画像質問応答に基づくキャプション生成への取り組み

理学専攻・情報科学コース 2140676 杜 静怡

1 はじめに

近年，生成されるキャプションを制御信号（コントロールシグナル）と呼ばれる情報を与えてキャプション内容を制御する研究が盛んになってきている．画像質問応答は，画像の内容に関する質問を与えることにより正解を導き出す技術である．質問は画像内に映る物体についての問い合わせ内容となっており，画像内の特定の情報に焦点を当てていると考えられる．このことから，本研究では VQA の質問をコントロールシグナルとして利用することを考え，画像質問応答に基づくキャプション生成を行う手法を開発する．

2 関連研究

近年，VQA のための注意ベースの深層学習法が多く提案されており，トップダウン型やボトムアップ型の注意法も含まれている． しかし，視覚的な質問に答えるには，視覚的な内容に関する情報だけではなく，常識も必要であり，人間が回答した限られた数の画像から直接に学習する他はないのが現状である． しかし，これまでの VQA 研究では，知識ベースを充実させることは比較的少なかった． その結果，生成されたキャプションは必ずしも質問に関連したものではなく，回答予測に必要な画像の特徴を無視している可能性がある．

最近の画像キャプションモデルも，注意ベースの深層学習モデルである． これらのモデルは，大規模な画像記述データセットからサポートを得て，顕著な結果を示している． しかし，ディープニューラルモデルは，依然として，最も重要なオブジェクトに基づくにもかかわらず、一般的なキャプションを生成する傾向がある． これまでの研究では，ターゲットを明確にしたキャプションを生成するよう促すキャプションモデルを構築してきたが，これらのキャプションは一般的に情報量が少なく，ほとんどのオブジェクトとその関係性について多様な説明を提供することができない． 本研究では，VQA 時に重要なオブジェクトに直接焦点を当て，VQA モジュールが答えを予測するのに役立つ情報を提供するキャプションの生成方法を考案する．

3 質問に基づくキャプション生成

図 1 に提案手法の概要図を示す．モデルは，テキスト情報抽出，画像情報抽出，相互注意メカニズム，回答分類，キャプション生成，物体検出の 6 つのモジュールから構成されている．特徴抽出部では，DistilBERT [1] と Vision Transformer [2] の 2 つから構成されている．DistilBERT は入力テキストの特徴を抽出し，Vision Transformer は入力画像の特徴を抽出するために使用される．Distill BERT モデルは，6 層から構成されており， 通常の BERT よりも層数が少なく，高速で，かつ精度が高いのが特徴である．Vision Transformer は，従来の CNN モデルよりも少ないパラメータで同じ精度を実現することができ，入力と出力が同じ次元であるため後続のモジュールの入力として利用しやすい特徴を備えている．相互注意メカニズムでは，テキスト特徴と画像特徴の相関を計算し，テキストから画像への注目特徴，画像からテキストへの注目特徴を獲得する．

デコーダ部では，回答選択モジュールへの入力は，テキストから画像への注目機構が出力する最初のトークンの特徴ベクトルであり，出力は質問に対する回答となる． キャプション生成モジュールにおいて，質問に基づいて画像内容に対応するキャプションを生成する．このモジュールへの入力は質問に対する画像の注目箇所における画像特徴ベクトルであり，出力は対応するキャプションとなる． オブジェクト検出モジュールは，入力された画像コンテンツとテキストコンテンツ情報に基づいて，画像内に映る物体のラベルを推論する．

4 実験

4.1 実験設定

データセット 実験では，COCO データセット[1]，および，それに基づき作成された画像質問応答用のデータセットである VQA v2.0[2]を用いた．質問に対するキャプションデータはないため，上記，2 つのデータセットを繋げ，質問に対するキャプションデータを作成した．該当データセットは以下の 4 つのステップで作成された．

Step1： COCO キャプションデータと VQA データにおいて，画像 id で両者のデータを照合し，キャプションと質問・回答の対応関係から，質問に対して適切なキャプションとみなせるものを選択し，質問-キャプションのペタデータセットを 2000 個生成した．

Step2： Step1 で作成した質問-キャプションのペアデータを正例とし，それ以外のペアデータを負例として，それぞれのデータを，訓練データ 8，評価データ 2 の割合で分割し，学習することによって質問に対して正しいキャプションを識別する識別器を構築した．black 識別器の精度は 0.81 となった．識別器の精度は十分とは言えないが，この識別器を用いることにより，質問に対する正解とみなせるキャプションの選別を容易にした．

