Masked Image Modeling を利用した情景画像中のテキスト認識

三ツ井悠翔　宮崎智　大町真一郎
東北大学大学院
yuto.mitsui.s1@dc.tohoku.ac.jp　tomo@tohoku.ac.jp
shinichiro.omachi.b5@tohoku.ac.jp

概要

既存のテキスト認識手法は実世界におけるデータセットのサンプル数が少ないため，合成データセットを用いて学習がされているが、実世界で発生する問題に対応できない。そこで，ラベルがない実画像の利用によってテキスト認識モデルの可能性を引き出すことが考えられており，テキスト認識に対する自己教師あり学習手法が検討されている。本研究では、Masked Image Modeling を利用し、文脈情報を考慮した新たなマスキング戦略を提案した。実験の結果，提案するマスキング戦略の有効性が実証された．

1　はじめに

情景画像中のテキスト認識は自然なシーンにおけるテキストの読み取りを目標としており，幅広い応用が存在するため重要なタスクとなっている．光学式文字認識 (OCR) の分野は大きく発展したが，実世界で発生するフォントや文字の形，撮影環境による問題によって，情景画像中のテキスト認識は現在も難しいタスクとなっている．既存のテキスト認識手法の多くは実世界におけるデータセットのサンプル数が少ないため，大規模な合成データセットを用いて学習がされている．しかし，実データと合成データのドメインギャップにより，実世界で発生する問題に対応できない．そこで，アノテーションされていない実画像の利用によってテキスト認識モデルの可能性を引き出すことが考えられており，テキスト認識に対する自己教師あり学習手法が検討されている．

これまでの研究では，Contrastive Learning（対照学習）の導入が試みられている．SeqCLR[1] は，テキスト認識のための sequence-to-sequence 対照学習手法を提案した．PerSec[2] はテキスト認識のための階層的な対照学習手法を提案した．

自然言語処理では，BERT[3] が Masked Language Modeling (MLM) による大きな成功を収めている．画像分野では BERT に触発され，データの一部を削除し，削除した部分を復元する考え方を画像に適用した Masked Image Modeling (MIM) が提案されている．2 つの手法は類似しており，PIXEL[4] は文をテキスト画像に変換し，MIM を用いることで BERT と同程度の結果を出している．

本研究では，MIM をテキスト認識に適用し，単語の文脈情報を反映した自己教師あり事前学習を行うことを目的としている．特に，MIM 手法として Masked AutoEncoder (MAE)[5] を利用し，マスキング戦略に対して変更を行った．具体的には，横方向には，ランダムなスパンに対してマスキングを行い，縦方向には一様なマスキングを行う．ランダムマスキングよりも，意味のある単位（文字や複数文字）をマスクすることで，抽象度の高いモデルを作るような効果があると考えられる．提案するスパンマスキングで事前学習を行うとランダムマスキングよりも 0.8%以上高い正解率が得られた．本論文の貢献は以下の 3 点である．

1. テキスト認識に対して初めて MAE を利用した事前学習を行った．
2. 文脈情報を考慮した新たなマスキング戦略を提案した．
3. 提案手法によって事前学習し，テキスト認識でのファインチューニングを行ったところ，事前学習無しの場合よりも 2.7%，他のマスキング戦略よりも 0.8%高い正解率を達成した．

2　関連研究

2.1　Masked Image Modeling(MIM)

MIM は自然言語処理における MLM の発展とともに，近年発展している．MIM 手法の一つである

MAE はマスクされていない画像パッチに対してのみ動作するエンコーダと潜在表現とマスクトークンから元の画像を構成する軽量なデコーダを用いて，入力画像の 75 %という高い割合でマスクを行うことで有効な事前学習ができることを可能にした．

本研究では，自己教師あり学習手法である MIM に着目し，計算効率と有効性の観点から MAE を事前学習に用いる．

2.2 自己教師ありテキスト認識

提案されている自己教師ありテキスト認識手法は，Sequence-to-Sequence モデルに基づいている．SeqCLR[1] は，テキスト認識のための対照学習手法を提案した．この手法では，順序に関する情報を維持したまま，シーケンスの個々の要素に対照学習を適用することができる．PerSec[2] もテキスト認識のための対照学習を提案している．提案された手法は，特徴量の各要素を高レベルと低レベルで比較する階層的な対照学習を行っている．

