修士学位論文

Visual Question Answering のための データ拡張と多種質問回答生成

平成 31 年 1 月 31 日

東京大学大学院
情報理工学系研究科 電子情報学専攻

48-176419　　築山 将央

指導教員　　相澤 清晴 教授

内容梗概

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．VQA における現在の課題として，ラベル無し画像データの活用がなされていないこと，データ拡張手法が無いこと，自然言語のエンコーダに未だ再帰型ニューラルネットワークを用いていることが挙げられる．これらを踏まえ，本論文では VQA において半教師あり学習を用いてラベルなし画像の活用の知見を得ることと，VQA のためのデータ拡張を行い，精度を向上させることを目的とした二つの手法を提案する．一つは 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回答ペアを生成するモデルの学習を行う手法である．VQA においては，ラベル無し画像に対して生成モデルを用いて合成ラベルとして質問回答ペアを付与することで，半教師あり学習に合成データを利用することが出来る．この手法の新規性として，画像に対応する質問回答ペアを一連の学習によって生成する点や，質問回答生成手法の評価尺度として半教師あり精度を用いる点が挙げられる．実験結果では，キャプションデータを有効活用した方法によって，半教師あり精度を 1.06% 向上させることができた．もう一つは VQA モデルにおける質問文のエンコーダとして Transformer を取り入れ，学習時に質問文内の単語をランダムに予測させることでデータ拡張を行う手法である．この手法では質問と回答を結合したものをモデルの入力とし，事前学習の際にランダムに文章中の単語をマスクし，それらを予測させることでデータ拡張を行う．この手法の新規性として，VQA モデルの入力エンコーダとして Transformer を採用した点と，VQA の学習時に入力文章をランダムにマスクし，画像だけでなく自然言語側でもデータ拡張を行う点が挙げられる．実験結果では，VQA モデルに Transformer を組み込みランダムマスキングによる予測タスクを解かせることで，比較対象のモデルと比較して 1.24% の精度向上が確認できた．

目次

図目次

表目次

第 1 章

序論

1.1 背景

深層学習の登場により，コンピュータビジョンの分野は急激な発展を遂げた．畳み込みニューラルネットワーク (Convolutional Neural Networks, CNN) [6] を用いた様々な深層学習のアーキテクチャが研究され，それらは大規模な画像データセットを学習することで，画像分類のタスクにおいて高い精度をあげている．コンピュータビジョンの分野では画像分類を始めとして，物体検出，セマンティックセグメンテーション，超解像や動画における動き認識など，様々なタスクが取り組まれているが，これらは基本的に，いかにして画像ドメインの知識を獲得するかという問題に帰着する．一方で，シンプルな画像分類が成熟して様々なタスクが提案されると同時に，キャプション生成をはじめとする，画像ドメインだけでなく自然言語ドメインの知識が必要になるタスクが登場した．こういった，画像と言語のマルチモーダル学習が必要となるタスク群を本論文では Vision Language Task と呼ぶ．Vision Language Task は，ドメインごとのタスクに特化した深層学習モデルだけでなく，様々なドメインの知識を持ち，それらを相互的に活用して多様な問題を取り扱える汎用人工知能の基礎研究としても注目されている．

画像と自然言語の双方を扱う Vision Language Task の一つとして，Visual Question Answering（VQA）が挙げられる [7]．このタスクは入力として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．このタスクを解くために画像のみならず自然言語ドメインの知識も必要なことはキャプション生成と共通だが，VQA では画像を前提として質問文の意味を正しく理解し，正しい回答を生成しなければならないため，キャプション生成と比較してより強い制約が掛かっているといえる．また，キャプション生成の応用例としてはタスク名の通り画像に

対するキャプション付けや解説が挙げられるが，VQA は次世代インターフェイスにおけるインタラクティブなナビゲーションや，視覚障害者の支援などに応用ができる．

VQA 自体のアーキテクチャや改善手法は既に数多く提案されている．特に VQA データセット [7] での精度を競う VQA Challenge コンペティションは毎年開催されており，最近では画像特徴量と言語特徴量の Fusion に着目した手法や，VQA に適した画像特徴量を取り出すための手法が提案された [8, 5]．また，ある画像領域に対応した一部の画像特徴量を選択的に取り出すアテンションという機構は Vision Language Task 以外にも様々なタスクで活用されているが，画像アテンションに加えて質問文中のどの単語に注目するかをも考慮できる Co-Attention Network は VQA において高い性能を示している [9, 10]．一方で，VQA の分野の中であまり注目されていない領域もある．

一つ目は，ラベル無し画像データの活用である．例えば少数の正ラベル付きデータとその他のラベルなしデータを用いて学習する半教師あり学習のアプローチは，VQA の分野ではなされていない．VQA の設定で半教師あり学習を行うとすると，ラベル（画像に対する質問と回答のペア）がない画像に対して質問と回答を生成し，それらのペアを合成データとして学習に用いることになる．さらに，VQA の派生タスクとして Visual Question Generation（VQG）[11] が提案されたが，入力画像から質問と回答の双方を生成する手法は未だ存在しない．

二つ目は，Data Augmentation（データ拡張）である．データ拡張は深層学習モデルの汎化を促進するために学習時の入力となるデータに揺らぎを持たせるものである．コンピュータビジョンにおける深層学習の分野ではランダムクロップやランダムフリップ，ノイズ重畳等のデータ拡張手法がしばしば用いられるが，VQA 特有のデータ拡張手法は提案されていない．

三つ目は，自然言語のエンコーダである．VQA ではモデルの入力が画像と質問文となるため，画像は CNN を用いて画像特徴量にエンコードされ，質問文は LSTM や GRU を用いて言語特徴量にエンコードされる [12, 13]．一方で近年，自然言語処理の分野では Transformer というアーキテクチャが登場し，LSTM や GRU に取って代わりつつある．しかし，VQA モデルに Transformer を取り入れたモデルは未だ提案されていない．

1.2 目的

本論文では，Visual Question Answering において半教師あり学習を用いてラベルなし画像の活用の知見を得ることと，VQA のためのデータ拡張を行い，精度を向上させることを目的とする．そのために，二つの手法を提案する．

一つは 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回答ペアを生成するモデルの学習を行う手法である．改めて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，正ラベル付きデータのみによる学習よりも精度を高めることを目的とした学習設定である．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）として生成された質問回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．図 1.1 に本手法のフレームワークの概観を示す．質問回答ペア生成のため，本論文では幾つかのアプローチを行っている．具体的には，敵対的学習を用いて生成モデルを作る方法，キャプション生成の手法を組み合わせて質問回答ペアを作る方法，事前学習されたキャプション生成モデルを VQA ドメインに適応させる方法，そしてテンプレートベースで質問回答ペアを得る方法である．これらについては提案手法の章で詳しく述べる．

本手法の新規性として，画像に対応する質問回答ペアを一連の学習によって生成する点や，Visual Question Answering 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点に加えて，質問回答生成手法の評価尺度として半教師あり精度を用いる点が挙げられる．

結果として，半教師あり学習における様々な質問回答生成手法を比較することで，VQA におけるラベルなし画像データ活用の知見を得ることができた．生成手法によって生成される質問回答の質や内容は大きく異なるが，特に Dense Captioning[4] を用いた手法ではキャプションデータで学習されたキャプション生成モデルを有効に利用し，ノイズの少ない質問回答ペアを生成することができ，半教師あり精度の向上をもたらした．

もう一つは VQA モデルにおける質問文のエンコーダとして Transformer を取り入れ，学習時に質問文内の単語をランダムに予測させることでデータ拡張を行う手法である．Transformer はエンコーダ・デコーダ型のアーキテクチャだが，本手法ではこのエンコーダ部分を発展させた Bidirectional Encoder Representations from Transformers(BERT) アーキテクチャを質問文のエンコーダとして用いる [14]．また，既存の VQA 手法では学習の際に質問文と画像を入力してそれぞれエンコードし，回答を予測することでモデルの最適化を行うが，本手法では質問と回答を結合したものをモデルの入力とし，事前学習の際にランダムに文章中の単語をマスクし，それらを予測させることでデータ拡張を行う．

本手法の新規性として，VQA モデルの入力エンコーダとして Transformer アーキテクチャを採用した点と，VQA の学習時に入力文章をランダムにマスクし，画像だけでなく自然言語側でもデータ拡張を行う点が挙げられる．また今回，データ拡張手法の有効性を検証するために VQA 精度を用いているが，この手法は様々な Vision Language Task の

モデルに適用可能であり，汎用性が高い，

Transformer を用いたデータ拡張の比較実験では，大規模コーパスで学習された言語モデルは VQA においてもよりよい言語特徴量の抽出に役立つことがわかり，マスキングによる事前学習を組み合わせることでさらに VQA 精度を高めることができた．従来は GloVe 等の学習済み単語ベクトルを Embedding レイヤーの初期値とした LSTM や GRU が VQA モデルにおけるエンコーダの主流であったが，今回の結果によって BERT 等の Transformer 系言語モデルが VQA において高い性能を示すことがわかった．

