ニューラルVQAのTOEIC写真問題への領域適応

Domain Adaptation of Neural VQA to TOEIC Photograph Questions

高里 盛良 Seira Takasato
三輪 誠 Makoto Miwa
佐々木 裕 Yutaka Sasaki
豊田工業大学大学院 工学研究科
Graduate School of Engineering, Toyota Technological Institute
{sd16417, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

1 はじめに

深層学習の発展により，機械翻訳や質問応答タスク等自然言語処理タスクの精度が大きく向上した．近年では，これら自然言語処理の手法が画像処理とも組み合わされるようになり，自然言語処理，画像処理を合わせたタスクの精度向上にも大きく貢献している．それらのタスクの１つに，画像質問応答 (Visual Question Answering; VQA) が存在する．これは画像と，その内容について尋ねる自然言語文での質問が与えられ，適切な回答を予測するタスクである．画像について様々な問題を設定することで，コンピュータの画像理解度を測ることができる．しかし，現在一般的に使用されているデータセットはコンピュータ向けに作られたもので，それが解けても実際にどれほどの理解力があるかの指標にはならない可能性がある．また，VQA において質の高いデータを大量に生成するコストは高く，新しい指標を作ることは容易ではない．

そこでこれまで我々はTOEICの写真問題を対象にキャプション生成モデルを適用してきた [7]．本研究では，人間向け問題を用いたVQAシステムの評価をし，VQAデータの人間向け問題への領域適応の可能性を明らかにする．

2 関連研究

2.1 画像質問応答

VQA [1] は与えられた画像とその内容に関する質問に答えるタスクであり，機械の画像理解の研究を促進するために提案された．単純な画像認識とは異なり，物体同士の関係を理由付けて答える必要があり，その精度はまだ人間に及んでいない．いくつかの問題形式が提案されており，その内の選択肢形式では画像と質問文，選択肢が与えられ，最も適するものを選ぶ．

タスクを解くためのVQAモデルは様々提案されており，多くがConvolutional Neural Network (CNN) とRecurrent Neural Network (RNN) を用いている．CNNは複数のフィルタを用いることで画像の特徴抽出を行う．RNNはニューラルネットワークを再帰的に接続した構造により時系列なデータ処理が可能となっており，文の特徴抽出を行う．CNNとRNNにより質問の画像と文が特徴ベクトルに変換されるため，これらの機構をエンコーダと呼ぶ．RNNとして，長い系列の記憶が可能なLong Short-Term Memory (LSTM) セルを持つLSTM-RNNが用いられることが多い．選択形式の問題の場合，得られた特徴ベクトルから，選択肢の質問に対する尤度を別のニューラルネットワークによって計算する．これをデコーダと呼ぶ．

2.2 領域適応

領域適応 (Domain Adaptation) は，他ドメインのデータが含む情報のうち，対象ドメインに対して有用な情報を抽出し学習への利用を試みる手法である．これによりデータ数が少ないタスクでの学習を補うことも可能である．

一般的に，モデルは学習ドメインのデータに特化するため，他ドメインデータでの評価精度は学習元のものよりも低くなる．そのためデータが大量にある場合にはそのドメインデータのみで学習することが望ましい．一方で対象とするタスクのデータが少ない場合や，異なるタスク間で共通の知識がある場合には領域適応が有効な手法の１つとされている．領域適応では，別

ドメインの大規模データを用いて共通した知識を学習しつつ，対象ドメインデータで調整し対象タスクに特化させることで，過学習を防ぎつつ学習を可能とする．領域適応におけるドメイン適応元をソースドメイン，適応先をターゲットドメイン，それぞれのドメインでのデータをソースデータ，ターゲットデータと呼ぶ．

近年では 2 つのモデルで互いに反する最適化を行う敵対的学習を用いる手法が提案されている．Ganin ら [3] は，分類問題においてエンコーダが 2 つのドメイン間の不変な特徴量を抽出可能となるように敵対的学習を用いた．抽出された特徴がソース，ターゲットどちらのデータから生成されたものかのドメイン予測を識別器と呼ばれるモデルが行い，その結果に対しエンコーダがドメイン予測の精度を 50%に近づけるように敵対的学習を行う．Chen ら [2] は，同じ構造を持つ 3 つのエンコーダを並列に並べ，2 つにそれぞれのタスク特化した特徴を，残りの 1 つにそれぞれのタスクに共通した特徴を抽出するよう学習する手法を提案した．共通な特徴抽出の強化には Ganin らと同様に敵対的学習が用いられている．

