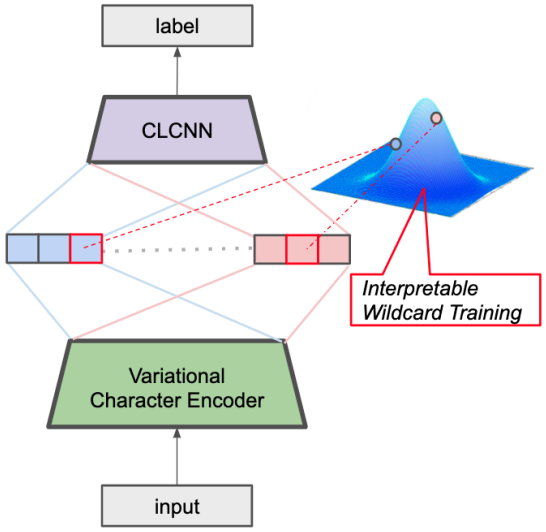

図 1: 提案文書分類手法の全体像．入力テキストを各文字の文字画像として扱い，β-VAE の encoder 部分の variational character encoder により文字表現を獲得する．本研究で新たに提案する interpretable wildcard training によりエンコードされた文字表現を関連する他の文字表現に擬似的に置き換えを行うことにより，分類器である CLCNN に対して学習時に data augmentation の効果を与える．

るよう学習が行われるため，得られる表現の各次元は互いに高い独立性が期待できる．このため、β-VAE を文字画像の学習に適用した場合， 部首やつくりといった特徴が，各次元に独立した表現として得られることが期待できる．こうした文字表現を用いることで、文字形状に特化した data augmentation の実現が可能であると考えられる．

深層学習モデルはその自由度の高さから過学習しやすいため，擬似的にデータを増やす data augmentation が広く使われている．自然言語処理における data augmentation では，一般的にテキストを単語分割したのちに，シソーラスを用いて置換可能な単語を類義語に置換する．しかし，こうした処理は正確に単語分割を行い，意味の解析を要するため容易ではない．この問題に対して，単語分割と意味解析が不要な wildcard training (WT) が提案されている [2]．WT は学習時に，文字表現の一部をランダムに dropout [8] することで任意の文字として扱うことを期待した手法であり，複数のタスクでその優れた効果が確認されている．しかし，WT は文字表現の一部をランダムに dropout しているので，意味的な構造を考慮しておらず，改善の余地が残されている．

本研究では学習データの低次元確率分布を学習する auto-encoder である β-VAE の長所を活用し，解釈性の高い文字表現の獲得可能な文書分類および，data augmentation である interpretable wildcard training (IWT) を提案する．新聞社の web 記事を用いた新聞社推定を評価実験として行い，IWT を有無による予測精度を比較ならびに、data augmentation としての解釈性を評価した．

2 提案手法

本論文では，文書分類問題において優れた文書分類能と，各文字の解釈性の高い低次元表現の両立を実現する文書分類モデルおよび，その効果を高める data augmentation である interpretable wildcard training (IWT) を提案する．提案する文書分類手法の全体像を図 1 に示す．提案手法は大きく分けて以下の 2 つの過程から構成されている．

- 各次元が独立した特徴を学習する β-VAE による文字表現の学習
- 文字表現を入力として文書分類を行う CLCNN による文書分類の学習

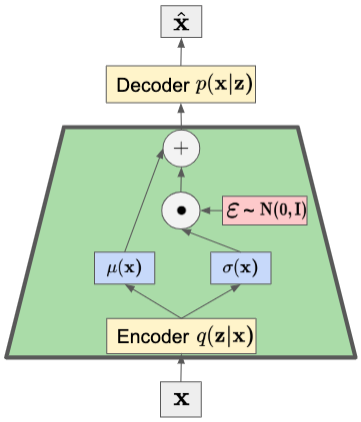

図 2: VAE の全体像

文字表現の学習では，β-VAE で文字形状を考慮しつつ潜在表現の各次元が独立した特徴を学習させる．文書分類の学習では，β-VAE でエンコードされる文字表現を元に，CLCNN を学習させる．

