2021 年 4 月 30 日

B4 伊藤 薫

進捗報告

1　今週やったこと

VisionTransformer の構造理解, グレースケール画像を含めた再学習と考察

2　VisionTransformer の構造

今までの ViT の実装で用いていたものは簡易な ViT を用いて, 一部 embedding に Conv2d を用いることで小型の ViT (DistillableVisionTransformer[1]) としていた (ResNet を蒸留できるモデルではあるが蒸留自体は使用していない. つまり Conv2d はいらないようなモデルとなっていたにもかかわらず入れていた. また, “ImageNet21k” を事前学習したモデルを転移学習している). 今回実装した ViT はシンプルに ViT-B_16 という ViT[2] を用いた (“ImageNet21k” を事前学習したモデルを転移学習している). またこの ViT は入力を画像のパッチにするのではなく, ResNet で得た特徴量を入力とする Hybrid Architecture も実装しているため, 今後試すのもあり. Hybrid Architecture はデータセットの規模が小さい場合にはわずかに ViT を上回り, 大きなものでは ViT のほうが良くなっていることが知られている. これは CNN が画像情報を捨象して要約することから, データセットが大きくなると必要な情報を捨ててしまう可能性を示している. 今回実装した ViT は“ImageNet21k” を事前学習したモデルを転移学習した ViT-B_16 というモデルであり, 基本的な構造は ViT[3] そのものである.

3　グレースケール画像を含めた再学習

まず初めに, 前回実装したテストデータにグレースケールを行ったものはコード中の画像の前処理に誤りがあったため結果が確かではなかった. 既に訂正済みである. 表 1 に訂正済みのコードを用いて訓練データを元画像とし, テストデータをグレースケール画像とした混同行列結果を示す.

今までの ResNet を蒸留した DistillableVisionTransformer, そして VisionTransformer を比較のため双方のネットワークで実行した. 以降, 名称を DistillableViT と ViT とする. 今回は訓練データにグレースケール画像, テストデータにグレースケール画像とし, 入力画像は 216 枚とした. 表 2 に DistillableViT を用いた縦軸を真値, 横軸を予測値とした混同行列結果を示す. 表 3 に ViT を用いた縦軸を真値, 横軸を予測値とした混同行列結果を示す.

表 1: 元画像の混同行列 (訂正済み)

真値	多義図形	50	11	11
	風景画	0	70	2
	肖像画	3	4	65
		多義図形	風景画	肖像画
		DistillableViT による予測値		

表 2: グレースケール画像の混同行列

真値	多義図形	55	7	10
	風景画	0	69	3
	肖像画	2	3	67
		多義図形	風景画	肖像画
		DistillableViT による予測値		

表 3: グレースケール画像の混同行列

真値	多義図形	61	4	7
	風景画	0	72	0
	肖像画	1	1	70
		多義図形	風景画	肖像画
		ViT による予測値		

表 2 より DistillableViT を用いた識別率は 88.4% となり, 多義図形の識別では 76.4% となった. 表 3 より ViT を用いた識別率は 94.0% となり, 多義図形の識別では 84.7% となった. 以上のことから訓練データをグレースケール画像とすると訓練データを元画像としていた場合の識別率 85.7% よりも高くなった. 多義図形の識別率に関しても 69.4% よりも高くなった. テストデータがグレースケール画像である場合, 訓練データをグレースケール化すると識別率の向上がみられた. ViT を用いた識別率 94.0% は訓練データ, テストデータ共に元画像とした場合と同等の識別率となった. 多義図形識別において DistillableViT より ViT のほうが 8.3% 高く, より良

い識別器と考えられる.

また, 訓練データを元画像＋グレースケール画像, テストデータをグレースケール画像とし, 入力画像を 216 枚とした実験を行った. 表 4 に DistillableViT を用いた縦軸を真値, 横軸を予測値とした混同行列結果を示す. 表 5 に ViT を用いた縦軸を真値, 横軸を予測値とした混同行列結果を示す.

表 4: グレースケール画像の混同行列

真値	多義図形	56	7	9
	風景画	0	71	1
	肖像画	0	4	68
		多義図形	風景画	肖像画
		DistillableViT による予測値		

表 5: グレースケール画像の混同行列

真値	多義図形	63	3	6
	風景画	0	72	0
	肖像画	2	1	69
		多義図形	風景画	肖像画
		ViT による予測値		

表 4 より DistillableViT を用いた識別率は 90.3 % となり, 多義図形の識別では 77.8 % となった. 表 5 より ViT を用いた識別率は 94.4 % となり, 多義図形の識別では 87.5 % となった. またしても ViT のほうが識別率が高く, より良い識別器といえる.

最後に訓練データを元画像＋グレースケール画像, テストデータを元画像とし, 入力画像を 216 枚とした実験を行った. 表 6 に DistillableViT を用いた縦軸を真値, 横軸を予測値とした混同行列結果を示す. 表 7 に ViT を用いた縦軸を真値, 横軸を予測値とした混同行列結果を示す.

表 6: 元画像の混同行列

真値	多義図形	53	10	9
	風景画	0	70	2
	肖像画	0	4	68
		多義図形	風景画	肖像画
		DistillableViT による予測値		

表 6 より DistillableViT を用いた識別率は 88.4% となり, 多義図形の識別では 73.6% となった. 表 7 より ViT を用いた識別率は 94.9% となり, 多義図形の識別では 88.9% となった. ViT を用いた識別率 94.9% は元画像を訓練データとテストデータとした DistillableViT を用いた識別率 93.1% を上回り, また, 元画像を訓練データとテストデータとした ViT を用いた識別率 94.0% をも上回った.

表 7: 元画像の混同行列

真値	多義図形	64	4	4
	風景画	0	71	1
	肖像画	1	1	70
		多義図形	風景画	肖像画
		ViT による予測値		

以上より ViT は DistillableViT よりも良い識別器であることがわかり, 訓練データを元画像＋グレースケール画像としたほうがテストデータの元画像の識別率が向上することが判明した. このことからグレースケール化は訓練データの DA として適切であることがわかった.

4 今後の方針

attention map の実装, 別の DA の実装, 実装コードの細かい調整 (識別率向上のため)

参考文献

[1] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

[2] VisionTransformer-PyTorch. `https://github.com/tczhangzhi/VisionTransformer-PyTorch`.

[3] 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.