ISS-P-009

位置関係を考慮した画像キャプション生成

守屋　響　　大和　淳司
工学院大学情報学部

1.　背景･目的

画像キャプション生成はコンピュータビジョンと自然言語処理を結びつけたものである[1]．画像を入力として，画像の中で行われている出来事，それに関わる人やものについて説明するキャプションを生成するタスクである．深層学習を用いた画像キャプション生成の精度が向上し，入力画像に対応したキャプションの出力が可能になっている．詳細なキャプションは画像内容の理解向上に繋がる．

本研究では画像キャプション生成における画像の詳細な描写のために，画像の中に登場する人やものの位置関係を含んだキャプションを生成することを目的としている．細かく描写するモデルでも画面内のどの位置に人物やものがあるのかを描写していないことが一般的である．位置関係を描写することで画像のより詳細な理解が可能になる．今回は左右の情報を描写するキャプションを生成することを目標とする．

2.　方法

位置関係を含んだキャプション生成は事前学習とファインチューニングの 2 段階で行なった．キャプションモデルの学習にはキャプションが既につけられているデータセットを使用することが一般的である．しかし，既存のデータセット Flickr や MS-COCO には目的としている位置に関する記述はない．そのため，新たに位置関係を追加したデータセットを作成し学習に使用する．

位置関係を含んだキャプションデータセットの作成は既にキャプションがつけられているデータセット, Flickr8k から画像を選び追記した．画像は 2 つの人やものが登場しているものを選択した．また，選んだ画像を反転して位置関係を逆にした画像も使用することでデータ数の増加と，左右のバランスを均等にした．

3.　学習

キャプションモデル学習にはエンコーダ・デコーダモデルを使用する. エンコーダには CNN ベースである EfficientNet[2] の Imagenet による学習済みモデル. デコーダには Attention のみを使い BERT など NLP モデルのベースとして利用されている Transformer[3]. 実行環境は Google Colaboratory を使用した.

Flickr8k を用いて事前学習を行ったモデルに左右の情報を追加したデータを 100 枚作成し，反転させたものを合わせて 200 枚で Epoch40 まで学習を行った.

4.　結果

学習を行なったモデルで図 1 の画像のキャプションを出力したところ下記のようなキャプションが生成された．犬が左でハードルが右にある画像であるが，犬の位置は正しいがハードルの位置については間違った記述をしている．また，図 1 は内容について説明できているが，テストした画像の多くは画像の内容と一致しないキャプションを出力していた．原因として考えられるのは学習回数の不足と，学習データの不足である．ファインチューニングにおいて学習データが少ないため物体の動作や位置のバリエーションが不足し，犬だったら左にいるとなっている可能性が考えられる．学習データを増やすことでバリエーションを増やし品質を向上させたい.

5.　まとめ

本研究では位置関係を含むキャプションの生成に取り組んだ．作成したデータセットを用いて学習をした結果，位置関係を描写できた画像もある．しかし，多くが生成したキャプションが画像の内容と一致しなかった．今後は画像のバリエーションを増やし複数の場面を用意することで画像の内容と一致する文の生成を行いたい.

また，評価指標の作成も課題である．現在は主観的な評価を行なっているが，定量的に評価する指標を作成する必要があると考えている．正解文章との類似度や位置情報の正解率を求め定量的な評価を行う予定である.

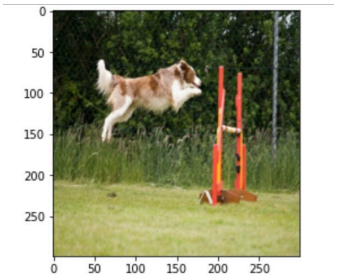

図 1 学習したモデルで生成したキャプション

出力：a dog on the left is jumping over a hurdle on the left

正解キャプション：A dog on the left is jumping over a gate on the right .

参考文献

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, ”Show and Tell: A Neural Image Caption Generator”
[2] Mingxing Tan, Quoc V. Le, ” EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”
[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, ” Attention Is All You Need”