CH-006

視覚的質問応答のための敵対的学習を用いた多種質問解答の生成
Adversarial Training of Diverse QA Generator for Visual Question Answering

築山 将央* 伊神 大貴* 入江 豪† 相澤 清晴*
Masao Tsukiyama Daiki Ikami Go Irie Kiyoharu Aizawa

1. はじめに

近年，キャプション生成をはじめとする，画像と自然言語間のマルチモーダル学習を行う研究が注目されている．画像と自然言語の双方を扱うタスクとして，Visual Question Answering（VQA）が挙げられる [1]．VQA は画像とその画像に対する質問文が自然言語で与えられ，質問に対する正しい回答を自然言語で出力する問題設定である．VQA 自体のアーキテクチャや改善手法は数多く提案されているものの，VQA におけるラベル無しデータの活用は行われていない．また派生タスクとして Visual Question Generation（VQG）[2] が提案されたが，入力画像から質問と解答の双方を生成する手法は存在しない．

本研究では，VQA における半教師あり学習のために，画像に対して多様な質問解答ペアを生成するモデルの敵対的学習を行う手法を提案する．半教師あり学習とは，モデルの学習の際に正ラベル付きのデータに加えてラベル無しのデータを利用することで，前者のみによる学習よりも精度を高めることを目的とした手法である．図 1 に本手法のフレームワークの概観を示す．VQA においては，ラベル無しの画像に対して仮のラベル（Pseudo-Label）として生成した質問解答ペアを付与することで，合成データとして VQA モデルの学習に利用することが出来るようになる．また本手法では，強化学習の手法で用いられる勾配方策法（Policy Gradient）を取り入れることで，質問解答ペア生成器を学習させている．

本研究では敵対的学習による生成手法に加えて，ベースライン手法やキャプションデータセットを活用した手法から合成データを生成し，それらを VQA における半教師あり学習によって比較検討する．

2. 関連研究

Visual Question Answering（VQA）は，画像とその画像に対する質問文が自然言語で与えられ，質問に対する正しい回答を自然言語で出力する問題設定である．2015 年に提案された新しいタスクであり [1]，画像中のシーンを理解し自然言語を出力しなければならない点ではキャプション生成と共通しているが，質問文が与えられることによって，出力の制約がより強くなっている．

Visual Question Generation（VQG）は，入力画像に対応するような質問文を出力する新しいタスクである．VQG は VQA の逆問題として Mostafazadeh ら [2] によって初めて提案された．Jain らは Variational Auto Encoders（VAE）を用いて，画像に対して多様な質問文を生成する手法 [3] を提案している．この手法では潜在変数のサンプリング方法を様々に比較することによって，多様性と正確性を両立する質問生成器を実現した．ただし，これらの研究では解答の同時生成は行っておらず，学習データが少数のケースについては触れられていない．

また，Liu ら [4] は，VQA の派生タスクとして Inverse VQA (iVQA) を提案している．これは入力として画像と解答が与えられ，そのセットに対する適切な質問文を生成する問題設定である．iVQA は本研究で提案する質問解答ペア生成と近いタスクであるが，入力とする画像と解答の組み合わせによっては正確な質問生成が困難となる場合があり，今回のように一枚の画像に対して多様な合成ラベルを得たい場合には向かない手法と言える．

