

## 進捗報告

### 1 SSIM と PSNR と RMSE による各単語にかけられた Attention の評価

表 1: 実験で用いたパラメータ

train data	826
eval data	551
test data	25
batch size	25
max epoch	600
snapshot_interval	50
optimizer	Adam
learning rate	0.0002
max_words_num	20
embedding_dim	300
captions_per_image	10
$\gamma_1$	4.0
$\gamma_2$	5.0
$\gamma_3$	10.0
attention image size	$272 \times 272$

表 1 に示した実験条件で、鳥画像での AttnGAN による Attention の観察を行った。図 1, 2, 3 に実験で得られた Attention の例 (GloVe と ConceptNet) と各単語への Attention 検証用自作データセットとの一部をそれぞれ示す。

図 3 にセグメンテーションデータを  $17 \times 17$  にリサイズし、skimage.transform.pyramid\_expand を用いて 16 倍に拡大したものを示す。これはもともとのサイズが  $17 \times 17$  である Attention を可視化する際の処理と同一の処理であり、'bird' という単語に対する Attention の理論上の最適解と同等であると考えられる。これを Attention との比較検証用データとする。

図 4～図 13 に各手法によって得られた epoch600 時点での Attention を可視化処理したものを示す。この処理の際に輝度を Attention として  $\text{attention} = (\text{attention} - \text{attention.min}()) / (\text{attention.max}() - \text{attention.min}())$  の式により、最小輝度を 0，最大輝度を 255 となるように調整している。

実験により得られた Attention と、CUB 公式のセグメンテーションデータおよび比較検証用データとの比較検証を SSIM と PSNR の値を用いて行った。図 14～図 21 に青いほど高い数値、赤いほど低い数値となるように Excel のカラスケールを用いて結果を示す。この時、epoch600 までに 50epoch ごとに snapshot したものの平均値とその中での最大値とを比較した。最大値を比較対象としたのは過学習を考慮したためである。

SSIM は構造的・部分的な類似度、PSNR は MSE を基にした再現度、RMSE は画素値の差分をそれぞれ示す値であり、得られた Attention と検証用自作データセットとを比較する指標となると考えられる。