本論文で述べる研究の貢献を以下に示す．

- VQA タスクにおいて，半教師あり学習の設定でラベルなし画像データを活用するフレームワークを提案した．
- VQA の半教師あり学習精度を，質問回答生成手法の新たな評価尺度として捉えた．
- キャプションアノテーションを活用した合成質問回答を生成することで，半教師あり VQA 精度を向上させた．
- VQA の分野で用いられていなかった言語モデル (Transformer, BERT) を取り入れ，精度を向上させた．
- VQA 以外の Vision Language Task にも適用可能な新たなデータ拡張の手法を提案し，さらに精度を向上させた．

1.3 構成

本論文では，VQA において半教師あり学習を用いてラベルなし画像の活用の知見を得るための手法と，VQA のためのデータ拡張を行い，精度向上を目指す手法を提案する．第 2 章では，本研究に関連する研究について述べる．主に Visual Question Answering の手法，画像に関連した自然言語を出力するモデル，VQA の派生タスク，敵対的学習，そして Transformer を始めとした言語モデルについて概観する．第 3 章では半教師あり学習のための質問回答ペア生成手法について述べる．生成から評価までのフレームワークと，今回比較のために用いた各手法について解説する．敵対的学習による生成手法に加えて，ベースラインとなる手法やキャプションデータセットを活用して合成データを生成した手法についても述べ，評価を行う．第 4 章では VQA のためのデータ拡張について述べる．具体的なデータ拡張の方法と Transformer を取り入れた VQA モデルについて述べ，ベースライン手法と VQA 精度を比較する．第 5 章では実験結果を踏まえた今後の課題と本稿のまとめを述べる．

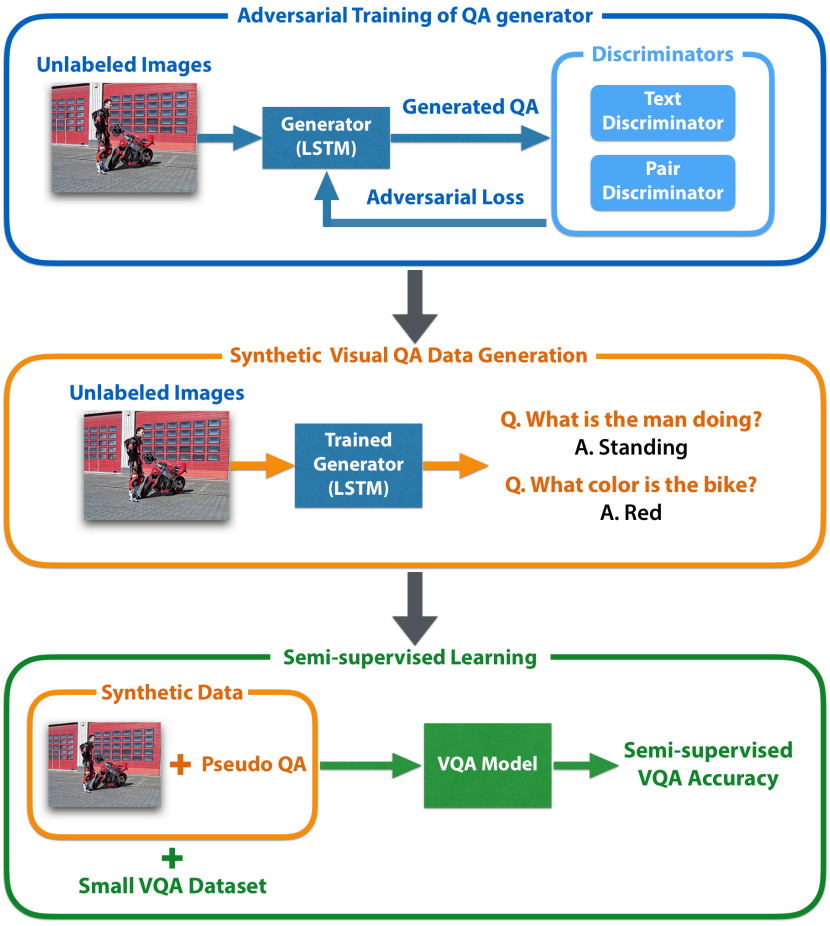

図 1.1: VQA における半教師あり学習のために，質問回答生成モデルを用いる我々のフレームワークの概要図．敵対的学習を用いてラベルなし画像に質問回答ペアを付与するモデルを学習させ，その出力を合成データとして半教師あり学習に用いる．敵対的学習の際に入力する画像群と，合成質問回答ペアを付与する画像群は同一のものである．

第 2 章

関連研究

2.1 Vision Language Task

コンピュータビジョンのタスクでは画像ドメインの理解が必要だが，それに加えて自然言語ドメインの知識も必要となるタスクを Vision Language Task と呼ぶ．これらのタスクを解くためには画像と自然言語間のマルチモーダル学習が必要となる．この節では代表的な Vision Language Task であるキャプション生成と，そのアプローチについて述べる．

キャプション生成とは，画像に対して自然言語による適切なキャプションを出力する問題設定である．入力は一枚の画像とし，出力は任意長（最大出力長はモデルによる）の自然言語のトークンとする．このタスクでは，画像分類や物体検出のように画像中に写っている物体を認識するだけでなく，それらの位置関係や画像全体の状況までを把握し，自然言語の形で出力する必要がある．そのためコンピュータビジョンの分野では，キャプション生成は深層学習モデルのシーン理解の一つの指標として研究が行われてきた．深層学習の登場以前はテンプレートを用いた手法 [15] が主であったが，深層学習の登場以後は CNN によって入力画像をエンコードし，LSTM 等に代表される再帰型ニューラルネット (Reccurent Neural Networks, RNN) によってキャプションをデコードするといった，エンコーダ・デコーダ型の手法 [16, 17] が高い精度をあげている．エンコーダ・デコーダ型のアーキテクチャの概要図を図 2.1 に示す．このような構造は，キャプション生成に限らず後ほど紹介する Visual Question Generation(VQG) を始めとする自然言語生成タスクで頻繁に用いられる．

Chen らは MSCOCO[18] のキャプションで事前学習されたキャプション生成器を，Policy Gradient[19] を用いた敵対的学習によって，CUB-200-2011 鳥画像データセット

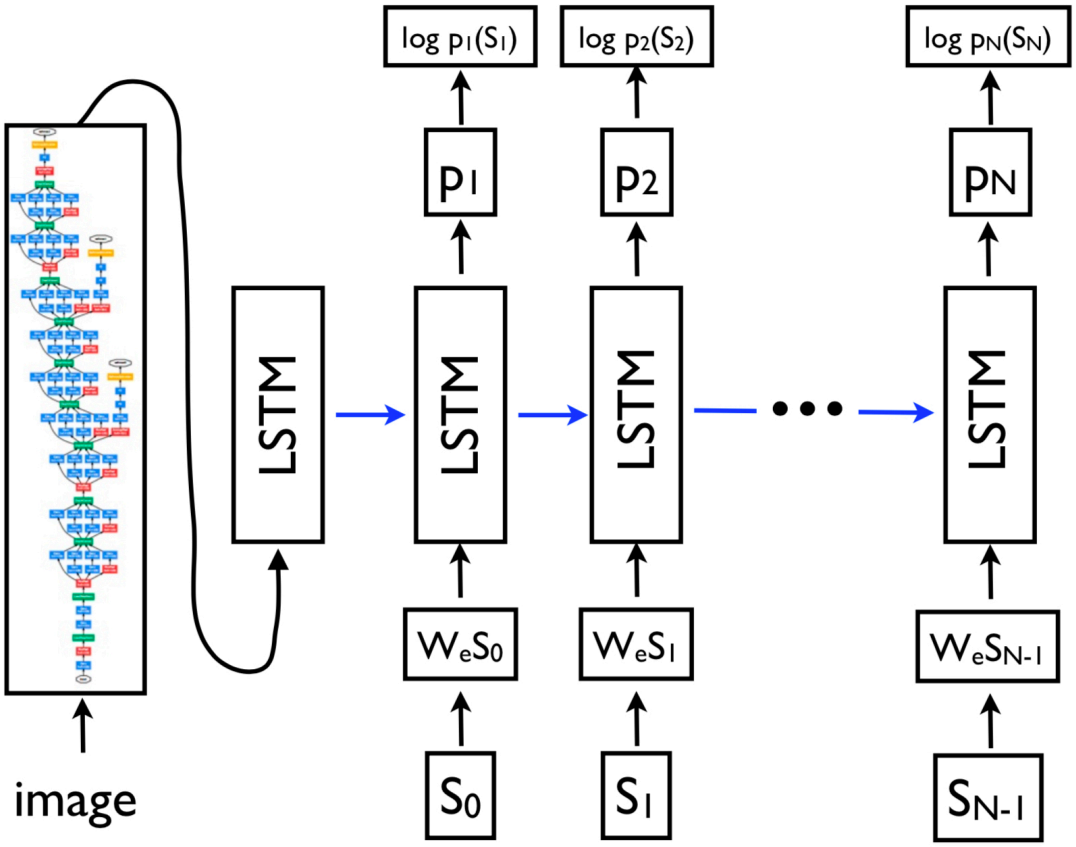

図 2.1: キャプション生成モデルの代表例であるエンコーダ・デコーダ型のアーキテクチャの概要図 [16]．CNN を用いて抽出（エンコード）した画像特徴量を RNN に入力し，自然言語を出力（デコード）するといった構造は，キャプション生成に限らず，Visual Question Generation など自然言語生成タスクで頻繁に用いられる．