3 提案手法

本論文では 2 つの手法を提案する．ひとつは人間向け問題を用いて VQA モデルの評価を行う． もうひとつは既存の大規模データセットを領域適応により人間向け問題に適用し，データ数を補う．

3.1 人間向け問題での VQA モデルの評価

本手法では，標準的な VQA モデルの評価を人間向け問題で行い，画像理解力を検証する．また，既存の大規模データセットでも評価し，既存データがどれ程人間向けデータと差異があるのかを検証する．

人間向け問題には TOEIC のパート 1 を用いる．パート 1 は，4 つの短文が音声で流れ，与えられた写真について述べているものを選択する形式となっている．本研究では音声を対象とせず，選択肢はテキストで与える．パート 1 の問題形式は多くの VQA とは異なり問題文はなく，画像と選択肢のみが与えられる．本研究ではこれに合わせて VQA モデルの変更を行う．

モデルには 2.1 節で述べた CNN，RNN をもつモデルに対して，選択肢のみの選択問題を解くためのモデル構造の変更を行ったものを用いる．概要を図 1 に示す．本モデルでは次のように画像に I 対する N 単語から成る選択肢 S_c の尤度 $p(S_c|I)$ を求める．

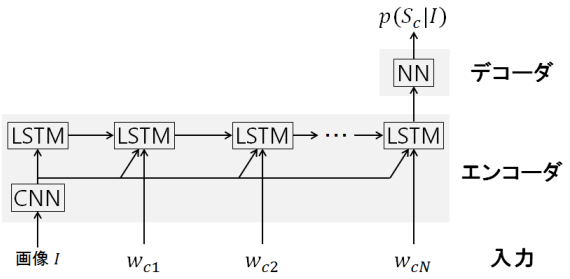

図 1: 問題文無し 4 択問題を解く VQA モデルの模式図

CNN によって画像の特徴を取り出し，それを RNN (本手法では LSTM-RNN を使用) の初期値 $\mathbf{h}_{c,0}$ とする．

$$\mathbf{I} = \mathrm{cnn}(I) \tag{1}$$

$$\mathbf{h}_{c,0} = f(\mathbf{I}) \tag{2}$$

ステップ t の RNN の隠れ状態 $\mathbf{h_{c,t}}$ はステップ t での単語 $w_{c,t}$ と $\mathbf{I}$ から得られる．W は行列を表す．OH は対応する単語の 1-0 ベクトルへの変換を表す．

$$\mathbf{h}_{c,t} = \mathrm{rnn}(\mathbf{w}_{c,t}) \tag{3}$$

$$\mathbf{w}_{c,t} = W_e \mathrm{OH}(w_{c,t}) + W_I \mathbf{I} \tag{4}$$

最終的な尤度は RNN の N ステップ目の隠れ状態をデコーダに入力したものとなる．

$$\mathbf{o}_c = W_h \mathbf{h}_{c,N} \tag{5}$$

$$p(S_c|I) = \mathrm{decoder}(\mathbf{o}_c) \tag{6}$$

他の選択肢にも同様の計算を行い，最も尤度の大きいものを正解とする．学習は正解選択肢の尤度を最大化するように行う．

3.2 領域適応による人間向け問題への既存データの利用

既存の大規模データを領域適応により人間向け問題と共に利用することで精度向上を図り，既存のデータがどれほど人間向け問題に利用可能かを検証する．領域適応には Chen らの手法を用いた．モデルを図 2 に示す．本提案手法では VQA モデルのエンコーダとして CNN と RNN を用いるため，それらを 3 つ並列に用い，それぞれのデータに特化するソースエンコーダとターゲットエンコーダ，共通して用いる共通エンコーダとする．デコーダには，ソースデータとターゲットデータの予測を行うソースデコーダ，ターゲットデコー

