G-005

マスク率を動的に変化させ事前学習した Masked Autoencoders による分類精度向上

Dynamic Modulation of Masking Rate in Pre-training of Masked Autoencoders Improves Classification Accuracy

畑 直宏† 多田村 克己† 佐村 俊和†
Naohiro Hata Katsumi Tadamura Toshikazu Samura

1. はじめに

自然言語処理において高い性能を示す Transformer [1] を画像処理に応用した Vision Transformer(ViT) [2] は高い分類精度を示す．しかし，事前学習に大規模なラベル付きデータが必要とされ，一般的ではない形式の画像を扱うことの多いリモートセンシング分野などでは，このデータセット準備の必要性が ViT の利用を阻むと考えられる．

そこで我々は，事前学習データセットの準備コストの低減につながる自己教師あり学習と ViT を組み合わせた Masked Autoencoders(MAE) [3] に着目する．MAE は encoder と decoder に ViT を用いた autoencoder であり，入力画像の再構築によって事前学習におけるラベル付きデータの準備を不要とする．また，入力画像の情報の一部を制限(マスク)することで計算量を削減できる．そのため，高いマスク率で分類精度を維持できれば，事前学習コストに加え，計算コストの低減も実現することができる．

本研究では，入力画像のマスク率を動的に変化させる学習手法を提案し，STL10 データセット [4] を用いた実験から，特に高いマスク率で事前学習した場合でのファインチューニング時の分類精度への影響を検証した．

2. 手法

2.1 Masked Autoencoders

MAE [3] は自然言語処理において高精度を示している masked language model [5] などに着想を得た画像処理における事前学習手法である(図 1)．入力画像を複数のパッチ画像に分割し，その中からランダムに抽出されたパッチに対してマスク処理を行い，マスクされていないパッチ(可視パッチ)のみを encoder へ入力する．encoder の可視パッチの出力とマスク処理を行ったパッチ(マスクパッチ)を decoder へ入力し，入力画像を再構築する事前学習を行う．この時，マスク率は常に一定で学習が行われる(静的手法)．

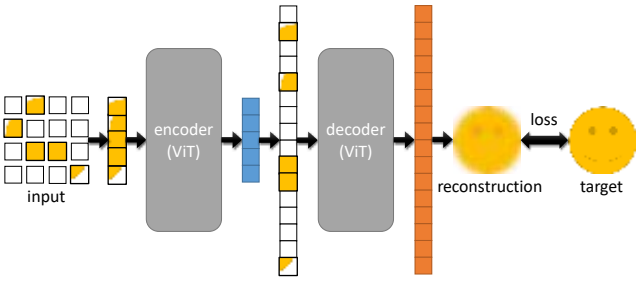

図 1　MAE の概要(静的手法)

2.2 マスク率の動的変化

本研究では，人間が教育システムを元に徐々に複雑な概念を学習することで効率的に学習する戦略を機械学習に取り入れたカリキュラム学習 [6] に着想を得て，MAE のマスクパッチの割合を動的に変化させることで再構築の難易度を徐々に上げる学習手法(図 2)を提案する(動的手法)．マスク率の割合が低い場合は再構築に用いるパッチ数が増えるため難易度が低く，反対に割合が高い場合は難易度が高いと考えられ，再構築の難易度を動的なマスク率によって設定した．

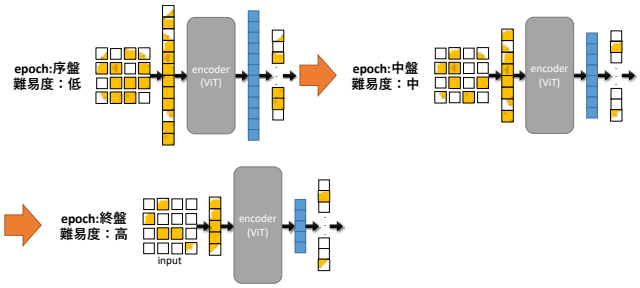

図 2　マスク率の動的変化(動的手法)

3. 実験条件

マスク率を固定とした静的手法とマスク率を動的に上昇させる動的手法の識別性能評価を行う為に STL10 データセット [4] を用いた実験を行った．

まず，既存の ViT-B [3] を参考に encoder と decoder block をそれぞれ 12 層，8 層とし，Attention のヘッド数を encoder では 12，decoder では 16 に設定し，埋め込み次元を 192 に変更したモデルを構築した．入力画像サイズ 96×96 の STL10 データセットに対して，パッチサイズを 6 に設定した(表 1)．

表 1　MAE モデルパラメータ

parameter	encoder	decoder
image size	96	96
patch size	6	6
patch embedding dimensions	192	192
depth	12	8
attention heads	12	16

次に，事前学習は表 2 に従う設定で，ラベルなしの訓練データ 10 万枚を用いて 100epoch 学習を行った．動的手法

†山口大学大学院創成科学研究科　Graduate School of Sciences and Technology for Innovation, Yamaguchi University

