

あなたの画像ツイートをお助け！

画像キャプション技術を応用した画像ツイート生成モデルの構築

概要

Encoder(CNN)-Decoder(LSTM)画像キャプションモデルを、Twitterから収集した画像ツイートのデータセットで学習し、画像ツイート生成モデルを構築。アンケートにより性能評価。

目的

画像の単なる状況説明でなく、ツイート閲覧者の共感を得る文章の生成



画像キャプション

- 状況の客観的で詳細な説明

例: 青い目をした猫がこちらを見つめている

ツイート風キャプションの特徴

- 状況を全て説明する必要がない
- 主観的な感想、ユーモアを含む
- 画像の中の世界に入り込んで良い

例: 子猫かわいい♡, じろり...

データセットの構築

Twitterデータセットの構築

Twitter APIにて指定した公開アカウントから画像ツイートを収集

データセット用アカウント選択にあたっての注意点

- 「動物」に関連するツイートを行うアカウントを選択
- ただし、以下3タイプのアカウントは除外
 - 画像に対して文章がないツイートをする
 - 画像からは分り得ない情報を含むツイートをする
 - 複数画像に同じ文章をツイートする

事前学習用データセット

STAIR captions [Yoshikawa+, 2017]

- 英語による画像キャプションデータセット COCO [Lin+, 2015]に対して、日本語話者が日本語キャプションを付与することで構築されたもの

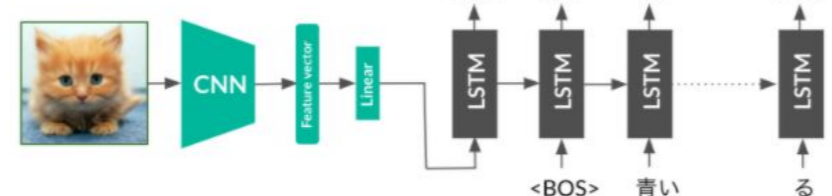
	STAIR	Twitter (訓練/検証/テスト)
データ数	820,310	4,420 / 620 / 10 (31アカウント)

実験

モデル

Encoder(CNN)-Decoder(LSTM)モデル [Vinyals+, 2015]を参考にPytorchで実装。
エンコーダはImageNetで事前学習したResnet152 [He+, 2015]で初期化。
デコーダでの推論時未知語 (<unk>) の出力確率を0に。

CNN-LSTMモデル



前処理

MeCabをトークナイザとして使用。MeCabの辞書として新語・固有表現に強いmecab-ipadic-NEologdを導入。使用語彙構築時にTwitterデータセットの語彙は低頻度語であっても全て残す。

学習の流れ

画像とキャプションの対応関係と言語モデルを学習させるため、STAIR captionsで事前学習しTwitterデータで追加学習。Twitterデータでの学習エポック数は検証データの結果をみて調整。

実験設定

	STAIR (事前学習)	Twitter (追加学習)
バッチサイズ	128	16
学習率	1e-03	1e-03
エポック数	10	2
LSTMの隠れ層の次元	512	512

最適化アルゴリズムはAdamを使用

実験コード: https://github.com/futakw/Twitter_Image_Captioning

結果

構築モデルの出力例

※Ground Truth: Twitterデータセットに含まれる実際のツイート内容

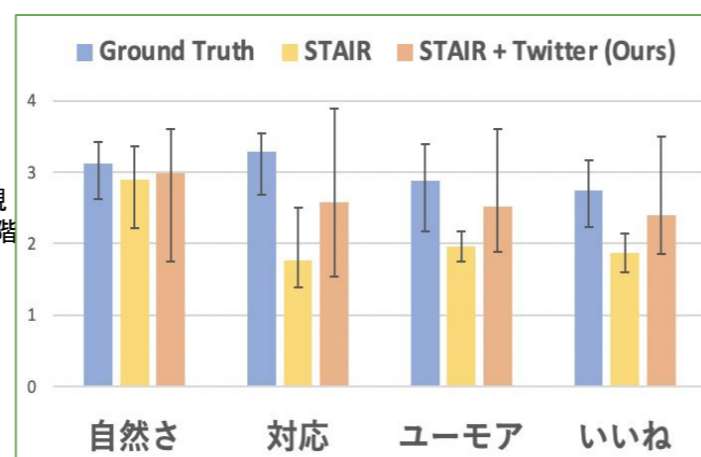


モデル	キャプション
Ground Truth	箱入り娘ですが...何か？
STAIR	猫がテレビの上に前足を乗せている
STAIR + Twitter (Ours)	猫の上目使いがかわいすぎる☆

アンケート調査

- 文章の自然さ(自然さ)
- 画像と文章の対応(対応)
- 文のユーモア(ユーモア)
- いいねしたいか(いいね)の観点で10サンプルに対して1~5段階評価。調査人数は35人。