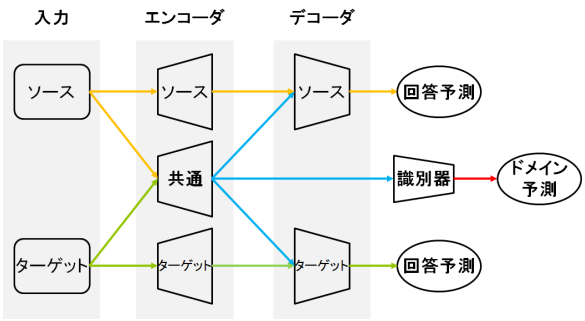

図 2: 領域適応を用いた VQA モデル

ダを用意する．ソースデータの回答予測は次のように行う．

ソースデータの画像 I_s，選択肢の 1 つ $S_{s,c}$ をソース，共通エンコーダ $\mathrm{encoder}_s$, $\mathrm{encoder}_c$ に入力し，それぞれの最終隠れ状態 $\mathbf{h}_{s,N}$, $\mathbf{h}_{c,N}$ を得る．

$$\mathbf{h}_{s,N} = \mathrm{encoder}_s(S_{s,c}, I_s) \tag{7}$$

$$\mathbf{h}_{c,N} = \mathrm{encoder}_c(S_{s,c}, I_s) \tag{8}$$

得られた 2 つの隠れ状態ベクトルを結合し，ソースデコーダ $\mathrm{decoder}_s$ に入力し，尤度を得る．concat はベクトルの結合を表す．

$$\mathbf{h}_{sc} = \mathrm{concat}(\mathbf{h}_{s,N}, \mathbf{h}_{c,N}) \tag{9}$$

$$p(S_{s,c}|I_s) = \mathrm{decoder}_s(\mathbf{h}_{sc}) \tag{10}$$

ターゲットデータの回答予測も同様に行う．

識別器は共通エンコーダの出力 $\mathbf{h}_{c,N}$ がソース，ターゲットデータどちらから生成されたものかを分類する．学習時は次の損失関数の最大化が行われる．

$$\max_{\theta^d} J^1_{adv} = \sum_{m=1}^{M}\sum_{i=1}^{N_m} \log p(m|X_i^{(m)}; \theta^d, \theta^c) \tag{11}$$

X は入力，M は全ドメイン，N_m は M のドメイン数，θ^d，θ^c はそれぞれ識別器と共通モデルのパラメータを表す．敵対的学習において共通モデルは識別器の分類精度が 50%になるように学習を行う．

$$\max_{\theta^c} J^2_{adv} = \sum_{m=1}^{M}\sum_{i=1}^{N_m} H(p(m|X_i^{(m)}; \theta^d, \theta^c)) \tag{12}$$

H はエントロピー誤差を表し p が 0.5 の時，すなわち識別器の分類精度が 50%の時に最大となる．

4 実験

本章では人間向け問題と領域適応を用いてコンピュータを評価する実験の設定と結果を述べる．

4.1 実験設定

人間向け問題には TOEIC Part1 の練習問題[1] (以下 TOEIC) 610 問を用いた．実際のテストでは音声で選択肢が読み上げられるが，本実験では音声を事前にテキストに変換し，画像とテキスト選択肢と答えを学習データとした．学習，開発，評価データそれぞれに 310 問，150 問，150 問用いた．

既存データは MS COCO (Microsoft Common Object Category) [4] を用いた．MS COCO は画像に対して 5〜6 個の説明文がついた大規模データであるが，これを，元から付与されている文の 1 つを正解とし，他の画像に付与されている文からランダムに 3 つ選び不正解とすることで，4 択問題に変換した (以下 COCO+)．

比較のため，3.1 節で述べた領域適応を行わない実験でのモデルの構造は 3.2 節で述べた領域適応を用いる実験のものと同じにし，敵対的学習を行わずに学習を行った．モデルの語彙は全て COCO+の学習データに含まれる単語の，頻度順上位 1 万単語を用いた．それ以外は未知語を示すタグに変換した．単語は全て小文字化し文の最後のピリオドは除去した．CNN には Antol らを参考に，Simonyan ら [5] の提案した，畳み込み層 16 層と 3 層層の全結合層を持つ VGG19 を用いた．Antol らは CNN に VGG19 より畳み込み層が 3 層少ない VGG16 を用いている．

モデルの学習の終了には早期終了を用い，最も高い開発スコアが次の 3,000 回の更新の間上回らなければ最も高い開発スコア時の結果を使用した．単語ベクトルは全て 512 次元，LSTM 隠れ層は全て 256 次元，デコーダの 2 層 NN は全て 256 次元とした．学習係数は TOEIC の学習で 3×10^{-6}，COCO，識別器では 10^{-3} とした．バッチサイズは全て 64 とした．識別器の学習では 1 回の更新で COCO+，TOEIC からそれぞれ半分の 32 ずつデータを選択した．