Chen らはある事前学習されたキャプション生成器を，勾配方策法を用いた敵対的学習によって，別のキャプションのドメインに適応させる手法 [5] を提案している．

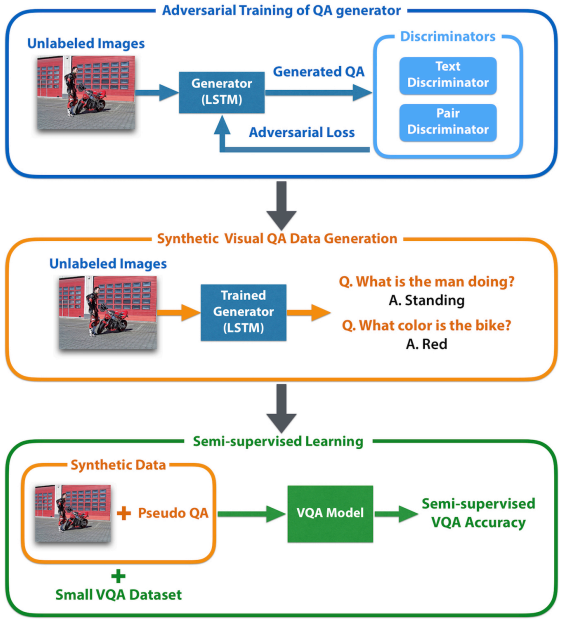

Fig.1: 本研究のフレームワークの概観

3. 提案手法

本研究では，VQA における半教師あり学習のために，画像に対して多様な質問解答ペアを生成するモデルの敵対的学習を行う手法を提案する．少数の正ラベル付きデータ（以下，Minimum Set とする）によって学習された生成器によって，ラベル無しの画像に対して

* 東京大学大学院 情報理工学系研究科 電子情報学専攻
† NTT コミュニケーション科学基礎研究所

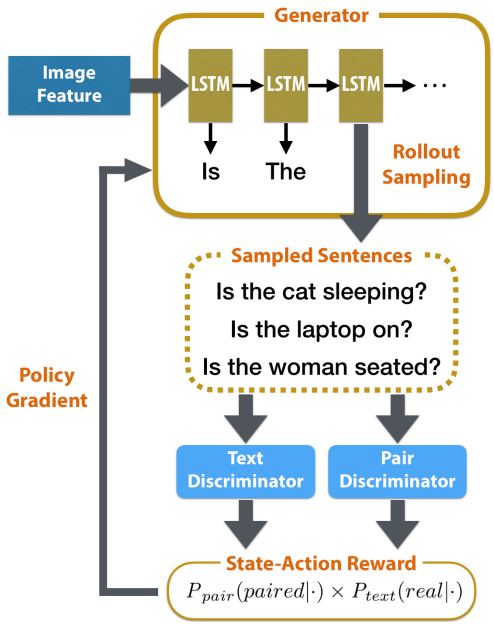

Fig.2: 本研究における敵対的学習の概観

Pseudo-Label として質問と解答を割り振る．これらの組み合わせを合成データとし，正ラベル付きデータに加えて VQA モデルに学習させることで，半教師あり学習を行う．

画像に対応する質問解答ペアを一連の学習によって生成するため，正ラベル付き学習データの前処理時に質問文と解答を結合している．この際，質問と解答の区切りとして固有のトークンを設けている．質問文を生成した後に固有トークンを出力し，その後解答を生成するという形式を容易に学習させる．

VQA における半教師あり学習の設定でモデルの比較評価を行うため，敵対的学習による手法に加えて，Long short-term memory（LSTM）のみを用いたシンプルなベースライン手法と，外部データとしてキャプションデータを利用した Dense Captioning による手法を用意した．以下ではそれらの手法について述べる．

3.1. 敵対的学習による生成手法

敵対的学習による質問解答ペア生成の手法の概観を図 2 に示す．本手法の敵対的学習では，画像に対する質問解答ペアの生成器と，そこから出力されるペアが VQA データの分布に沿っているかどうかを識別する分類器を交互に学習させる．すなわち Generator として質問解答生成器を用い，Discriminator として Generator の出力が妥当かどうかを見分ける二つの分類器を用いる．これらの詳細は後述する．敵対的学習を用いる利点は，Generator の学習を行う際の入力として Pseudo-Label を付与したいラベル無し画像を利用できる点と，損失計算の際のサンプリングによって，出力されるペアの多様性向上が期待できる点である．今回は敵対的学習が進むにつれて，Generator からラベル無し画像に対して VQA データの分布を再現しつつ多様な質問解答ペアが生成されることを期待する．