※棒グラフ: 10サンプル平均値
エラーバー: 最大・最小値



考察


全ての評価観点で $STAIR < Ours < Ground Truth$ 。
STAIRのみで学習した場合と比較し、文章の自然さを保ちつつ、ツイートの自然さに適したキャプションを生成できているが、人間のツイートには勝てない。

※Twitterデータセットのみでの学習ではほとんど出力が得られず。データの質により画像と文章の対応や言語モデルが学習できないためだと考えられる。




課題と今後の展望

- ドメインが限定されている
 - イヌネコ以外の動物ではうまくいかないことが多い
 - =>データセットの偏り解消
 - =>他ドメインでも検証(例: 飯テロ、風景ツイート等)
- 定量的な評価が難しい
 - ツイート文は自由度が高く、多様性が許容されるため、特定の基準による定量評価ができない。エポック数も恣意的。
- 1つの候補しか生成できない
 - デコーダではGreedy Searchによる生成をしており候補のみ生成
 - =>ビームサーチ等の導入で複数候補を出力を可能にする
 - ツイート調(言葉遣い・方向性等)も限定的
 - =>ツイート調ごとに候補を出力する仕組みの導入

うまくいった例 1



	Ground Truth	STAIRのみ	STAIR + Twitter (ours)
画像ツイート			
文の自然さ	2.97	2.21	3.61
画像と文章の対応	2.69	1.39	3.11
ユーモア	2.40	2.04	2.86
「いいね」したいか	2.23	1.93	2.64

うまくいって例 2

	Ground Truth	STAIRのみ	STAIR + Twitter (ours)
画像ツイート			
文の自然さ	3.43	3.18	3.29
画像と文章の対応	3.54	1.61	3.89
ユーモア	3.26	2.07	3.61
「いいね」したいか	3.17	1.93	3.50

微妙な例 1

文が不自然

	Ground Truth	STAIRのみ	STAIR + Twitter (ours)
画像ツイート			
文の自然さ	2.97	3.04	1.75
画像と文章の対応	3.20	1.93	1.93
ユーモア	2.34	1.75	2.00
「いいね」したいか	2.29	1.71	1.89

微妙な例 2

画像にあまり対応していない

	Ground Truth	STAIRのみ	STAIR + Twitter (ours)
画像ツイート			
文の自然さ	3.00	3.14	2.76
画像と文章の対応	3.40	2.07	2.29
ユーモア	3.40	2.18	2.29
「いいね」したいか	3.14	2.14	2.14

その他の結果

Animal @!waaa! 1m
猫の表情☆



Animal @!waaa! · 1m
これは、僕のだよ。



Animal @!waaa!1m
僕の鼻を取っている



Animal @!waaa! · 1m
お風呂に入れてほしいた



Animal @!waaa! 1m
何かを見ている



Animal @!waaa! 1m
海の中で、何か？



その他の結果

Animal @!waaa! 1m
僕の鼻は、僕のだよ



Animal @!waaa! · 1m
飼い主のくせになっている



Animal @!waaa! · 1m
にゃおーん#猫



Animal @!waaa! 1m
お散歩になってもいい？



Animal @!waaa! · 1m
僕の鼻をつけてくる



Animal @!waaa! · 1m
僕の子



その他の結果

Animal @!waaa!1m
犬の口が開いている



Animal @!waaa!1m
お昼寝中



Animal @!waaa! · 1m
僕の顔を試してみた。



Animal @!waaa!1m
ペンギン親子



Animal @!waaa! · 1m
お散歩になるのか。



Animal @!waaa! · 1m
犬と一緒に寝ている犬



その他の結果

Animal @!waaa!1m
シロクマの親子



Animal @!waaa! 1m
お昼寝中



Animal @!waaa! · 1m
にゃおーん#猫





参考文献

- [1] Yoshikawa et al. STAIR Captions: 大規模日本語画像キャプションデータセット In NLP2017, 2017
- [2] Lin et al. Microsoft COCO: Common objects in context. In ECCV, 2014
- [3] Vinyals et al. Show and Tell: A Neural Image Caption Generator In CVPR, 2015
- [4] He et al. Deep Residual Learning for Image Recognition In IEEE, 2016

Twitter データセットの 例



ごめんなさい。。



馬の筋肉はほれほれしますね



明日はいいことあるよ☆



手のひらの上でぐっすり・・・



ハムスターのお尻が可愛すぎる。



三兄弟。



こういう状態を「ふくらすずめ」と言い、俳句の季語にもなっています。もちろん冬の季語です。。。羽の間に空気を入れた、防寒対策ですね。。。