4.2 実験結果

最初に，提案手法で用いるモデルの一般的な VQA データでの評価結果を表 1 に示す．学習と評価データには Zhu らの Visual 7w [6] (7w) を用いた．比較として Antol らのモデルの評価結果 [6] を掲載している．7w は選択肢付き 4 択問題であるため，提案手法を用いる際は問題文の末尾に選択肢の単語を結合して 1 つの選択肢文とすることで，問題文無し 4 択問題へと変換した．7w には Zhu らが新たに提案した “pointing” と

[1] http://www.english-test.net/toeic/listening/#photographs

呼ばれる，画像内の領域を回答する問題形式も存在するが，両手法の結果はそれを除いたものとなっている．結果として Antol らの手法を上回り，提案手法には一般的な VQA を解く能力があることが確かめられた．

次に，TOEIC での評価結果を表 2 に示す．「人間」は人手による評価を表す．4 人の日本人により評価を行った平均は 86.0%となった．評価は音声をテキストにして行った．提案手法において，領域適応を行わずに COCO+のみを学習した結果の 24.7%はランダムで選んだ場合の 25%に近く， COCO+のみでの学習では TOEIC を全く解けないことが分かった．領域適応を用いると TOEIC のみで学習した場合より開発，評価データで 2.0%ポイント，1.3%ポイント結果が向上した．また最終的な結果の 39.3%は表 1 で示した一般的な VQA データでの評価値 56.2%よりも低くなった．

5 考察

TOEIC での評価が 39.3%であることより TOEIC は一般的な VQA データよりも難しいといえる．また COCO+で学習したモデルが TOEIC を全く解けていないことから，これらのデータの差異は大きいと考えられる．実際にそれぞれの 1 文当たりの単語数や画質，文形式が異なる．領域適応により結果が向上したことより，COCO+のデータを TOEIC の学習に適応できたと考えられるが，大きな向上ではない．原因は，TOEIC と COCO+の間に利用できる共通な知識が少なかったため，もしくは今回の手法ではそれを十分に抽出できなかったためと考えられる．COCO+で学習をしたモデルの TOEIC 評価が低いことからも，2 つのデータの差異が大きかったと考えられる．

6 おわりに

本研究では，コンピュータの画像の理解度を検証することを目的に，VQA モデルと領域適応を用いて人間向け画像付き問題である TOEIC を解いた．結果として，TOEIC が他の VQA データセットよりも難しい問題であること，今回の VQA モデルには TOEIC の画像練習問題を良い精度で解く能力がないことが示された．一方で COCO+から TOEIC への領域適応により精度向上が見られた．しかし，その向上は僅かであり，2 つのデータには共通部分が少ないという可能性を示した．今後は今回の結果について詳細な解析を行う必要がある．様々な VQA モデルを用いて今回の結果がモデルとデータどちらに依存しているかを検証すること，様々な学習データを用いて今回の領域適応の手法が VQA に対して有効であったか，COCO のどのようなデータが TOEIC の学習の助けになったのかを検証することを今後の課題とする．

表 1: VQA [1] と提案手法の Visual 7w での評価結果

手法	正答率 [%] 開発	評価
VQA [1]	—	52.1 [6]
提案手法 (領域適応 無し)	56.4	56.2

表 2: 提案手法の TOEIC での評価結果

手法	学習データ	正答率 [% (問)] 開発	評価
人間	—	—	86.0 (129)
領域適応 無し	COCO+	24.7 (37)	24.7 (37)
	TOEIC	41.3 (62)	38.0 (57)
領域適応 有り	COCO+ & TOEIC	**43.3 (65)**	**39.3 (59)**

参考文献

[1] Antol et al. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[2] Chen et al. Adversarial multi-criteria learning for chinese word segmentation. In *ACL*, 2017.

[3] Ganin et al. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, Vol. 17, No. 1, pp. 2096–2030, January 2016.

[4] Lin et al. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

[5] Simonyan et al. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[6] Zhu et al. Visual7W: Grounded Question Answering in Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[7] 高里ら．ニューラルキャプション生成モデルによる画像説明文の選択．言語処理学会第 22 回年次大会, pp. 103–106, 2016.