[20] のキャプションのドメインに適応させる手法 [21] を提案した．この手法ではキャプション生成の各ステップに対して報酬を設定するため，学習時に Monte Carlo Rollout によるサンプリングを行っている．我々はこの手法における Generator の学習方法を参考にし，State-Action Reward の設計を行った．彼らはキャプション生成モデルのドメイン適応のために敵対的学習を用いているが，我々のフレームワークではラベルなし画像に対してより自然な質問回答ペアを付与するために敵対的学習を用いて LSTM を学習している．

Mathews らは，画像と対になっていないキャプションデータセットから，一枚の画像

に対して様々なスタイルのキャプションを生成できるモデルを学習する SemStyle を提案した [22]．彼らの提案した手法はエンコーダ・デコーダ型であり，まず画像を入力として幾つかの Semantic Terms（写っている物や状況を表すトークン）に変換するエンコーダ，Semantic Terms を入力としてスタイルに合わせた自然言語を出力するデコーダからなる．通常キャプション生成モデルは，画像とキャプションのペアでの学習が必要だが，彼らは Semantic Representation 空間を経由させることで対になっていないデータでのスタイル別キャプション生成を実現した．

2.2 Visual Question Answering

Visual Question Answering (VQA) とは，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 タスクの概要図を図 2.2 に示す．2015 年に提案された新しいタスクであり [7]，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．VQA には，Multiple-Choice 型と Open-Ended 型という二つのタイプが存在する．Multiple-Choice 型は短文からなる回答の候補が事前に一定数用意されており，その中から正しいものを選択する設定であるため，一般の画像分類と同様に多値分類問題に落とし込むことができる．Open-Ended 型は回答文の候補が用意されておらず，任意の回答文を生成する必要がある．本研究では，これらのうち Open-Ended 型の VQA を扱う．

VQA の手法群を紹介する準備として，VQA を解くための一般的なモデルの構成について述べる．他の Vision Language Task と同様に，まず画像と言語の特徴量を抽出するエンコーダをそれぞれ用意する．次に何らかの方法でそれらの特徴量を結合し，出力となる回答を得るデコーダが続く．Open-Ended 型の VQA ではデコーダとして自然言語をステップ毎に出力する RNN を選ぶことも考えられる．しかし，VQA で高い精度を上げている手法群ではデータセットによく登場する回答を数万個あらかじめ固定し，それらの多値分類問題を解かせるというアプローチを取っている．そのため，特徴量結合後のデコーダ部分は Fully-Connected 層が連なる場合が多い．ほとんどの VQA 手法は，回答の出力を自然言語文章の予測として行っておらず，全回答候補から正しいものを選択するという多値分類問題として扱っており，そのアプローチの方が精度も高い．

現在 VQA のタスクで State-of-the-Art となっている手法群は，画像特徴量を抽出する際に画像中の物体領域 (Object Region) に注目した Bottom-Up Attention を用い，言語特徴量の抽出には LSTM もしくは GRU を用いているものが多い．さらに，最新の手法

群は画像特徴量と言語特徴量の結合部分に重きを置いている．

Gan らは，質問に答える手がかりとなる画像中の物体領域のピクセルワイズアノテーションを独自に収集し，それを用いて VQA モデルの学習時に明示的にアテンションを貼る VQS という手法を提案した [23]．Vision Language Task で用いられるアテンションマップは通常，学習の過程で暗黙的に最適化されるが，彼らは VQA において，答えとなる部分のセグメンテーションマスクがアテンションマップの教師データとなることを仮定し，学習時に明示的に最適化を行った．また彼らは VQA の発展タスクとして，質問文と画像を入力として答えとなる部分のマスクを出力する Question-Focused Semantic Segmentation(QFSS) という新たなタスクも提案した．

Kim らが提案した Bilinear Attention Network(BAN) は，画像特徴量と言語特徴量双方にアテンションを貼る Co-Attention の計算量を削減するために Low-rank Bilinear Pooling[25] を取り入れ，さらにアテンションマップの数を増やすために残差学習を行ったものである [5]．彼らの Multimodal Residual Networks（MRN）では，最大 8 つの Co-Attention Map を使い学習することが可能となり，VQA 精度を改善した．

Johnson らは，Neural Module Network (NMN)[26] を VQA に応用し，質問文を機械的にパースして VQA のタスクを複数のサブタスクに分割して解く手法を提案した [27]．また彼らは NMN を用いた手法の評価のため，3DCG を用いた CLEVR という合成データセットを公開しており，特に画像中の物体の形，色，そして位置関係を問う問題は，NMN の適用によって解きやすくなっていることを示した．CLEVR に含まれる質問回答ペアは上記カテゴリのものに限定されており，質問と回答は 3DCG 画像とともに全て機械的に生成されている．

2.3 VQA の派生タスク

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 の発展タスクとして Mostafazadeh ら [11] によって初めて提案された．前節で述べたように VQA の手法は回答の予測を多値分類問題として扱っているが，VQG では出力部分は RNN を用いて一連の自然言語トークンの形で出力する必要がある．そのため VQG は VQA と比較すると制約が緩く，正確な評価も難しいタスクである．以下に紹介する VQG の手法は基本的に画像と回答文が与えられ，妥当な質問文を生成するという設定で行われるが，我々のフレームワークではラベルなし画像に対して多種の質問と回答のペアを生成する必要があるため，入力は画像のみという設定である．

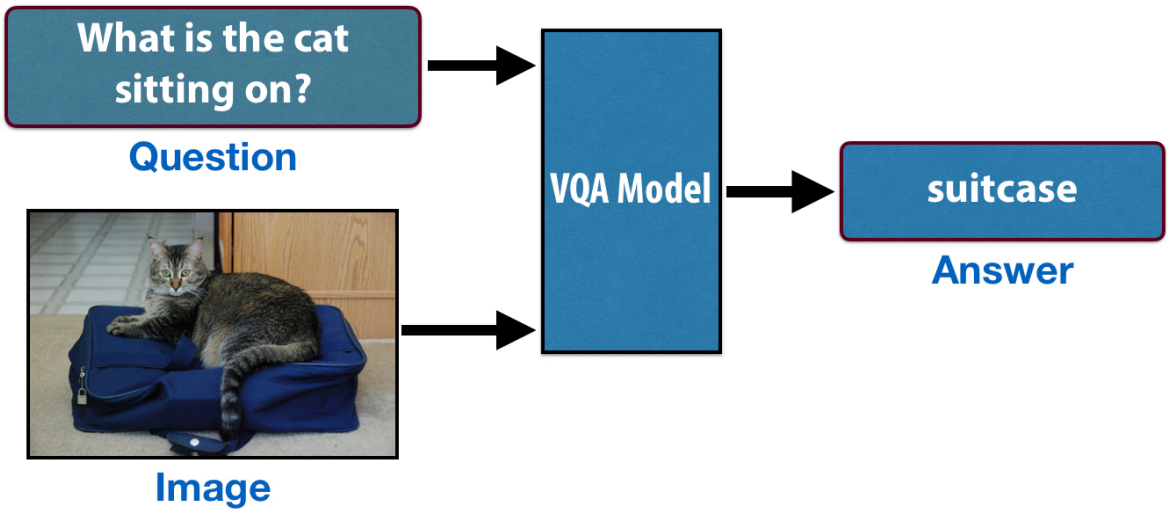

図 2.2: Visual Question Answering(VQA) の概要図．質問文と画像を入力とし，モデルは妥当な回答を予測することが求められる．

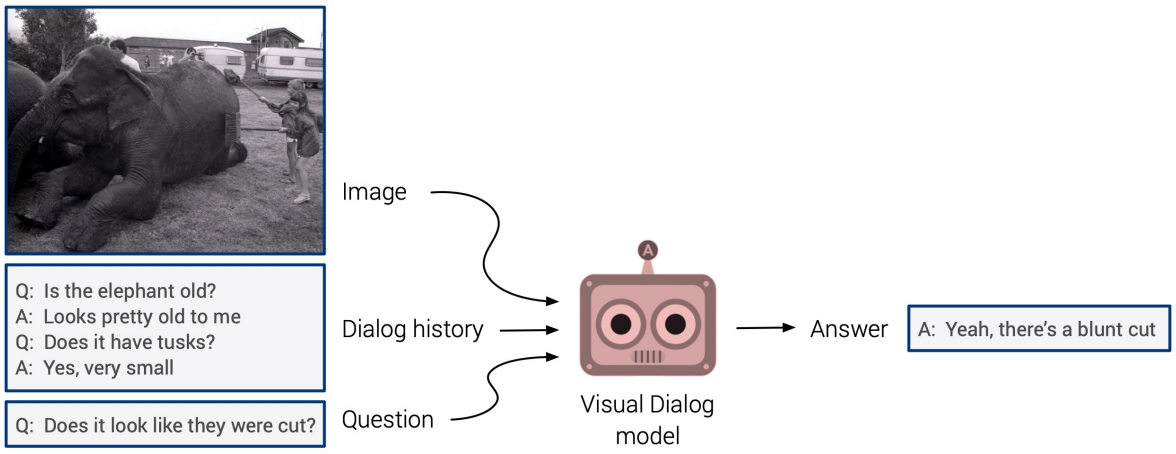

図 2.3: Visual Dialog の概要図 [1]．モデルは一連の質問に答える必要があるが，過去のやり取りから手がかりを得ないと回答できない質問も含まれる．