Step3： Step2 で学習したモデルを用いて識別した質問とキャプションのペアデータに対して，手動でフィルターをかけ正解データの質を保つ． Step2 を繰り返し，学習データを増やす．

Step4： 全ての対象データをペアリングし，最終的に目的とする質問-キャプションの学習用ペアデータセットを取得する．

black 最終的に得られたデータセットの 2 割を評価用，8 割を訓練用に使用して学習を行った．

4.2 実験結果

表 1 は，本研究において作成したデータセットを用いて，コントロールシグナルを用いないキャプション生成システムである CATR [3]，質問応答システムである bottom-up [3]，提案モデルにおける質問をコントロールシグナルとして与えた際のキャプション生成の BLEU および質問に対する回答の精度（Accuracy）の

[1]https://cocodataset.org/home
[2]https://visualqa.org/
[3]https://github.com/saahiluppal/catr

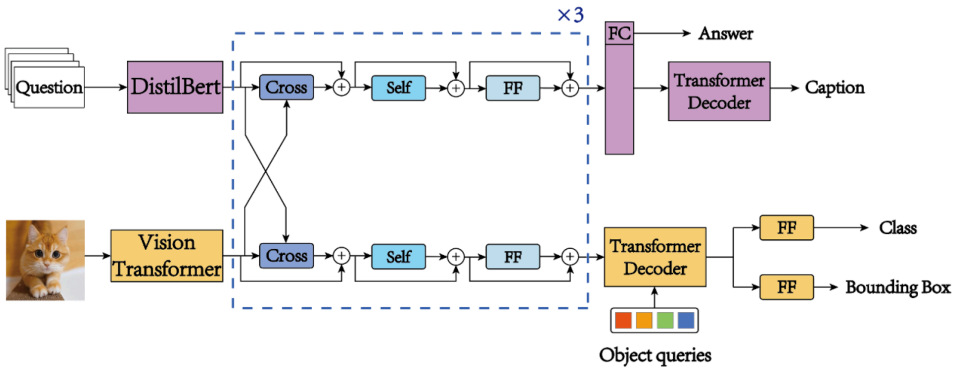

図 1: 質問に基づくキャプション生成

表 1: 比較結果

モデル	精度	BLEU スコア
CATR	-	31.1
bottom-up [3]on caption	-	31.9
Ours	-	32.4
bottom-up on VQA	71.3	-
Ours on VQA	72.4	-
Ours on VQA+caption	71.8	31.9

スコアである．それぞれのスコアが高いほど，アルゴリズムの性能が高いことを表す．

CATR は，画像キャプションに Transformer を導入した手法であり，本研究のベースラインとなる比較対象として取り上げる．画像質問応答の枠組みで質問をコントロールシグナルとして用いた Image caption に関する既存研究は少なく，bottom-up は本研究に比較的類似した方向性を有しているため，比較対象として選別した．bottom-up は，Faster R-CNN を用いて画像中の物体を認識し，その領域の画像特徴量を獲得，LSTM を用いて画像キャプションを行う手法である．

表 1 において，BLEU スコアだけ表記されているものは通常の画像キャプショニングを作成したデータを用いて行なった結果となっている．また，BLEU スコアがないものは質問応答における回答のみに対する精度を示している．

表 1 からわかるように，作成したデータセットにおいて，本手法は，画像に対するキャプション生成のみの単独の目的で学習させた場合，BLEU スコアが CART に対して，1.3，bottom-up に対して 0.5 向上した．また，画像質問応答を対象に，質問に対する回答のみの学習においては，bottom-up に対して 0.9 精度が向上し，回答およびキャプション生成をおこなった際にはキャプション生成においては，bottom-up は同じ BLEU スコアとなったが，回答においては 0.5 の精度向上が観測された．

質問応答のコントロールシグナルに基づきキャプションと回答の 2 つの損失を考慮して学習した場合，BLEU スコアと精度はそれぞれ 31.9 と 71.8 となった．2 つの損失を考慮した場合，キャプションを評価する BLEU スコアは少し下がってしまったが， buttom-up による回答の精度よりスコアが上昇していることより，提案手法は質問応答の回答の精度向上にも貢献していることを確認した．

図 2 に実験結果を示す．

図 2: 実験結果（Ours on VQA+caption）

5 おわりに

本論文では，画像質問応答における質問を制御信号として採用し画像に対するキャプション生成を制御する枠組みを提案した．COCO データセットおよび VQA v2.0 に基づき構築したデータセットを用いた実験において，提案手法がキャプション生成を評価する BLEU スコアと質問に対する回答の精度の両指標において比較手法よりも高精度となった．

参考文献

[1] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[2] 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.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.