2.1 β-VAE による文字表現の学習

β-VAE は VAE の目的関数の正則化項に係数 β を導入し，潜在変数が事前分布に従う制約を強めている．VAE はデータ分布 $\mathrm{p}(\boldsymbol{x})$ を推定する生成モデルである．事前分布のガウス分布 $\mathrm{p}(\boldsymbol{z})$，事後分布 $\mathrm{q}(\boldsymbol{z}|\boldsymbol{x})$，生成モデル $\mathrm{p}(\boldsymbol{x}|\boldsymbol{z})$ とし， 以下の目的関数を最小化する．

$$L = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}|\boldsymbol{z})] - \beta D_{\mathrm{KL}}[q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})] \quad (1)$$

第一項は画像の再構成誤差，第二項は潜在変数が事前分布に従うように学習することを意味する．平均 μ，分散 σ として VAE の全体像を図 2 を示す．潜在変数を直接確率分布とすると encoder まで誤差逆伝播できないため，近似手法の reparameterization trick を使用している．今回使用した β-VAE における encoder と decoder を表 1a に示す．β-VAE によって学習させた encoder 部分を本研究では variational character encoder と呼び，エンコードされる潜在表現が独立した特徴を獲得した文字表現として利用する．

2.2 CLCNN による文書分類の学習

入力テキストを各文字画像として扱い，学習済み variational character encoder からエンコードされた文字表現を埋め込み，一次元方向に畳み込んで学習をする．今回使用した CLCNN のアーキテクチャを表 1b に示す．学習時には variational character encoder 部分を固定して，CLCNN のパラメータのみをクロスエ

表 1: 本研究で用いる β-VAE および CLCNN のアーキテクチャ (カーネルサイズ k，出力サイズ o，ストライド s)

(a) β-VAE の encoder と decoder のアーキテクチャ

Layer	Encoder	Decoder
1	Conv(k=(4，4)，o=32，s=2) → ReLU	Linear(o=256) → ReLU
2	Conv(k=(4，4)，o=32，s=2) → ReLU	Linear(o=1024) → ReLU
3	Conv(k=(4，4)，o=64，s=2) → ReLU	Deconv(k=(4，4)，o=64，s=2) → ReLU
4	Conv(k=(4，4)，o=64，s=2) → ReLU	Deconv(k=(4，4)，o=32，s=2) → ReLU
5	Linear(o=256) → ReLU	Deconv(k=(4，4)，o=32，s=2) → ReLU
6	Linear(o=2×10)	Deconv(k=(4，4)，o=1，s=2) → Sigmoid

(b) CLCNN のアーキテクチャ

Layer	CLCNN
1	Conv(k=(1，3)，o=512) → ReLU
2	Maxpool(k=(1，3)，s=3)
3	Conv(k=(1，3)，o=512) → ReLU
4	Maxpool(k=(1，3)，s=3)
5	Conv(k=(1，3)，o=512) → ReLU
6	Conv(k=(1，3)，o=512) → ReLU
7	Linear(o=classes)

ントロピー誤差関数を目的関数として誤差逆伝播法によって最適化する．

2.3 Interpretable wildcard training

IWT は β-VAE により得られた表現の各次元が独立した表現になることに着目し，得られた文字表現のある一つの次元に対してガウス分布に従ったノイズを付加させることで，関連する他の文字表現になることを期待した新たな data augmentation である．このとき，文字表現は各次元に独立した特徴を持っていることから，変化する文字表現の解釈が可能となる．

3 実験

評価実験では文字画像を元にした文書分類問題において、識別能と、IWT による文字表現の解釈性および、data augmentation としての効果、つまり分類能向上の評価を行った．

3.1 実験設定

文字表現および文書分類の学習に用いた β を 20 とした β-VAE と，CLCNN のパラメータの 最適化には Adam [9] を用いた．CLCNN の学習に使用する wildcard traning における wildcard 率は 0.1 に設定した．

比較対象として，先行研究 [2] で用いられている，CAE を用いた文字表現の学習とその文字表現を用いた CLCNN による文書分類の学習手法 (CAE+CLCNN) ならびに，その中で提案されている wildcard training (WT) を適用したモデル (CAE + CLCNN + WT) を用い，文書分類能を比較した．

また，提案手法である VAE + CLCNN + IWT の有効性を確認するために IWT に WT に変更した VAE + CLCNN + WT についても比較した．以下に文字表現の学習と文書分類の学習それぞれの実験設定の詳細を示す．

文字表現の学習　文字表現の学習するにあたり，対象の文字として平仮名，片仮名，漢字　(JIS 第一・二水準)，英数字，記号を含む計 6,625 文字の常用日本語を使用した．β-VAE の入力には常用日本語の文字を 64×64pixels のグレースケール画像に変換して入力した．