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 [28] 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．しかし，回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．

Uehara らは，画像中のアノテーションされていない未知のオブジェクトについて，情

報を得るために VQG を利用した手法を提案している [29]．彼らのフレームワークは，画像から物体候補を得る部分，物体候補から未知物体を選出する部分，未知物体領域についての質問文を生成する部分からなり，生成された質問文を人間に提示して未知物体の情報を答えさせることで，未知物体についての学習データを得ることを目的としている．物体候補から未知物体を選出する部分では WordNet を用いて，物体のラベル候補の上位語を得て，それを手がかりに質問文を生成している．

Liu ら [31] 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回答が与えられ，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回答ペア生成と近いタスクであるが，入力とする画像と回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．

Li らは，VQG を VQA の逆問題として見なし，潜在空間のパラメータの重みを共有して VQA モデルと VQG モデルの学習を同時に行うことで，両方のタスクで精度を向上させる手法を提案した [32]．VQA は質問文と画像から回答を予測するタスクであるのに対し，VQG は画像と回答から質問文を予測するタスクであるため，この二つは逆問題として捉えることができる．

Visual Dialog は，入力画像とそれについての一連の会話において，妥当な返答を行う派生タスクである [1]．Visual Dialog タスクの概要図を図 2.3 に示す．このタスクでは，初期ステップで画像とその画像についてのキャプションがエージェントに与えられ，さらに一つ目の質問がなされる．エージェントが一つ目の質問に回答すると，二つ目の質問が与えられると同時にこれまでの会話の履歴も与えられる．こうして決められた数の質問と回答を繰り返し，これを 1 ラウンドとする．Visual Dialog の評価はラウンド中の各回答の正解率を，全ラウンドで平均した値でなされる．二つ目以降の質問は，画像とその文章だけでは正しい回答を導けず，初めのキャプションや過去のやり取りが手がかりとして必要になる場合がある．したがって Visual Diaglog タスクでは，モデルがいかに History を保持するかということが重要になる．

Embodied Question Answering（EQA）は，エージェントに三次元空間上で VQA を解かせるという派生タスクである [2]．EQA タスクの概要図を図 2.4 に示す．具体的には，家屋の 3DCG のデータセットである House3D[36] で構成された三次元空間上でランダムにエージェントの初期位置が定められ，エージェントは初期位置の一人称視点（レンダリングされた画像）と質問文が与えられる．その後エージェントは前進，回転など決められた幾つかのアクションを実行していき，そのたびに更新された一人称視点の画像を得る．これを任意ステップだけ繰り返し，最終的にエージェントは初めの質問に対する回

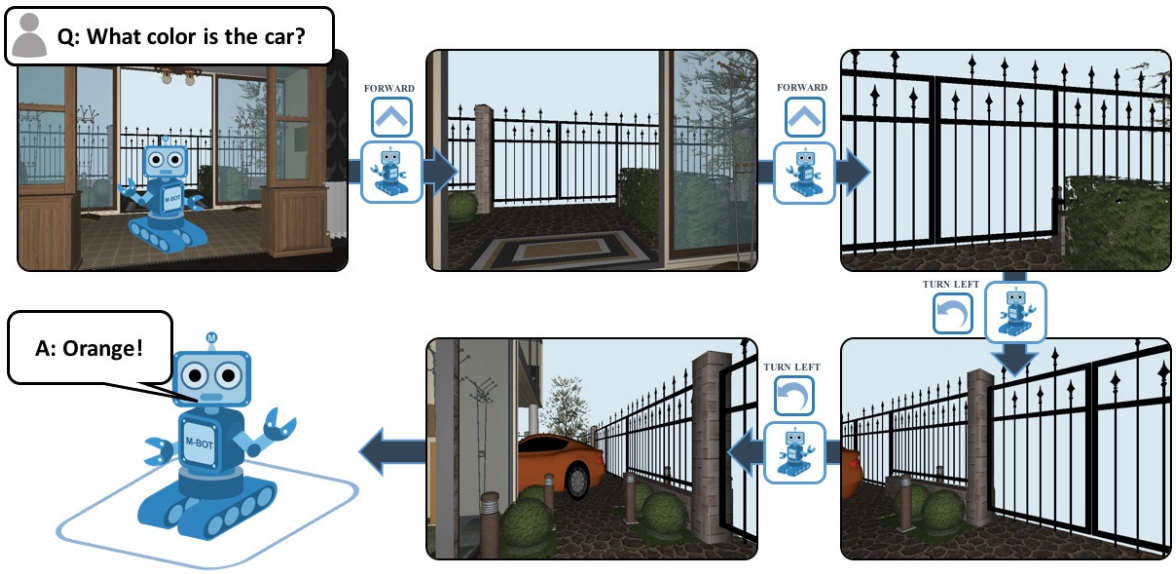

図 2.4: Embodied Question Answering(EQA) の概要図 [2]．エージェントは質問に答えるため，三次元空間上を移動する必要がある．一人称視点画像は毎アクションごとに更新され，モデルに入力される．

答を出力する．EQA の評価は最終的なエージェントの回答の正解率によってなされる．EQA は初期位置の視点だけでは質問に回答することができないため，エージェントが質問を理解し，自律的に行動して回答を導くことが望まれる．EQA タスク用の機械的に生成された質問と回答のデータセットも公開されている．

Das らは，NMN にアイデアを得て EQA をサブタスクに分割し，サブタスクごとにエージェントの行動の最適化を行う Neural Modular Control を提案した [37]．この手法では EQA タスクでエージェントが取る必要のある幾つかのアクション，例えば部屋から出る，ある部屋を探す，あるオブジェクトを探す等をサブタスクとして扱い，別々に強化学習を行うことで EQA 精度を向上させている．

2.4　言語モデル

自然言語処理の分野では Recurrent Neural Network(RNN) が長らく用いられてきた．長期的な入力を考慮するために Long Short Term Memory(LSTM) Network が登場した [12]．長年 LSTM が用いられていたが，2010 年代に入って Gated Recurrent Unit(GRU) というアーキテクチャが登場した [13]．さらに近年，Transformer と呼ばれる新たなアー

キテクチャが登場した [3]．Transformer のアーキテクチャを図解したものを図 2.5 に示す Transformer はエンコーダ・デコーダ型のアーキテクチャで, 自己注意 (Self Attention) という構造を持つことが重要である．自然言語処理の代表的なタスクである機械翻訳分野では既に多くの手法で RNN 系のアーキテクチャは用いられず，Transformer に取って代わられている．

Devlin らは，Transformer のエンコーダ部分を発展させた Bidirectional Encoder Representations from Transformers(BERT) を提案した [14]．これは自然言語処理の様々なタスクに適応可能な汎用性の高い言語モデルであり，大規模言語コーパスを用いて事前学習され，タスクごとにファインチューニングを行う．構造としては Transformer の前段のエンコーダを幾つか連結したものになっている．また BERT は双方向 (Bidirectional) な Transformer と捉えることもできる．これは従来の Transformer では事前学習で未来の単語を予測するタスクを解いており，現在の単語以降をマスクする必要があったため，一つの文章について前から後ろという単方向の学習しかできなかったが，BERT では学習の際にランダムに文章中の単語を予測させているため，周辺の単語情報全てを用いて学習が可能になったためである．自然言語処理タスクに汎用的な事前学習という，BERT と似たアプローチを行っている ELMO も，文章の始まりから終わりへの順方向学習と，終わりから始まりへの逆方向学習を別々に行う必要がある [38].