(dynamic)は 20epoch ごとにマスク率を徐々に上げて学習を行う 6 条件を定義した(表 3)．また，動的手法 6 条件の最終的なマスク率を学習初期から固定して設定する静的手法(static)を 6 条件定義する．各条件において 3 回試行し，学習済みモデルを作成した．再構築 Loss として 平均二乗誤差 を設定した．

最後に，事前学習したモデルの encoder に線形識別機を追加し，ラベルありの訓練データ 5000 枚を用いてファインチューニングを行い，検証データ 8000 枚に対する accuracy で評価を行った(表 2)．各学習済みモデルに対してそれぞれ 2 回試行した．

表 2　学習パラメータ

parameter	pre-train	fine-tuning
train data	100000	5000
validation data	-	8000
batch size	50	50
epoch	100	100
base learning rate	1e-3	1e-3
optimizer	AdamW	AdamW

表 3　動的手法の各条件におけるマスク率の遷移

マスク率 [%]	epoch				
	0~19	20~39	40~59	60~79	80~99
70	20	40	50	60	70
75	25	45	55	65	75
80	20	40	60	70	80
85	25	45	65	75	85
90	20	40	60	80	90
95	25	45	65	85	95

4. 結果

事前学習における各条件の最終 epoch の 再構築 Loss の平均値を表 4 に示す．最終 epoch のマスク率は静的手法と動的手法で等しいが，全てのマスク率において，わずかに動的手法の Loss が小さい．また，マスク率が高まるにつれ差分が大きくなっている．そのため，事前学習の入力画像の再構築において，マスク率を動的に上昇させる動的手法がわずかに優れていると考えられる．

表 4　事前学習の再構築 Loss

マスク率 [%]	手法		動的 – 静的
	静的	動的	
70	0.02804	0.02793	-0.00011
75	0.03286	0.03240	-0.00046
80	0.03980	0.03938	-0.00043
85	0.05001	0.04908	-0.00093
90	0.06670	0.06538	-0.00133
95	0.10222	0.09927	-0.00294

ファインチューニングにおける各条件の最終 10epoch の検証データに対する Accuracy を図 3 に示す．静的手法を用いた学習済みモデルにファインチューニングを行った場合，事前学習のマスク率 75%を頂点として，マスク率の上昇に伴い分類精度が下がる．一方，動的手法を用いた学習済みモデルにファインチューニングを行った場合，事前学習のマスク率が高い条件でも分類精度を維持された．マスク率

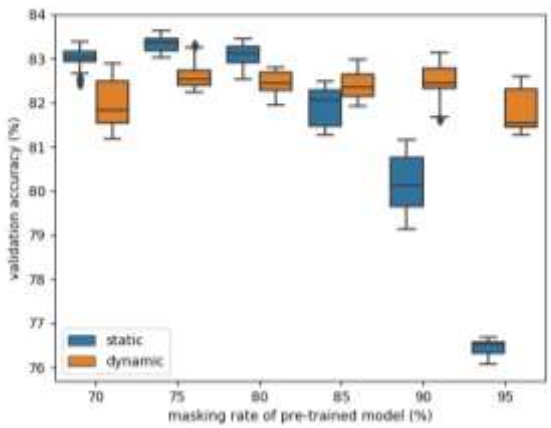

図 3　ファインチューニング時の Accuracy

85%を境に動的手法を用いた学習済みモデルが静的手法の学習済みモデルを上回り，マスク率が 85%以上では動的手法の分類精度が顕著に高い．

5. おわりに

本研究では MAE の事前学習においてマスク率を動的に変化させる学習手法を提案した．静的手法を用いた学習済みモデルでは高いマスク率の場合に識別能力が低下するが，動的手法を用いた学習済みモデルは高いマスク率においては，低いマスク率と同程度の識別能力を維持できることを示した．事前学習のマスク率が高いほど可視パッチを減少でき，encoder の計算量を削減できる．そのため，高いマスク率の場合に高い識別性能を示す動的手法は，MAE を用いた識別における計算コスト削減に有効であることが示唆される．

一方，ファインチューニング時のピーク性能は静的手法を用いた事前学習モデルの方が高い結果であった．そのため，動的手法を用いた事前学習モデルのピーク性能を改善させることが必要であると考える．

謝辞

本研究の一部は，JSPS 科研費 JP20H02417 の助成を受けて実施された．

参考文献

[1] Ashish Vaswani et al., “Attention Is All You Need”, Advances in Neural Information Processing Systems 30 (NIPS 2017), (2017).
[2] Alexey Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”, The Ninth International Conference on Learning Representations (ICLR), (2021).
[3] Kaiming He et al., “Masked Autoencoders Are Scalable Vision Learners”, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), (2022).
[4] Adam Coates, et al. “An Analysis of Single Layer Networks in Unsupervised Feature Learning”, Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, PMLR, Vol.15, (2011).
[5] Jacob Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, Proceedings of NAACL-HLT 2019, Vol.1, (2019).
[6] Yoshua Bengio et al., “Curriculum Learning”, Proceedings of the 26th Annual International Conference on Machine Learning, pp.41 – 48, (2009).