ISSN 2186-5647

2-53

MLP-Mixer のハイパーパラメータの最適化

日大生産工　○本橋　卓也　　日大生産工　山内　ゆかり

1. まえがき

近年、畳み込みニューラルネットワーク(Convolutional Neural Network: CNN)の構造が画像認識の代表的なモデルとして研究されている。しかし、最近の画像処理ではVision Transformer(ViT)[1]という畳み込みを用いずにattention機構を用いる構造が注目されている。このViTの構造が研究されていく中で、多層パーセプトロン(Multilayer Perceptron: MLP)のさらなる研究[2]が提言されてきている。TolstikhinらはMLPのみを用いたモデルとしてMLP-Mixer[3]を提案し、畳み込みやattention機構を用いたモデルと比較しても十分競合できるモデルであると報告されている。画像認識の分野では認識する対象の大きさは画像によって様々である。しかし、MLP-Mixerでは一部の重みが共有されているために、MLPのパラメータ数による特徴混合の性能が十分に発揮できていない。

本研究では、共有されている重みの設定を個別に行うことでMLPのパラメータ数による特徴混合の性能を十分に発揮させ、精度の向上を目指す。

2. 従来手法

2.1 畳み込みニューラルネットワーク

畳み込みニューラルネットワークとは人間の視覚をモデルに考案されたニューラルネットワーク(Neural Network: NN)である。基本的な構造は通常のNNに畳み込み層とプーリング層を加えたものとなっている。

$$f(x) = \sigma(b + \sum_{l=0}^{n} \sum_{n=1}^{\infty} \left(w_{ljm} + x_{ljm} \right)) \quad (1)$$

2.2 多層パーセプトロン

多層パーセプトロンとはパーセプトロンを何層にも重ねたものである。MLPは基本的に入力層、隠れ層、出力層から構成されている。ノードの層間にはそれぞれ重みパラメータが設定されており、入力が重みパラメータにより伝達されることにより、出力が決定される。

$$y = f\left(\sum_{i=0}^{h-1} w_i , x_i + b \right) \quad (2)$$

2.3 MLP-Mixer

MLP-Mixerは入力画像を画像パッチとして複数枚に切り分け、それらのパッチに対して空間方向とチャンネル方向にMLPで混合させる手法である。図1に基本的なモデルの全体像を示し、概要を説明する。

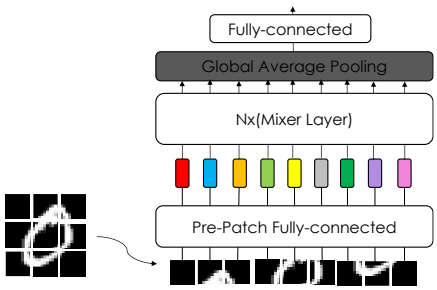

Figure 1　MLP-Mixer

初めに入力画像をP * Pの画像パッチに分割する。入力画像の画素数をSとして画像パッチの縦をH,横をWとし、式を次に示す。

$$S = HW / P^2 \quad (3)$$

次に、3次元の各パッチを線形変換して2次元のパッチに再形成します。入力画像をx、2次元のパッチをx_Pとし、式を次に示す。

$$x \in \mathbb{R}^{H \times W \times C} \rightarrow x_P \in \mathbb{R}^{N \times (P^2, C)} \quad (4)$$

その後、各パッチに対してMixer-layerを繰り返し行い特徴を学習させる。Mixer-layerについての全体像を図2に示す。MLPブロックについての構造を図3に示す

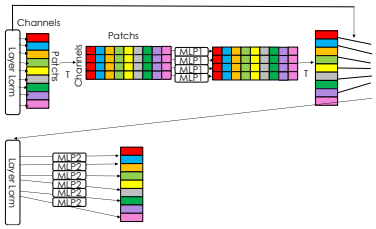

Figure 2　Mixer-layer

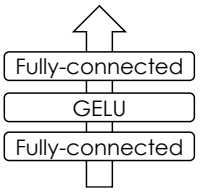

Figure 3　MLP ブロック

MLP-Mixer with Expanded Representation
by individual Weights

Takuya MOTOHASHI and Yukari YAMAUTI

Mixer-layerは上側のtoken-mixingと下側のchannel-mixing という2つの部分から構成されている。

token-mixingでは初めに各パッチに対して転置を行う。次に転置した行列をMLPブロックに入力する。MLPブロックから出力された行列を転置し元の画像パッチの形に直している。画像パッチの行列をXとし、MLPブロックの1回目の全結合層の出力をW_1、2回目の全結合層の出力をW_2とし、MLPブロック中の活性化関数GELUをσ、MLPブロックからの出力をUとし、式を次に示す。