3.1.1.Generator

以下，Generator の入力となる画像を $\mathbf{x}$，出力となる固定長の文章を $\mathbf{y} = [y_1, ..., y_t, ..., y_T]$ とし，T は文に含まれる単語数，y_t は各単語であるとする．キャプション生成においてよく用いられる CNN-LSTM 型の生成器の学習では，目的関数 $J(\theta)$ は以下のように表される．

$$J(\theta) = -\sum_{n=1}^{N}\sum_{t=1}^{T} log\pi_\theta(\hat{y}_t^n|\mathbf{x}^n, \hat{\mathbf{y}}_{t-1}^n)$$

ここで $\hat{\mathbf{y}}_{t-1}^n = [y_1^n, ..., y_{t-1}^n]$ であり，$\pi_\theta(\hat{y}_t^n|\mathbf{x}^n, \hat{\mathbf{y}}_{t-1}^n)$ は，事前に $\mathbf{x}^n, \hat{\mathbf{y}}_{t-1}^n$ が与えられた際にパラメータ θ を持つ生成器が $\hat{y}_t^n$ を出力する確率である．また，$\hat{\mathbf{y}}$ は $\mathbf{x}$ と紐付いた文章の Ground Truth であり，N は総データ数を表す．本手法では文章生成器の最適化を敵対的学習を用いて行うため，目的関数として，強化学習の手法で用いられる勾配方策法（Policy Gradient）の目的関数を用いた．この目的関数は以下のように表される．

$$J(\theta) = \sum_{n=1}^{N}\sum_{t=1}^{T} E_{\mathbf{y}_t^n}[\pi_\theta(y_t^n|\mathbf{x}^n, \mathbf{y}_{t-1}^n)Q((\mathbf{x}^n, \mathbf{y}_{t-1}^n), y_t^n)]$$

ここで $Q((\mathbf{x}^n, \mathbf{y}_{t-1}^n), y_t^n)$ は State-Action Reward であり，State $(\mathbf{x}^n, \mathbf{y}_{t-1}^n)$ において Action y_t^n を選択した時の報酬値を表す．今回は文章生成器のケースを考えるため，Action を選択することは出力する単語を選ぶことに等しい．従って，Policy Gradient を用いた文章生成器の最適化では，各単語出力時の State-Action Reward がより高い値となるような，モデルのパラメータ θ を学習によって探索することを行う．State-Action Reward の設計については次節以降で述べる．

3.1.2.Text Discriminator と Pair Discriminator

本手法では，二種類の Discriminator を使用する．一つは Text Discriminator であり，入力文章が生成された偽のものであるか，そうでないかを判定する二値分類器である．もう一つは Pair Discriminator であり，入力の画像と文章ペアが正しく紐付いたものであるか，紐付いていないものであるか，それとも生成された偽のものであるかを判定する三値分類器である．

Pair Discriminator の学習では，Minimum Set 内の画像質問解答の組に $paired$ ラベルを，Minimum Set 内のランダムな画像と QA の組に $unpaired$ ラベルを，ラベル無し画像とそれに対する Generator の出力の組に $generated$ ラベルを付与し，Ground Truth として用いる．Pair Discriminator の Softmax 層の出力は $P_{pair}(class|\mathbf{x}, \mathbf{y}), class \in \{paired, unpaired, generated\}$ となる．各教師データに対する Softmax 層の Cross-Entropy を損失として，三値分類を正しく行えるように Pair Discriminator のモデルパラメータの最適化が行われる．

Text Discriminator の学習では，Minimum Set 内の質問解答文に $real$ ラベルを，ラベル無し画像に対する Generator の出力文章に $fake$ ラベルを付与し，Ground Truth として用いる．Text Discriminator の Softmax 層の出力は $P_{text}(class|\mathbf{y}), class \in \{real, fake\}$ とな

る．Pair Discriminator と同様に入力文章の二値分類を正しく行えるようにモデルパラメータが最適化される．

二つの Discriminator の出力から，Reward は $R(\cdot) = P_{pair}(paired|\cdot) \times P_{text}(real|\cdot)$ と表される．そして，これがバッチ中の全 State-Action について平均されたものが $Q(\cdot)$ となる．

$$Q((\mathbf{x}^n, \mathbf{y}^n_{t-1}), y^n_t) \simeq \frac{1}{K}\sum_{k=1}^{K} R([\mathbf{y}_{t-1}, y_t, \mathbf{y}_{(t+1):T_k}]|\mathbf{x})$$