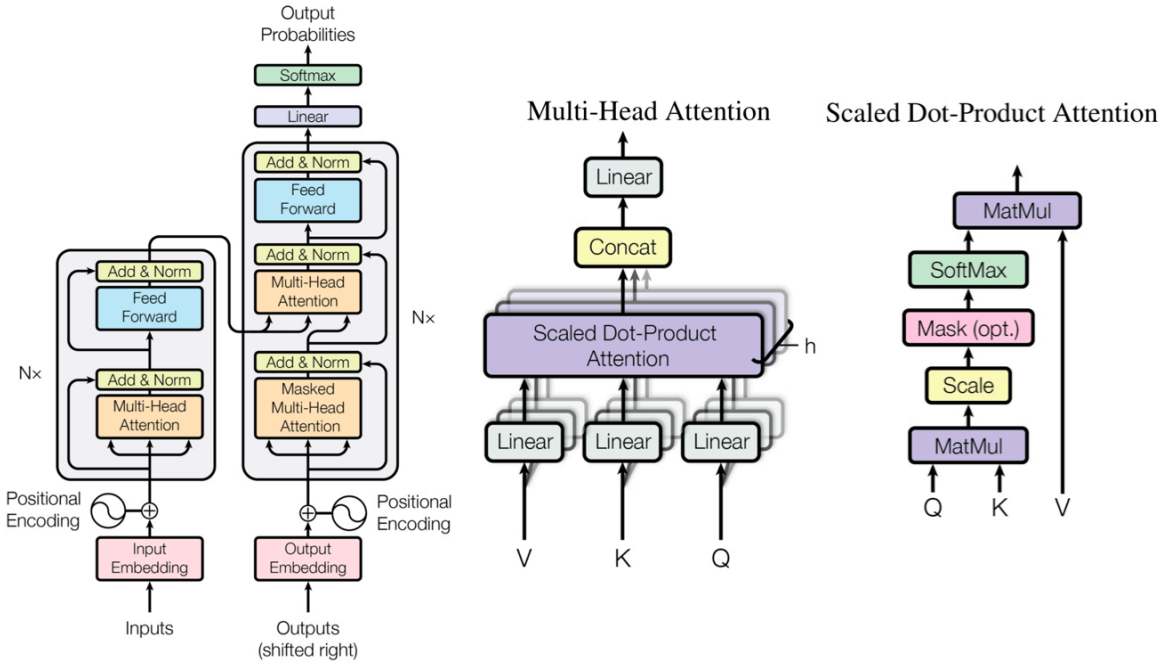

図 2.5: Transformer のアーキテクチャ図 [3]. Multi-Head Attention による自己注意機構を持つエンコーダ・デコーダ型のアーキテクチャである.

第 3 章

半教師あり学習のための多種質問回答生成

本論文では，Visual Question Answering において半教師あり学習を用いてラベルなし画像の活用の知見を得ることと，VQA のためのデータ拡張を行い，精度を向上させることを目的とする．このチャプターでは，Visual Question Answering 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回答ペアを生成する手法について述べる．

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，正ラベル付きデータのみによる学習よりも精度を高めることを目的とした学習設定である．今回は VQA において，少数の正ラベル付きデータ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 Pseudo-Label として質問と回答を割り振る．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．

質問回答ペア生成のため，本論文では大きく分けて 4 つのアプローチを行っている．具体的には，敵対的学習を用いて生成モデルを作成する方法，キャプション生成手法である Densecap を利用して質問回答ペアを作成する方法，事前学習されたキャプション生成モデルを VQA ドメインに適応させる方法，そしてテンプレートを元に質問回答ペアを作成する方法である．さらに，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，これらの手法に加えて Long short-term memory（LSTM）[12] のみを用いた単純なベースライン手法も用意した．本節ではこれらの手法について順に述べる．

3.1 敵対的学習による質問回答生成手法

敵対的学習による質問回答ペア生成の手法の概観を図 3.1 に示す．本手法の敵対的学習では，画像に対する質問回答ペアの生成器と，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 として質問回答生成器を用い，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回答ペアが生成されることを期待する．また，敵対的学習の際に入力する画像群と，合成質問回答ペアを付与する画像群は同一のものである．

質問回答ペアを生成する Generator の本体には LSTM を用いる．パラメータが初期化されたままの LSTM を Generator として敵対的学習を行うと学習が不安定となるため，半教師あり学習で利用可能とした少数の正ラベル付きデータを LSTM の事前学習に用いている．

また画像に対応する質問回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回答を結合している．この際，質問と回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，その後回答を生成するという形式を容易に学習させることが出来ている．

3.1.1 Generator

本手法の Generator を図解したものを図 3.2 に示す．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 [16] 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た．

$$J(\theta) = -\sum_{n=1}^{N}\sum_{t=1}^{T} log\pi_\theta(\hat{y}_t^n|\mathbf{x}^n, \hat{\mathbf{y}}_{t-1}^n) \tag{3.1}$$

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

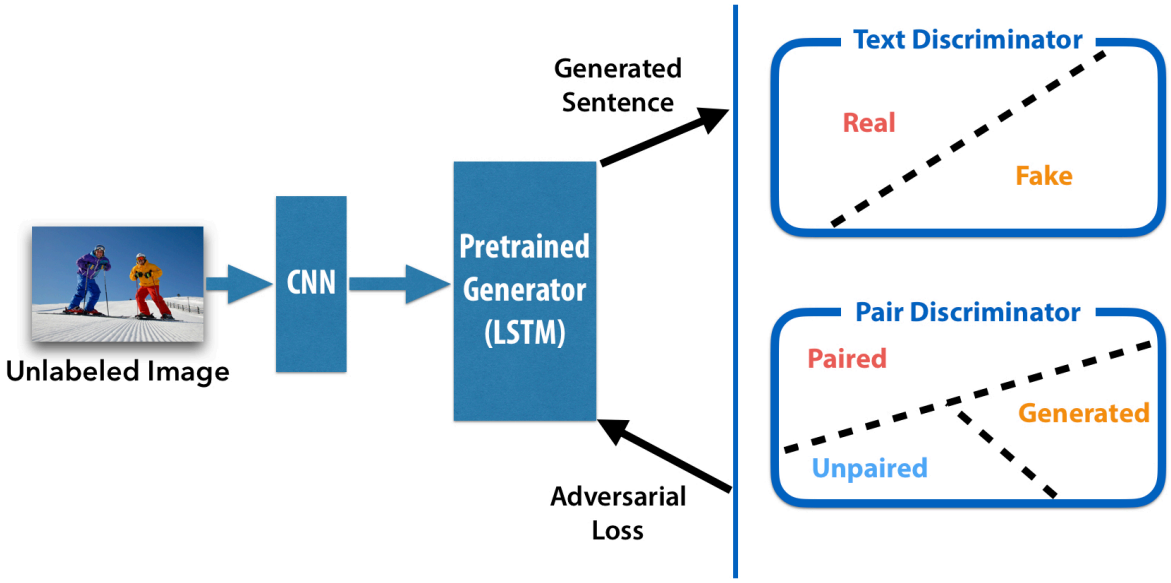

図 3.1: 我々の敵対的学習手法の概要図．少数の正ラベル付きデータで LSTM を事前学習した後，ラベルなし画像を入力として二つの Discriminator を用いて敵対的学習を行う．

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[19]）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．

$$J(\theta) = \sum_{n=1}^{N} \sum_{t=1}^{T} E_{\mathbf{y}_t^n}[\pi_\theta(y_t^n | \mathbf{x}^n, \mathbf{y}_{t-1}^n) Q((\mathbf{x}^n, \mathbf{y}_{t-1}^n), y_t^n)] \tag{3.2}$$

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．

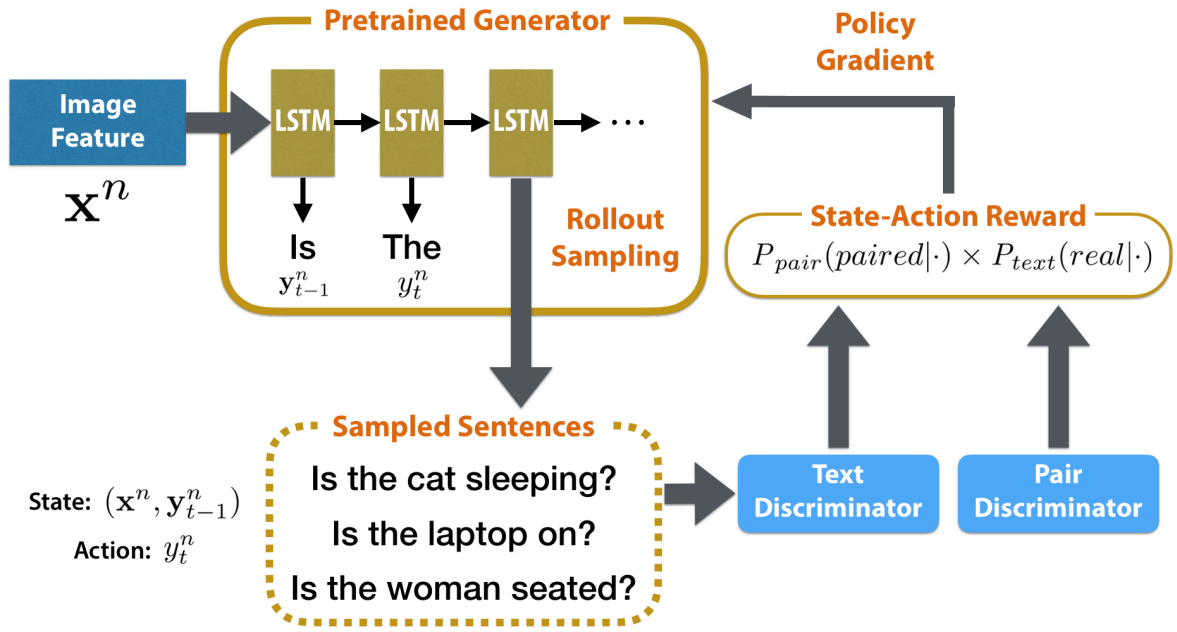

図 3.2: 生成モデルの敵対的学習の具体的な方法．Rollout Sampling を用いて LSTM の出力をサンプリングし，Discriminator に与える．報酬は二つの Discriminator のスコアをかけ合わせたものである．