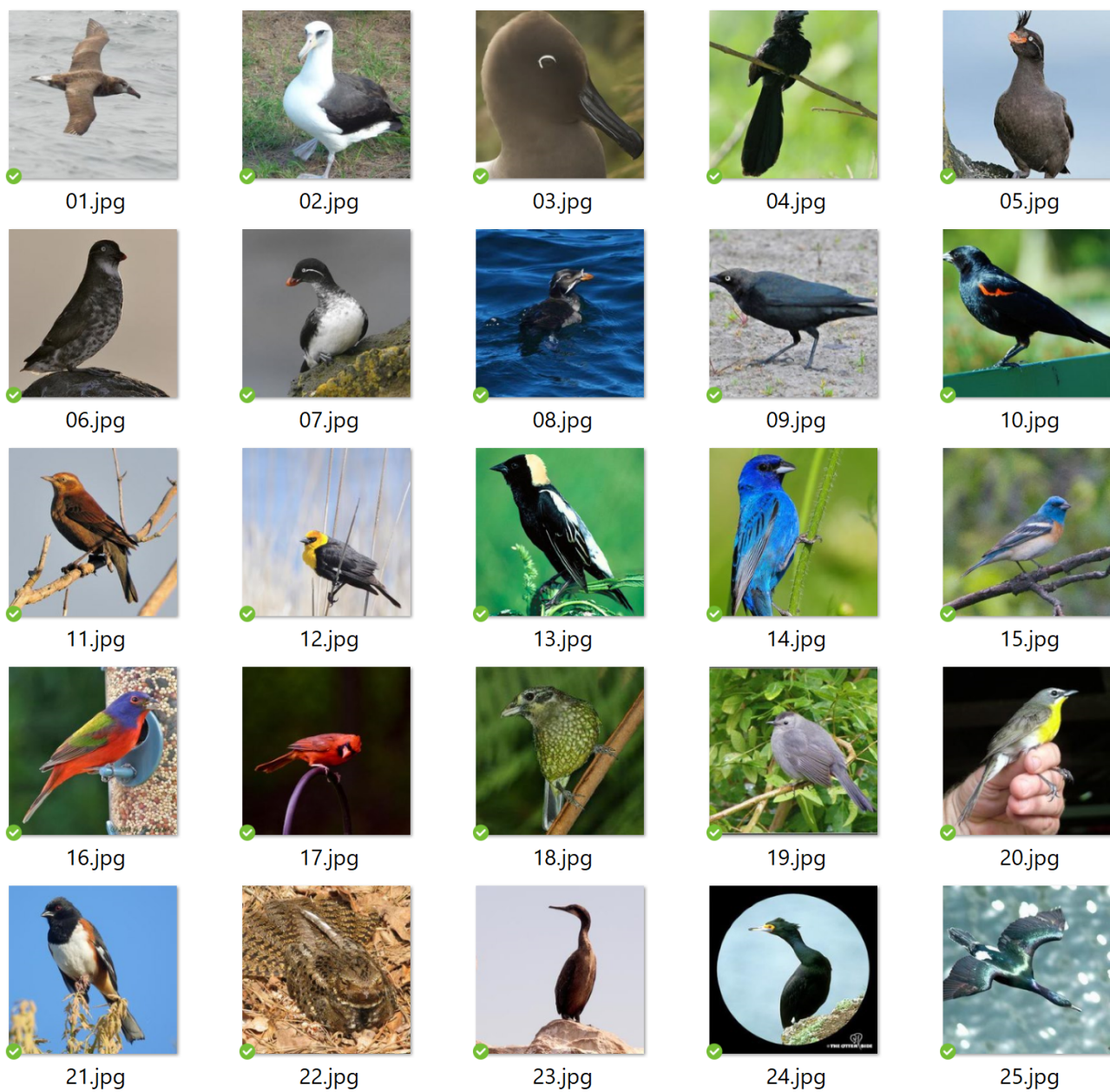


図 1: テスト用の鳥画像データ (正方形にクロップ済み)

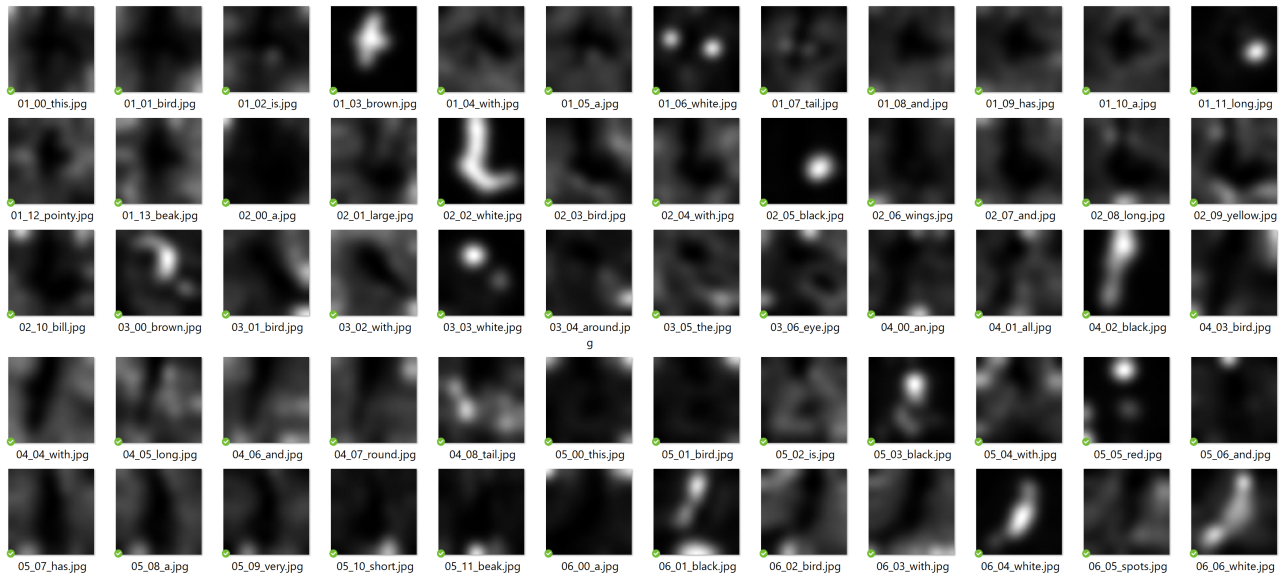


図 2: 実験で得られた各単語への Attention の例 (一部, 手法 : GloVe(LSTM))