ここで，K は Rollout Sampling の数である．Rollout Sampling では，各 State-Action から T_k まで生成される文章をサンプリングし，それらの Reward の平均をその時点の State-Action Reward として近似している．

3.2.Dense Captioning を用いた手法

Johnson らが提案した Dense Captioning[6] は，画像中のオブジェクトに対応する多様なキャプションを生成する手法である．この手法を用いてラベル無し画像に多種のキャプションを付与し，それらを質問文に変換することで合成データとするのが本手法である．Densecap の事前学習済みモデルは学習データとして Visual Genome データセット [7] の Region Descriptions を用いているため，本手法で生成された合成データによる学習は半教師あり学習ではないが，外部データが利用できる場合の精度を確かめるために今回の実験で実装を行った．Densecap によって生成される肯定文を質問文へと変換するため，今回は Ren らの手法 [8] 中で使われている方法を用いた．

4. 実験

本研究の実験では，提案手法の章で述べた各手法による半教師あり精度を比較する．各手法からラベル無し画像に対して質問解答ペアを生成させ，その組み合わせを合成データとして，Minimum Set に加えて VQA モデルに学習させた精度を半教師あり精度とする．本実験では VQA データセット [9] を用いた．本章では，本実験で用いたデータセット，VQA データに対する前処理，また各学習ステップで用いたパラメータについて述べる．

4.1. 実験の詳細

4.1.1. 前処理

本実験では敵対的学習による手法とベースライン手法において，VQA データ中の登場頻度が低い単語を削減し，LSTM を学習させる際の Vocabulary を 10,067 単語に固定した．この中には文章の開始を示す BOS（Begin-of-Sentence）トークンと文章の終わりを示す EOS（End-of-Sentence）トークン，質問と解答の区切りを示す固有トークンが含まれている．

また，ペアとして学習に利用できる VQA データ 658,111 件を Training 604,940 件，Validation 53,171 件に分割し，Training Set の内の 30,000 件を半教師あり学習で利用できる正ラベル付きの Minimum Set とした．最終的に合成データを加えた半教師あり精度の評価は Validation Set によって行った．

4.1.2. 事前学習

本実験で質問解答ペアの生成器として用いたのは，CNN-LSTM 型の文章生成器であり，CNN によって抽出された画像特徴量が LSTM の初期ステートに入力される形を取る．LSTM の Hidden State 数は 512 とした．また，LSTM の入力として用いる画像の特徴量は，事前学習済みの Resnet-101 による pool5 層特徴量とした．

最適化手法には Adam を選択し，Minimum Set によって学習させた．Minimum Set は事前学習に加えて，敵対的学習の際の Discriminator の学習にも利用している．この学習で Validation 精度の最も高いモデルを敵対的学習の事前学習済みモデルとし，ベースラインの手法とし本実験の比較に加えた．

4.1.3. 敵対的学習

敵対的学習においても最適化手法は Adam を用い，初期学習率は 5×10^{-5} とした．Rollout 数は $K = 3$ としたため，Batch Size を B，Sentence Length を S とすると，各バッチごとに $K \times B \times S$ 個の State-Action 値 $Q(\cdot)$ が計算される．

敵対的学習においては Generator と Discriminator の学習を交互に行う割合を，目的関数に対する最適化が安定するように選ぶ必要がある．本実験では Generator の学習を 1 iteration 行う間に，Text Discriminator と Pair Discriminator の学習を 20 iteration 行った．

4.1.4. 半教師あり学習

半教師あり学習を行い VQA 精度を比較するために，本実験では Bottom-Up Attention と Top-Down Attention を用いた既存の VQA モデルを利用した．

合成データを生成する手法との比較のため，上限値として VQA データセットの Training Set を全て学習させた精度と，Minimum Set のみを学習させた精度を求めた．ベースライン手法，敵対的学習による手法，Densecap による手法それぞれによって生成された合成データを，Minimum Set に加えて VQA モデルに学習させた精度を，各手法の半教師あり精度として本実験の比較に用いる．