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 の概観を図 3.3 に示す．Pair Discriminator の学習では，Minimum Set 内の画像質問回答の組に *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 層の出力は以下のようになる．

$$P_{pair}(class|\mathbf{x}, \mathbf{y}), class \in \{paired, unpaired, generated\}$$

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

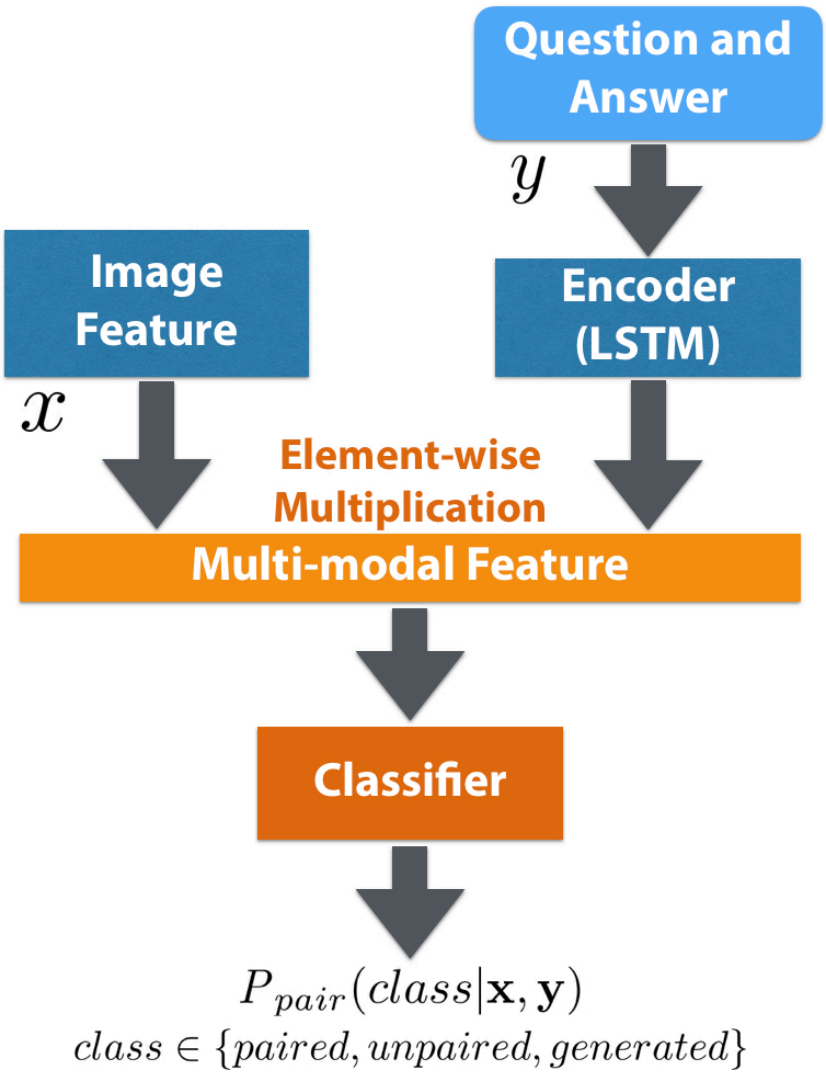

図 3.3: Pair discriminator の概要図．画像特徴量と言語特徴量を Concat するのではなく，要素ごとにかけ合わせてマルチモーダル特徴量とする．

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 内の質問回答文に $real$ 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 層の出力は以下のようになる．

$$P_{text}(class|\mathbf{y}), class \in \{real, fake\}$$

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 は以下のように表せる．

$$R(\cdot) = P_{pair}(paired|\cdot) \times P_{text}(real|\cdot) \tag{3.3}$$

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．

$$Q((\mathbf{x}^n, \mathbf{y}^n_{t-1}), y^n_t) \simeq \frac{1}{K}\sum_{k=1}^{K} R([\mathbf{y}_{t-1}, y_t, \mathbf{y}_{(t+1):T_k}]|\mathbf{x}) \tag{3.4}$$

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 キャプション生成モデルのドメイン適応による質問回答生成手法

本節では，事前学習されたキャプション生成モデルを VQA ドメインに適応させる方法について述べる．この手法は前節で述べた敵対的学習を用いた手法と学習の流れは同様だが，LSTM の事前学習にキャプションデータセットを用いる部分が異なっている．

事前学習で用いることのできるキャプションデータは，従来のキャプション生成手法と同様に MSCOCO の train データセットとした．事前学習を終えた段階では LSTM すなわち Generator は画像に対するキャプション生成器として働くが，敵対的学習が進むにつれて Generator が質問回答ペア生成器へと適応する．

この手法では，キャプション生成の知識を持った Generator を適応学習させることで，キャプションデータの知識を活かしつつ画像に対する質問回答ペアを生成できることを期待する．すなわち VQA データセットのみで学習された生成器と比較して，外部データセットを活用してより質の高い質問回答ペアを生成することを目指した．事前学習と本学習の詳細な条件については実験の章で述べる．

3.3 Dense Captioning を用いた質問回答生成手法

Johnson らが提案した Dense Captioning(Densecap)[4] 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．従来のキャプション生成手法は，一枚の画像全体を表すキャプションを生成するように学習されていたが，この手法は写っている各物体について密なキャプションを付けることができる．Densecap によって生成さ

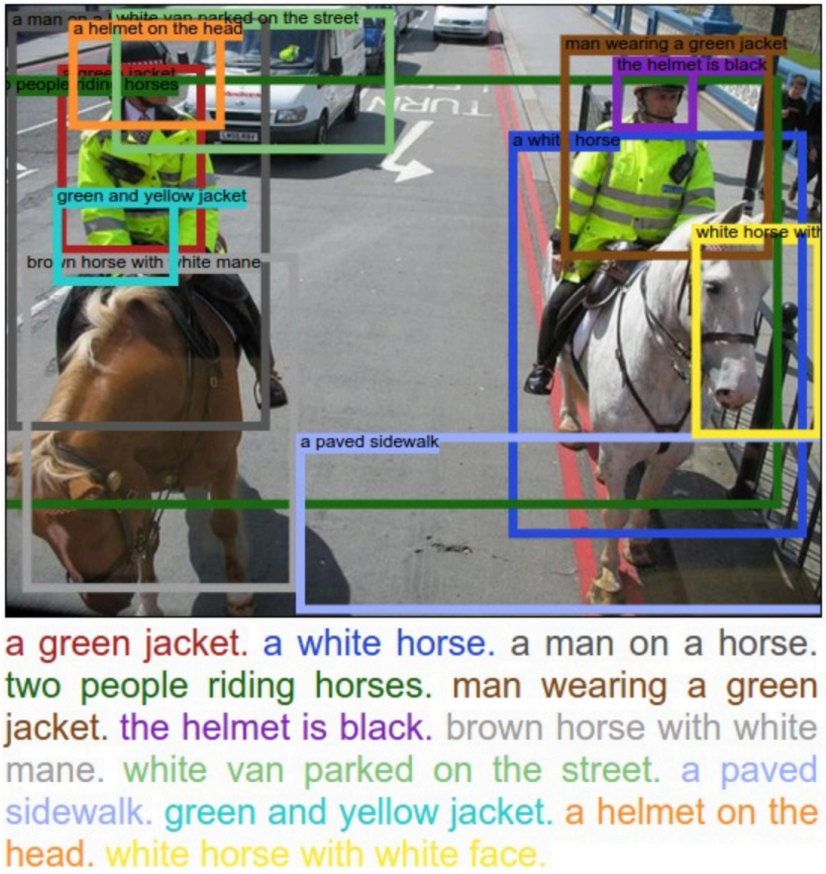

図 3.4: Dense Captioning によるキャプションの生成例 [4]．一枚の画像に対して物体ごとの密なキャプションを生成できる．

れたキャプションの例を図 3.4 に示す．Densecap 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．この手法の概要図を図 3.5 に示す．

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 [39] の画像中の各物体に対応したキャプションを用いているため，本手法で生成された合成データによる学習は厳密に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キャプションは画像中のオブジェクトに対応した肯定文であるため，VQA の合成データとして用いるにはこれを質問文と回答文のペアに変換する必要がある．今回は肯定文から質問文への変換に，Ren らの手法 [40] 中で使われている方法を用いた．

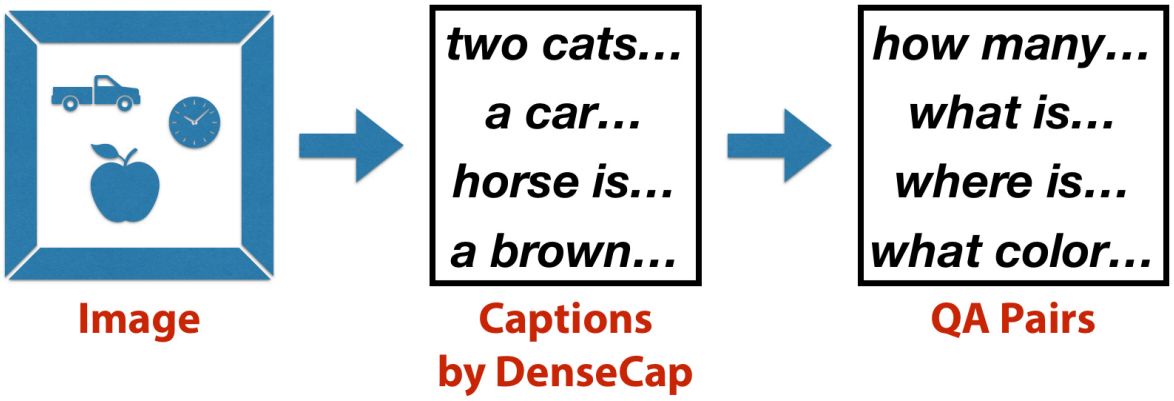

図 3.5: Dense Captioning を利用した我々の質問回答生成手法の概要図．画像に対して Densecap で生成されたキャプションを，それぞれ質問文と回答文に変換して合成データとする．