$$U_{*,1} = X_{*,i} + W_2\sigma\left(W_1 LayerNorm(X)_{*,1}\right) \quad \text{for: } i=1\ldots C \tag{5}$$

また、MLPブロックで使用されている活性化関数GELUについて式を次に示す。

$$GELU = 0.5x + \left\{1 + tanh\left(\sqrt{2/\pi}(x + 0.44715x^3)\right)\right\} \tag{6}$$

channel-mixingではtoken-mixingからの出力をMLPブロックに入力する。MLPブロックから出力された行列を出力とする。画像パッチの行列をUとし、MLPブロックの1回目の全結合層の出力をW_3、2回目の全結合層の出力をW_4とし、MLPブロック中の活性化関数GELUをσ、MLPブロックからの出力をYとし、式を次に示す。

$$Y_{j,*} = U_{j,*} + W_4\sigma\left(W_3 LayerNorm(U)_{j,*}\right) \quad \text{for: } i=1\ldots S \tag{7}$$

事前に設定した回数N回までMixer-layerに画像パッチの行列を入力として繰り返す。

次に、Mixer-layerにかけ終わった画像パッチの各行列に対してGlobal Average Pooling(GAP)を行う。最後に、GAPで抽出した特徴を全結合層にかけることで出力が算出される

3. 提案手法

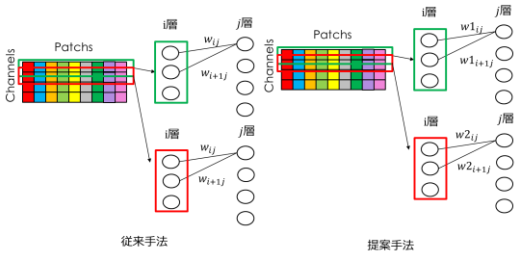

Fig.4 重みの設定

本研究では、特徴の混合を行うときにより細かく混合を行うため、MLP-Mixer内の全結合層間での重みを個別に設定することで精度の向上を目指す。従来研究では、MLPブロック内のFC層の構造がsingle-channel depth-wise Convolutionと1×1 Convolutionと同等のものであることから、層間の重みを共有していた。提案手法では全結合の重みパラメータを個別に設定することで、MLPが持つパラメータの幅による特徴判別の能力を活かす。

4. 実験方法および測定方法

Table.1 ネットワークの設定

	パターン1	パターン2
Number of layers	12	24
Patch resolution P×P	32×32	16×16
Hidden size C	768	1024
Sequence length S	196	196
MLP dimension DC	3072	4096
MLP dimension DS	384	512

提案手法の妥当性を図るために、従来手法と提案手法を同様のデータセットで画像認識の学習を行い、精度の比較検討を行う。学習及びテスト時に用いられるデータセットはCifar-10を使用し、学習用データ50,000枚とテスト用データ10,000枚を用いて学習及びテストを行う。また、ネットワーク構造は表1にある2つのパターンを実験で使用する。

5. まとめ

本研究ではCNNの１つのパターンの構造をMLPで表現した従来手法に対して、MLPの特性を活用することで精度の向上を目指した。

参考文献

1) Alexey Dosovitskiy,Lucas Beyer, Alexanbor Kolesnikov,Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthimer,…,Niel Houlsby “AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE“,arXiv:20929(2020)

2) Hanxiao Lio,Zihang Dai,David R. So,Quoc V. LePay “Attention to MLPs“,arXiv:2105.08050v1(2021)

3) Hya Tolstikhin,Neil Houlsby,Alexanbor Kolesnikov,LucasBeyer,XiaohuaZhai,ThomasUnterthimer,…,AlexeyDosovitsky“MPL-mixer:Anall-MLP Architectture for vision“,arXiv:2105.01601v4(2021)