ラベル無し画像と，それに対して生成された質問解答ペアの組み合わせを合成データとする．ここで言うラベル無し画像とは，Training Set に含まれ，かつ Minimum Set では使用されていない VQA データセット内の画像を指す．

4.2. 結果

本実験によって得られた半教師あり精度の比較を表 1 に示す．この表における Upper-bound とは，利用できる VQA データを全て学習させた場合の VQA 精度である．上限値と Minimum Set のみを用いた精度には 13% 以上の開きがあるものの，敵対的学習による合成データを用いた結果は Minimum Set のみを用いた精度を下回る結果となった．ベースライン手法による手法，敵対的

表 1: 各手法による VQA 精度の比較

Method	VQA Accuracy
Baseline	45.14
Adversarial Training	**46.17**
Minimum-Set Only	48.90
Densecap	**49.96**
Upper-bound	62.09

Fig.3: 各手法によって生成された質問解答ペアの例

学習による手法ともに，生成される合成データの解答に正しくないものが含まれていることが，学習時にノイズとして働いて精度を下げる原因となっていると思われる．加えて，敵対的学習による手法の現時点での問題として，Minimum Set に含まれない新規の質問文（Novel Question）を出力できる確率が低いことが挙げられる．各手法による合成データの多様性を定量的に比較することについては，今後の課題とする．

一方で外部データで学習された Densecap を用いた手法は，Mininum Set による精度を 1.06% 上回った．この手法では，外部データを有効活用した新規の質問解答ペアが精度向上に寄与していると考える．

各手法によって実際に生成された質問解答ペアの例を図 3 に示す．VQA データセットの質問解答ペアを比較すると，敵対的学習による手法では多様な質問解答ペアを生成できていることが見て取れるが，誤った解答も含まれている．Densecap を用いた手法では，画像に対して定型的でシンプルな質問と解答のペアを生成できており，多様性は低いが，解答の誤りは殆ど存在しなかった．そのため，敵対的学習による手法と比較して高い精度が観測された．

5. むすび

本研究では，VQA における半教師あり学習のために，画像に対して多様な質問解答ペアを生成するモデルの敵対的学習を行う手法を提案した．入力画像に対して複数の質問を生成する研究は既に幾つか存在するが，画像に対応する質問解答ペアを一連の学習によって生成する手法は存在しない．また VQA の分野における半教師あり学習の設定でモデルの評価を行う点も，本研究の新規性の一つである．

各モデルから生成された質問解答ペアを合成データとして半教師あり学習の精度を比較した．実験の結果，敵対的学習による手法はベースラインを上回ったが，正ラベルのみを用いた学習結果を下回る精度となった．一方で，キャプションが利用できることを仮定した Densecap による手法は，正ラベルのみの精度を上回った．

今後の課題としては，Densecap を用いた手法を手がかりにキャプションを有効活用した生成手法の提案や，質問により正確に対応した解答を生成する方法の検討，テンプレート生成によるタスクの細分化などが考えられる．また今回比較に用いた半教師あり精度の他に，Novel Question 数や Unique N-gram 数による多様性の比較も検討したい．

謝辞

本研究は，本研究の一部は CREST（JPMJCR1686）と NTT CS 研からの共同研究による支援を受けた．

参考文献

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pp. 2425–2433, 2015.

[2] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. In *ACL*, Vol. 1, pp. 1802–1813, 2016.

[3] Unnat Jain, Ziyu Zhang, and Alexander G Schwing. Creativity: Generating diverse questions using variational autoencoders. In *CVPR*, pp. 6485–6494, 2017.

[4] Feng Liu, Tao Xiang, Timothy M Hospedales, Wankou Yang, and Changyin Sun. ivqa: Inverse visual question answering. *arXiv preprint arXiv:1710.03370*, 2017.

[5] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *ICCV*, Vol. 2, 2017.

[6] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, pp. 4565–4574, 2016.

[7] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, Vol. 123, No. 1, pp. 32–73, 2017.

[8] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *NIPS*, pp. 2953–2961, 2015.

[9] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and Yang: Balancing and answering binary visual questions. In *CVPR*, 2016.