詳細は実験の章で述べるが，本研究の実験では VQA データセット内の全質問回答ペアの数と，半教師あり学習で用いられる正ラベル付きデータの数が定まっている．Densecap によって一枚の画像に対して生成されるキャプションの数は画像によってばらつきがあるが，半教師あり学習に用いる合成質問回答ペアの数は実験設定に合わせて変更している．

また，Densecap の性質上，物体に対応してはいるもののキャプションとは呼べない文章が出力される場合がある．例えば”An horse.”や”Green jacket.”などである．このような文章は Ren らの手法によって正しく質問回答ペアに変換できないため，合成データには含まれないように除外している．さらに，同じカテゴリの質問ばかりが合成データに含まれないよう，質問の種類を調整している．具体的には，物体の色を問う質問 (what color)，数を問う質問 (how many)，物体自体に関する質問 (what) などがそれぞれ合成データに存在するならば，合成データセットを生成する際にそれらが一つずつ含まれるよう調整を行った．

3.4 テンプレートベースの質問回答生成手法

本節では，テンプレートベースで合成質問回答ペアを生成する手法について述べる．この手法は文章の生成に LSTM 等の深層学習モデルを用いず，幾つかの質問と回答のテンプレートを用意し，物体検出アノテーション等を利用してテンプレートを完成されて合成データとするものである．

Kafle らは既存の VQA 手法をより正確に比較するために，VQA データセットに含ま

れる質問の種類を以下の 12 タイプに大別した [41].

1. Object Presence: 物体の存在認識 (e.g., ‘Is there A in the image?’)
2. Subordinate Object Recognition: 物体の種類認識 (e.g., ‘What kind of A is in the picture?’)
3. Counting: 数え上げ (e.g., ’How many A are there?’)
4. Color Attributes: 色認識 (e.g., ‘What color is A?’)
5. Other Attributes: その他の特性認識 (e.g., ‘What shape is A?’)
6. Activity Recognition: 活動認識 (e.g., ‘What is the A doing?’)
7. Sport Recognition: スポーツ認識 (e.g.,‘What are A playing?’)
8. Positional Reasoning: 位置認識 (e.g., ‘What is to the left of A on B?’)
9. Scene Classification: シーン認識 (e.g., ‘What room is this?’)
10. Sentiment Understanding: 感情理解 (e.g.,‘How is A feeling?’)
11. Object Utilities and Affordances: 物体の効用認識 (e.g.,‘What object can be used to A?’)
12. Absurd: 意味をなさない質問

彼らは VQA データセットから各カテゴリごとにテンプレートの元となる質問文を抽出し，MSCOCO データセットのアノテーションを用いて，改めてカテゴリごとに半自動的に質問回答ペアを作成した.

このアプローチからアイデアを得て，半教師あり学習の合成データとしてテンプレートを用いることを考えた．簡単のため，今回は 12 タイプのうち 5 タイプのみ (Object Presence, Subordinate Object Recognition, Counting, Color Attributes, Sport Recognition) に絞り，MSCOCO データセットの画像と各アノテーションからテンプレートベースの合成質問回答ペアを生成した．Counting タイプの質問を例に具体的な生成方法を述べる．まず，テンプレートとして ”Question: How many A in this photograph? Answer: B.” のような簡単なテンプレートを質問タイプごとに幾つか用意する．続けて，MSCOCO の物体検出アノテーションをキャプションアノテーションを用いて，テンプレートの空き部分を埋める．この部分のアルゴリズムは質問タイプによって異なるが，Counting タイプでは物体検出アノテーションを参照し，画像中の各物体の数を数え上げ，それぞれについてテンプレートを埋めた質問回答ペアを生成する．こうして生成された 5 タイプの質問回答ペアを，正ラベル付きデータに加えて合成データとし，VQA モデルの半教師あり学習を行う．この手法がこれまで紹介した他の手法と大きく異なっているのは，合成データの生成に深層学習モデルを用いていない点である．Kafle らの論文でも

述べられているように，今回選んだ 5 タイプは比較的 VQA モデルにとって易しい Easy Questions である．これらについて，テンプレートベースの簡単な，しかし VQA データセットには含まれない新規な質問回答ペアを合成データとして学習させた場合，精度の向上が見られるのか，知見を得るためにこの手法を取り入れた．

3.5 実験

本章では，提案手法の章で述べた各手法の性能を評価するための実験について述べる．まず各実験で用いるデータセットについて述べた後，半教師あり学習の設定で VQA の合成データ生成手法を比較する実験について述べる．今回は各手法からラベル無し画像に対して質問回答ペアを生成させ，その組み合わせを合成データとして正ラベル付きデータに加えて VQA モデルに学習させた精度を半教師あり精度とする．最後に具体的な出力例を挙げながら結果を示し，合成データの内容を比較しつつ考察を行う．各節で，データに対する前処理や事前学習の条件，各実験で用いた初期パラメータについても明記する．

3.5.1 データセット

MS COCO [18] は，約 20 万件の画像と様々なタスクのためのアノテーションを含むデータセットである．本実験で用いる 2014 年版の MSCOCO データセットには，Training 用の画像が約 8 万件，Validation 用の画像が約 4 万件，Test 用の画像が約 8 万件含まれている．MSCOCO には物体検出のためのバウンディングボックスを始めとして，画像一枚につき 5 件ずつの人間によるキャプション，セマンティックセグメンテーションのためのマスクなど，多くのアノテーションが存在し，様々な分野で評価用のデータセットとして用いられている．このデータセットの画像の傾向として，複数のオブジェクトが同時に写り込んでおり，それらの位置関係も単純ではないため，複雑なシーンを持つものが多い．

VQA データセット [43] は．MSCOCO データセットの画像に対応する質問と回答ペアのデータセットであり，Version1 で提供されている総質問回答ペアは 658,111 件である．Multiple-Choice 型では各質問ごとに 18 件の回答候補が用意され，Open-Ended 型では各質問に対する人間による回答が 10 件用意されている．VQA データセットの精度評価では，この 10 件の回答のうち，3 件以上同じ回答が存在するものについては正答と定義している．一つの質問回答ペアについて，正答できた際の Accuracy を 1 とすると，10 件中 2 件存在する回答の場合の Accuracy は 0.66，10 件中 1 件のみ存在する回答の場合の

Accuracy は 0.33，それ以外の回答は 0 である．これらを全質問回答ペアについて平均したものが VQA データセットでの Accuracy となる．

3.5.2 前処理

本実験では深層学習モデルを用いた各手法において，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，質問と回答の区切りを示す固有トークンも含まれている．

VQA データセットは画像とそれに対応する質問と回答のペアからなるが，画像は既に抽出済みの Bottom-Up Attention Feature を用いた．質問文と回答文中の単語はそれぞれ，あらかじめ作成した辞書のインデックスに置き換えた．そして，トークナイズされた質問文と回答文を $[TKN]$ トークンで結合し，モデルの入力とした．

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データとした．これを以下では Minimum Set と呼ぶことにする．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．

3.5.3 事前学習

本実験で質問回答ペアの生成器として用いたのは，Vinyals らが提案した CNN-LSTM 型 [16] 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 の初期状態に入力される形を取る．今回の実験では，LSTM の Hidden State 数は最も性能がよい 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[44] による pool5 層特徴量とした．

最適化手法には Adam[45] を選択し，正ラベル付きの VQA データは 30,000 件によって学習させた．この 30,000 件の VQA データを固定し，敵対的学習の際の Discriminator の学習と，半教師あり学習の際の正ラベル付きデータとして用いた．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．

3.5.4 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．

3.5.5 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，本実験では実装が公開されている中で最も性能が良い，Bottom-Up Attention と Top-Down Attention を用いた既存の VQA モデル [8] 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．Baseline は，敵対的学習を行わず CNN-LSTM 型のモデルに直接画像から質問回答の出力を学習させ，その後ラベルなし画像に対する出力を合成データとしたものである．VQA Pre-train Adversarial Training は，LSTM の事前学習を 30,000 件の正ラベル付きデータで行い，その後敵対的学習を行ったモデルである．Caption Pre-train Adversarial Training は，LSTM の事前学習を MSCOCO のキャプションデータで行い，その後敵対的学習を行ったモデルである．VQA Pre-train Input Caption Data は，LSTM の事前学習を 30,000 件の正ラベル付きデータで行い，その後の敵対的学習でキャプション．Template Based Method は，5 タイプの質問テンプレートから合成質問回答ペアを生成したものである．Densecap は，ラベルなし画像に対して Dense Captioning によって多種キャプションを生成し，それらを質問回答ペアに変換したものである．Upper-bound は利用可能な VQA 学習データを全て用いて VQA モデルを学習させたものであり，上限値とみなす．