文書分類の学習　文書分類の評価として，朝日，毎日，産経，読売新聞の政治，経済，国際カテゴリの記事で各社 5,610 件 計 22,440 件の web 新聞記事を使用した．これらの記事のうち 8 割を学習用，2 割を評価用に分割を行った．前処理として，文字列長を 128 になるようにランダムクロップし，常用日本語以外は空白として各文字を 64×64pixels のグレースケール画像に変換した．モデルの学習時には文書中から連続する 128 文字分だけ取り出して学習に使用し，評価時には 128 文字を 1 つずつスライドさせ，文書全体を入力として使用し，評価を行った．

3.2 実験結果

実験結果では β-VAE によって学習された文字表現，およびそれらを用いた CLCNN による文書分類について示す．また interpretable wildcard training による data augmentation の効果について述べる．

β-VAE による文字表現の学習　学習した β-VAE で道，進，遠の文字表現の 9 次元目を-3 から 3 まで 0.5 ずつ動かしたときの再構成画像を図 3 に示す．β-VAE が学習により獲得する各次元の値は平均 0，分散 1 の正規分布に近くなるため，この範囲で動かすことでこ

図 3: 道，進，遠の文字表現の 9 次元目を-3 から 3 まで変化させた再構成画像

表 2: Web 新聞記事の新聞社推定の結果

Model	**Acc.** [%]
(Ours) β-VAE + CLCNN + IWT	83.00
(Ours) β-VAE + CLCNN + WT	83.11
(Ours) β-VAE + CLCNN	81.27
CAE + CLCNN + WT [2]	**83.27**
CAE + CLCNN [2]	78.32

の次元の影響をほぼ観察することができる．この例では，対象次元の値を-3 にしたときは水などを意味する “さんずい”，3 にしたときは草などを意味する “くさかんむり” に変化した．このように，ガウス分布に従ったノイズを付加することで意味のある文字表現に変換することが確認できた．

Web 新聞記事の新聞社推定　Web 新聞記事の新聞社推定能の比較を表 2 に示す．文字表現として CAE でエンコードした低次元表現を使用した場合と比較して，β-VAE でエンコードした埋め込みを使用した場合，3%程度の精度向上が確認できた．これは CAE で学習される埋め込みと比べて，β-VAE による各次元が独立した特徴を学習する優れた埋め込みによるものだと考えられる．また，本研究で新たに提案した data augmentation である IWT を導入することで，2%程度の精度向上が確認できた．また，解釈性の高い data augmentation が可能な IWT が WT と同程度の予測精度を達成した．

4　おわりに

本研究では学習データの低次元確率分布を学習する auto-encoder である β-VAE を活用し，より優れた文字表現の獲得と解釈可能な識別モデルおよび、それを活かした効果的な data augmentation である interpretable wildcard training (IWT) を提案した．実験により IWT は意味のある他の文字表現に変換が可能であり，得られる文字表現は各次元が独立した特徴を持っていることで解釈性が高い data augmentation であることが確認された．

参考文献

[1] X. Zhang, J. Zhao, and Y. LeCun, “Character-level convolutional networks for text classification,” in *Proc. of NIPS*, 2015, pp. 649–657.

[2] D. Shimada, R. Kotani, and H. Iyatomi, “Document classification through image-based character embedding and wildcard training,” in *Proc. of IEEE Big Data 2016, Big Data and Natural Processing Workshop (BigNLP 2016)*. IEEE, 2016, pp. 3922–3927.

[3] S. Kitada, R. Kotani, and H. Iyatomi, “End-to-end text classification via image-based embedding using character-level networks,” in *Proc. of IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. IEEE, 2018, pp. 1–4.

[4] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, “Stacked convolutional auto-encoders for hierarchical feature extraction,” in *International Conference on Artificial Neural Networks*. Springer, 2011, pp. 52–59.

[5] F. Z. Dai and Z. Cai, “Glyph-aware embedding of chinese characters,” *Proc. of the First Workshop on Subword and Character Level Models in NLP*, 2017.

[6] F. Liu, H. Lu, C. Lo, and G. Neubig, “Learning character-level compositionality with visual features,” *Proc. of ACL*, 2017.

[7] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, “beta-vae: Learning basic visual concepts with a constrained variational framework.” *ICLR*, vol. 2, no. 5, p. 6, 2017.

[8] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” *CoRR preprint arXiv:1207.0580*, 2012.

[9] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR preprint arXiv:1412.6980*, 2014.