このようなテキスト認識における対照学習手法の成功に対して、本研究では MIM 手法を活用し，その有効性について検討している．

3 方法

3.1 Masked Autoencoder

本研究では，自己教師あり事前学習として MIM 手法である Masked Autoencoder(MAE)[5] を利用しており，以下の 4 つの主要な要素から構成されている．

スパンマスキング 画像を非重複パッチに分割し，一部のパッチをマスクする（除去）．マスキングの方法として本研究で提案するスパンマスキングと MAE で用いられているランダムマスキングを用いた．ランダムマスキングは一様分布に従ってランダムなマスクを生成する．一方，スパンマスキングはテキスト画像が水平方向に対してのみ文脈情報を保持していると仮定し，文脈情報を考慮して垂直方向と水平方向で異なるマスキングを行う．垂直方向に関しては文脈情報を持たないため，一様なマスクを生成する．水平方向は文脈情報を持つため，PIXEL[4] のマスキング戦略を利用し，意味のある単位でマスクを生成する．具体的には，スパン間にいくつかのマスクされていないパッチを残しながら，最大 $S = 6$ パッチの連続画像スパンに対してマスキングを行う．マスクの割合に関しては，それぞれの戦略について 25, 50, 75%の値を適用して実験を行った．マスク画像の例を図 1 に示す．

(a)Random Masking (b)Span Masking

図 1 マスク画像の例

エンコーダ ViT[6] を利用し，マスクされていないパッチのみを処理する．これによって必要メモリを削減して学習速度を上げることに加え，事前学習とファインチューニング間のミスマッチを減少させている．

デコーダ デコーダには，エンコーダの出力とマスクトークンが入力され，画像の再構成タスクが実行される．デコーダは，6 層のトランスフォーマーブロックと全結合層によって構成され，計算量はエンコーダに対して小さくなるように設計されている．

再構成ターゲット デコーダの出力は再構成された画像に再形成され，再構成画像と元画像の平均二乗誤差（MSE）が損失関数として利用される．再構成の対象としてマスクされたパッチの画素値を正規化したものを用いることで性能が向上することがわかっているため [5]，この方法を利用する．また，テキスト認識時には RGB 画像ではなく白黒画像を利用するため，再構成の対象として白黒画像を用いる．

3.2 テキスト認識

テキスト認識モデルとして Transformer Encoder と予測ヘッドで構成される ViTSTR[7] を用いる．Transformer Encoder 部分で，MAE によって事前学習されたパラメータを継承し，ファインチューニングを行う．その際，予測ヘッドは初期化する．

4 実験

4.1 データセット

合成データセット 代表的な合成データセットである MJSynth(MJ)[8] と SynthText(ST)[9] を組み合わせて使用した．MJ は 9M 個のテキスト画像で構成されている．ST は 800k の画像から 8M のテキスト

表 1 学習設定

Train dataset: MJ + ST	Batch size: 192
Iterations: 300k	Schedular: Cosine learning rate scheduler
Optimizer: Adadelta	Learning rate: 1.0
Adadelta ρ: 0.95	Adadelta ϵ: $1e^{-8}$
Loss: Cross Entropy	Gradient clipping: 5.0
Image size: 224×224	Channels: 1 (grayscale)

画像を切り出したものである．

情景画像中のテキスト認識ベンチマーク 情景画像中テキスト認識ベンチマークで評価した．ベンチマークはテキストの難易度やテキストのレイアウトによって regular データセットと irregular データセットに区別することができる．

まず，regular データセットは，文字の感覚が均等で，水平にレイアウトされた比較的容易なテキスト画像が含まれる．IIIT5K-Words(IIIT)[10]，Street View Text(SVT)[11]，ICDAR2013(IC13)[12] が該当する．

次に，irregular データセットには曲がったテキストや回転，歪んだテキストなど難しいケースが含まれている．ICDAR2015(IC15)[13]，SVT Perspective(SP)[14]，CUTE80(CT)[15] が該当する．

4.2 実験詳細

自己教師あり事前学習 MAE のエンコーダとして ViT-Tiny を用いた．ViT-Tiny は Patch Size が 16，Embedding Size が 192，Head の数が 3 の ViT であり，基本的な ViT と比較してパラメータ数が小さいモデルである．事前学習には合成データセット（MJSynth と SynthText）の両方を用いた．計算資源の都合上，データ量は 50%とした．MAE の論文と同様に学習を行い，ハイパーパラメータについても論文に従った．

