2021 年 6 月 4 日

B4 伊藤 薫

進捗報告

1　今週やったこと

ViT との比較のため, CNN における Grad_Cam, Grad_Cam++ による視覚化

2　Grad_Cam

VGG-16 という CNN モデルを用いて, 元画像とグレースケール画像を学習した重みを使用した. 以降, ViT と CNN の混同行列及び識別率, 視覚化について比較していく. 表 1 に縦軸を真値, 横軸を CNN による予測値とした混同行列を示す. 表 2 に縦軸を真値, 横軸を ViT による予測値とした混同行列を示す. CNN における識別率は 91.2% であり, ViT における識別率は 94.9% である. また, 多義図形に関する識別においては CNN が 81.9% であり, ViT が 88.9% である. 識別率より ViT が 今回用いた CNN モデルよりも多義図形識別により良いモデルと言える.

表 1: CNN による混同行列

真値	多義図形	59	6	7
	風景画	0	72	0
	肖像画	1	5	66
		多義図形	風景画	肖像画
		CNN による予測値		

表 2: ViT による混同行列

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		

図 1 に CNN モデルにおいて元テストデータ及び, attention の視覚化である Grad_Cam, Grad_Cam++ の例を示す. 図 2 に ViT における元テストデータと attention map を適用した画像を示す.
まず, Grad_Cam と Grad_Cam++ に関して, より広範囲に多義図形を捉えているのは Grad_Cam++ と見て取れる. 図 1 の最下段の画像は一見すると図 2 の最下段の attention map と同等に見えるかもしれないが, 細かく見ると, Grad_Cam と Grad_Cam++ は左に移っている男性に attention がかかっている. その証拠に CNN モデルにおいて最下段の元テストデータは多義図形であるにも関わらず肖像画と誤識別している. これに対して, 図 2 より ViT では多義図形のみに注視領域があり, 実際に多義図形と識別していることから正しく識別できていると言える.

3　今後の方針

attention map カラー版の実装 (必要がある場合), attention map を layer ごとに視覚化 (説明に使うことがある場合), 他のモデル (VGG-19 等) でも試す, 多義図形画像を探す, 別の DA の実装

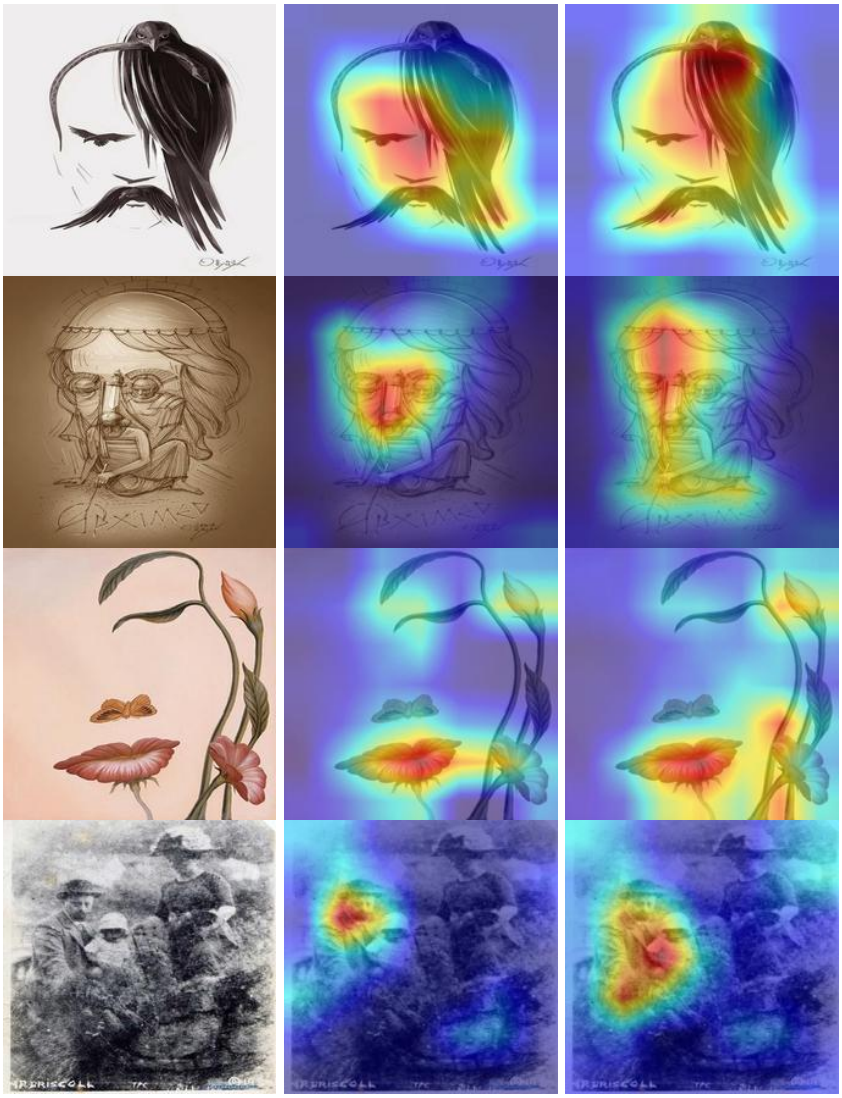

図 1: 左：元のテストデータ (多義図形), 中央：Grad_Cam 適用画像, 右：Grad_Cam++ 適用画像

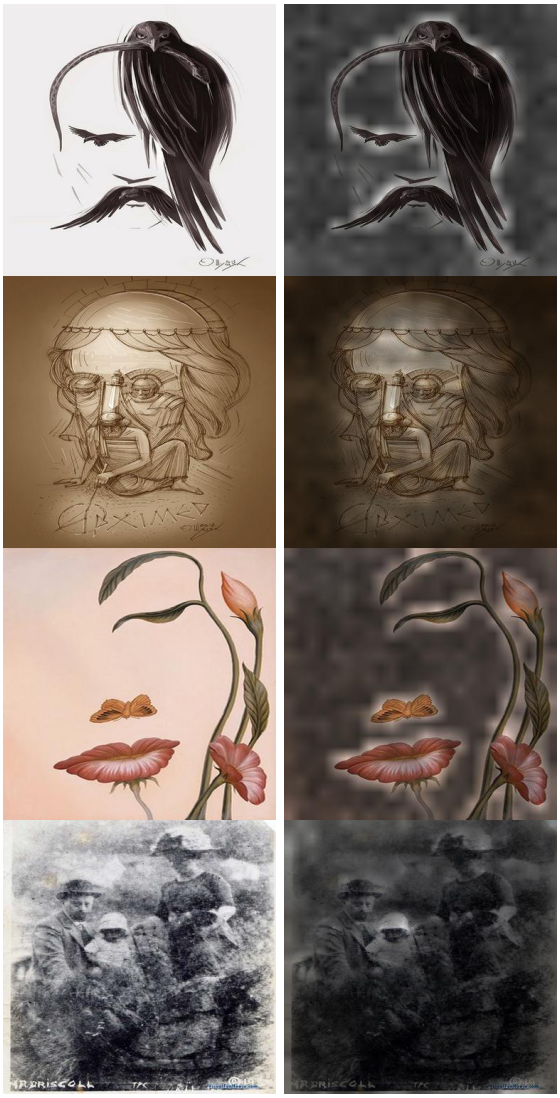

図 2: 左：元のテストデータ (多義図形), 右：attention map 適用画像