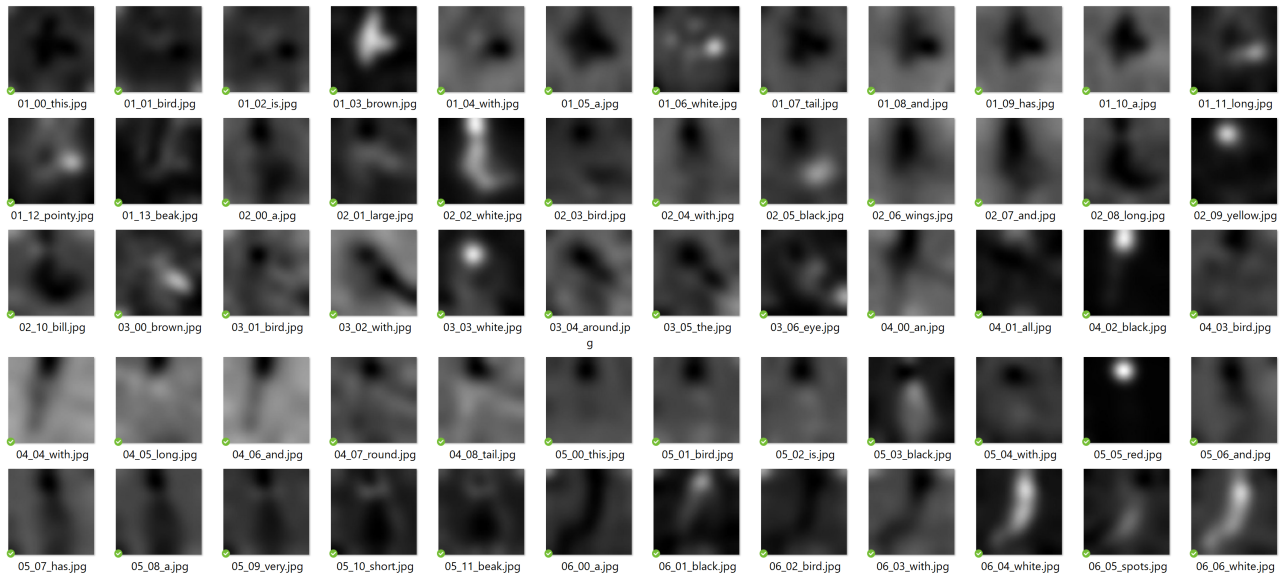


図 3: 実験で得られた各単語への Attention の例 (一部, 手法 : ConceptNet(LSTM))

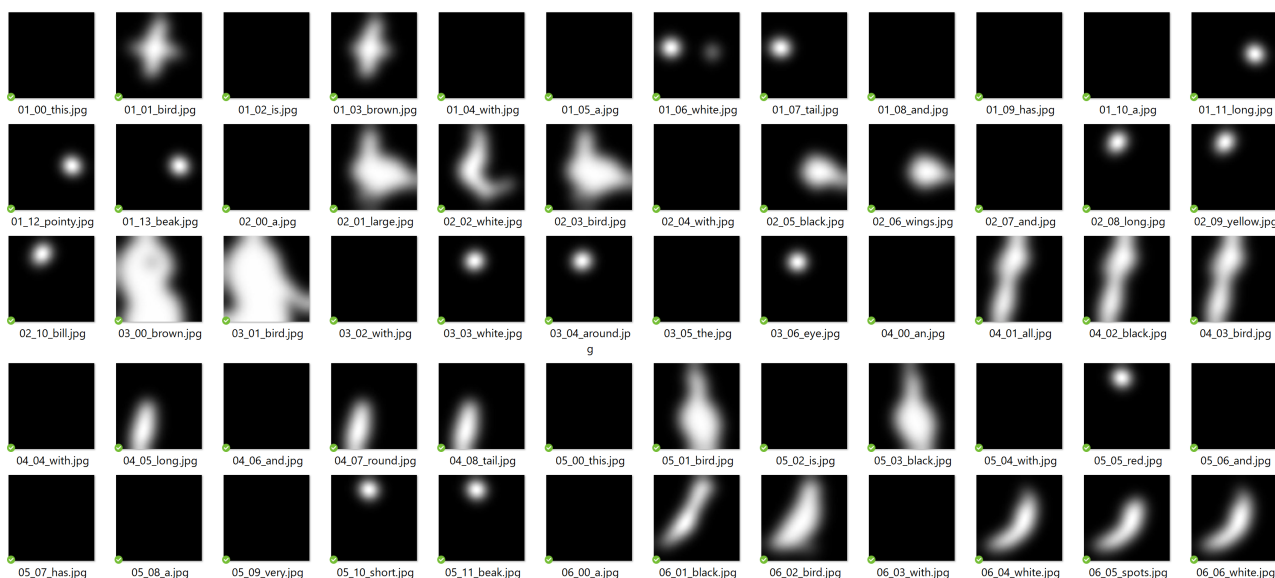


図 4: 各単語への Attention 検証用自作データセット (一部)

## 2 現段階での結果と考察

現段階での、全体における Attention の検証結果を「attn\_whole.xlsx」に、単語ごとの Attention の検証結果を「attn\_words.xlsx」に記す。

「attn\_words.xlsx」において、色を示す単語は紫色の太字で背景を黄色、stopwords(nltk に準拠)は薄紫のイタリック体で示している。

全体における Attention については、特に 12 と 17 の画像へは正しく Attention がなされ、特に 3 と 22 の画像へは誤った Attention がなされやすい傾向にあることが数値的に表れた。これらのことから、周囲と異なる色の部分を局所的に有していると、Attention の精度が上がしやすいと考えられるが、15 や 20 に対する精度が高くないため、一概に言える条件ではない模様である。

また、単語への Attention に関しては、色を示す単語に関して、強く Attention が付与されるということと、その精度に関して、具体的な数値を示すことができた。しかし、それ以外の単語に対する評価指標としては、改善の余地があると考えられる。

## 3 現段階での課題

- 「bird」などの一部の単語が stopwords と同等以上の高頻度語となっている
- 各手法による Attention の分散度合いに関する考慮はあまりしていない