テキスト認識でのファインチューニング ViTSTR で用いられていた学習設定をそのまま使用した．(表 1) 学習データとして合成データセット（MJSynth と SynthText）の両方を用いた．テストデータには，4.1 節で紹介した情景画像中テキスト認識ベンチマークを用いた．

4.3 事前学習におけるマスク画像の再構成

事前学習において一枚のマスク画像を再構成した際の結果を図 2 に示す．ランダムマスキングの場合，マスクの割合が 75%という不鮮明な場合でも，ほとんどの場合で文字列を再構成できていた．ここから，モデルが優れた特徴表現を学習していると考えられる．一方で，スパンマスキングでは，マスク

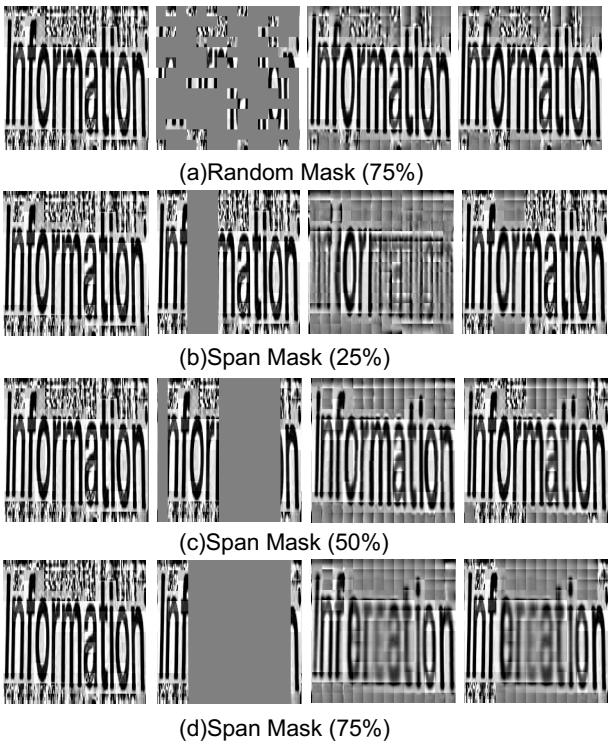

図 2 マスク画像の再構成結果（左から原画像，マスク画像，出力画像，マスク+出力画像）

の割合が高くなるにつれて成功率が低くなっていた．マスクの割合が 25%の時はマスク部分の文字を予測できていることが多かった．ここから，見えなくなっている文字を予測しており，文脈情報を事前学習で獲得していることがわかる．75%になると，例のようにマスク部分の文字が不鮮明になってしまっている例が多かった．隠される文字が多すぎると，正しい文脈情報を獲得することはできていないと考えられる．

4.4 テキスト認識

事前学習したモデルをテキスト認識タスクでファインチューニングし，評価を行った．表 2 に学習データのすべて（100%）を使用した結果，表 3 に学習データの 1%を使用した結果を示す．評価指標には Top1 Accuracy を用いた．

学習データ全てによる学習では，スパンマスキングで 75%の割合をマスクした時，他の場合よりも 0.8%以上高い正解率が得られた．ランダムマスキングに着目すると，50%の時に最も良い正解率が得られており，MAE による ImageNet での結果 (75%) とは異なっていた．通常の画像とテキスト画像では情報の密度が異なっていることが原因と考えられる．スパンマスキングでは，マスクの割合が高くなるに

表 2 テキスト認識の結果（学習データ 100%）

Pretrain Method	Regular			Irregular			Avg.
	IIIT	SVT	IC13	IC15	SP	CT	
Scratch	81.9	81.9	89.4	70.7	71.9	61.8	76.3
Random(25%)[5]	82.7	82.2	89.8	71.3	74.6	65.6	78.2
Random(50%)[5]	83.5	83.5	89.3	72.4	74.4	66.0	74.4
Random(75%)[5]	81.7	80.1	88.0	66.4	69.3	61.1	76.0
Span(25%)(Ours)	82.4	82.5	88.8	69.1	71.5	61.5	76.0
Span(50%)(Ours)	**84.3**	83.5	89.7	70.8	75.2	66.0	78.2
Span(75%)(Ours)	83.8	**84.5**	**90.2**	**72.4**	**75.3**	**67.7**	**79.0**

表 3 テキスト認識の結果（学習データ 1%）

Pretrain Method	Regular			Irregular			Avg.
	IIIT	SVT	IC13	IC15	SP	CT	
Scratch	58.0	54.3	71.3	37.1	37.8	28.1	47.8
Random(25%)[5]	59.5	52.9	69.8	38.4	40.6	28.8	48.3
Random(50%)[5]	63.1	59.2	73.4	44.0	44.7	31.9	52.7
Random(75%)[5]	55.7	51.3	65.2	34.2	34.7	22.9	44.0
Span(25%)(Ours)	67.2	65.1	76.0	47.2	47.4	34.0	56.1
Span(50%)(Ours)	74.5	**74.3**	**83.5**	**57.8**	**60.6**	**48.3**	**66.5**
Span(75%)(Ours)	**75.2**	73.6	83.1	57.4	59.2	44.4	65.5

つれて正解率が上がっている．マスク部分のスパンの長さが長くなり，より意味のある単位でマスクされたことによりこの結果になったと考えられる．しかし，事前学習で復元された画像の品質とは結果が異なっており，2 つの結果に相関がないと考えられる．

学習データの 1%での学習では，スパンマスキングで 50%の割合をマスクした時にスクラッチやランダムマスキングよりも 10%以上高い正解率が得られた．スパンマスキングが，MAE で利用されているランダムマスキングよりも平均 10%以上正解率が高くなっており，提案手法の有効性が実証された．学習データ全てを用いた時とは傾向が異なっており，マスクの割合が 50%の時に正解率が最も高くなった．事前学習時に，75%では文字をはっきりと予測できていないことが関係していると考えられる．

5 結論

本論文ではテキスト認識のための自己教師あり学習として Masked Image Modeling に着目し，新たなマスキング戦略を提案した．提案したマスキング戦略は，MAE に用いられているマスキング戦略よりも有効であるということが分かった．

今後は事前学習での実世界データセットの利用やモデルサイズの変更，既存手法との比較実験を行いたいと考えている．

参考文献

[1] Aviad Aberdam, Ron Litman, Shahar Tsiper, Oron Anschel, Ron Slossberg, Shai Mazor, R Manmatha, and Pietro Perona. Sequence-to-sequence contrastive learning for text recognition. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 15302–15312, 2021.

[2] Hao Liu, Bin Wang, Zhimin Bao, Mobai Xue, Sheng Kang, Deqiang Jiang, Yinsong Liu, and Bo Ren. Perceiving stroke-semantic context: Hierarchical contrastive learning for robust scene text recognition. AAAI, 2022.

[3] Zhimin Bao Mobai Xue Sheng Kang Deqiang Jiang Yinsong Liu Hao Liu, Bin Wang and Bo Ren. Bert: Pre-training of deep bidirectional transformers for language understanding. **In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019**, p. 4171–4186.

[4] Phillip Rust, Jonas F Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. Language modelling with pixels. **arXiv preprint arXiv:2207.06991**, 2022.

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 16000–16009, 2022.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. **arXiv preprint arXiv:2010.11929**, 2020.

[7] Rowel Atienza. Vision transformer for fast and efficient scene text recognition. In **International Conference on Document Analysis and Recognition**, pp. 319–334. Springer, 2021.

[8] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. **arXiv preprint arXiv:1406.2227**, 2014.

[9] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In **Proceedings of the IEEE conference on computer vision and pattern recognition**, pp. 2315–2324, 2016.

[10] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In **BMVC-British machine vision conference**. BMVA, 2012.

[11] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In **2011 International conference on computer vision**, pp. 1457–1464. IEEE, 2011.

[12] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In **2013 12th international conference on document analysis and recognition**, pp. 1484–1493. IEEE, 2013.

[13] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In **2015 13th international conference on document analysis and recognition (ICDAR)**, pp. 1156–1160. IEEE, 2015.

[14] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In **Proceedings of the IEEE International Conference on Computer Vision**, pp. 569–576, 2013.

[15] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. **Expert Systems with Applications**, Vol. 41, No. 18, pp. 8027–8048, 2014.