これらの手法で生成された合成質問回答ペアを 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．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

MS COCO データセットの画像を指す．

表 3.1: VQA データセットの Validation Set における，各手法の半教師あり精度．Minimum-Set Only は 30,000 件の正ラベル付きデータのみを用いてモデルを学習させた精度である．Dense Captioning を利用した手法は，その精度を約 1% 上回った．

Model	VQA Score
Caption Pre-train Adversarial Training	43.28
Baseline	45.14
VQA Pre-train Adversarial Training	**46.23**
Template Based Method	**48.17**
Minimum-Set Only	48.90
Densecap	**49.96**
Upper-bound	62.09

3.5.6 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 3.1 に示す．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，敵対的学習による手法ともに，生成される合成データの回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．

テンプレートベースの手法は，Minimum Set のみの精度をわずかに下回った．この結果から，今回用意した 5 タイプのテンプレートはいずれも Easy Question であり，VQA データセットに含まれない新規な質問回答ペアを生成できても，VQA 精度の向上には寄与しないことが確認できた．テンプレートベースの手法は敵対的学習を用いた手法と比較してノイズが無いため，精度が大きく低下してはいないと考えられる．

キャプションデータで LSTM の事前学習を行ってから適応学習を行う手法は，ベースライン手法を下回った．LSTM がキャプション生成器としてはたらく状態で敵対的学習

を行うと，現状ではノイズが非常に多くなってしまうということが確認できた．

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 Densecap の学習に Visual Genome データセット [39] の Region Descriptions を用いており，外部データを有効活用した新規の質問回答ペアが精度向上に寄与していると考える．

各手法によって生成された質問回答ペアと，実際の VQA データセットのサンプルを図 3.6 に示す．敵対的学習による手法では，ラベル無し画像に対応した多様な質問回答ペアを生成できていることが見て取れるが，誤った回答も含まれている．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回答のペアを生成できており，多様性は低いが，回答の誤りは殆ど存在しない．テンプレートベースの手法では，テンプレートに沿って非常に簡単な質問と回答のペアを生成できており，ノイズも殆ど含まれないが，この難易度の質問では精度の向上には影響がなかった．

質問の種類の分布を比較するため，VQA データセットに含まれる質問文と Densecap を用いた手法で生成された質問文を，それぞれ Word N-grams をもとに可視化した．その結果をサンバースト図として示したのが図 3.7 と図 3.8 である．VQA データセットの分布と比較して，Densecap を用いた手法による生成データには Yes/No 質問が殆ど含まれず，分布に偏りがあることがわかる．しかし，これらには Minimum Set に含まれない新規の質問回答ペアが多く含まれ，誤った回答の数が非常に少ないため，本実験では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．

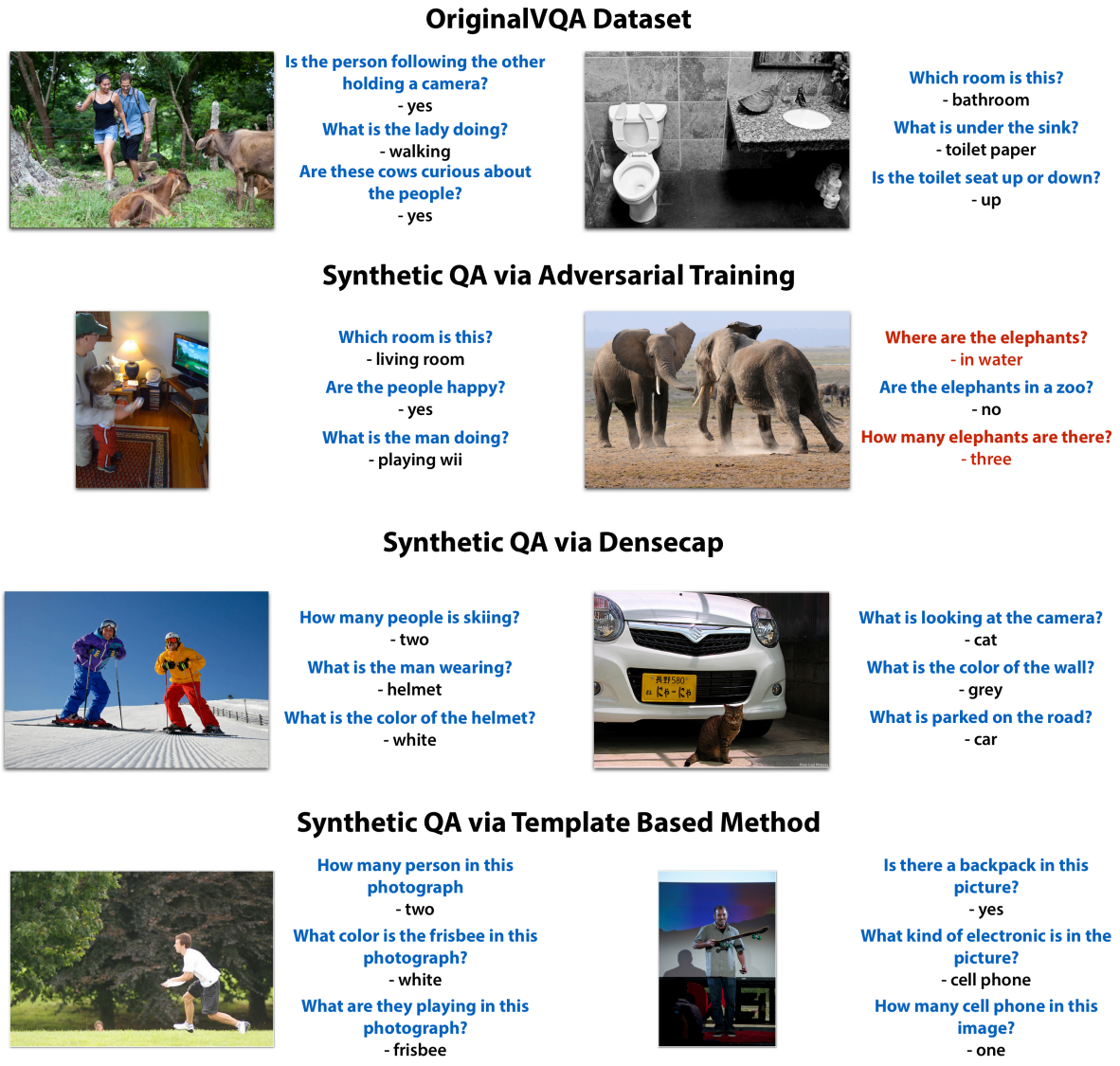

図 3.6: 我々の手法でラベルなし画像に対して生成された合成質問回答ペアと，実際に VQA データセットに含まれる質問回答ペアの比較．赤字はノイズとなった（妥当でない）ペアを示しており，敵対的学習を用いたモデルはノイズを含むことがわかる．Densecap を用いた手法の出力はシンプルな質問だが，ノイズはほとんど含まれない．テンプレートベースの手法も，出力は簡単でワンパターンだが，ノイズは含まれない．

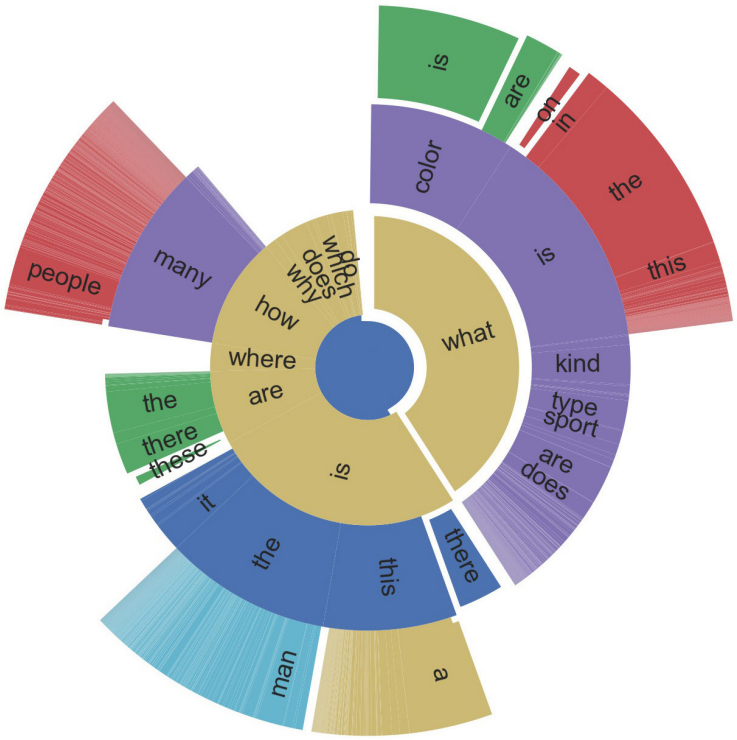

図 3.7: VQA データセットの Word N-grams．様々な種類の質問がバランスよく含まれている．

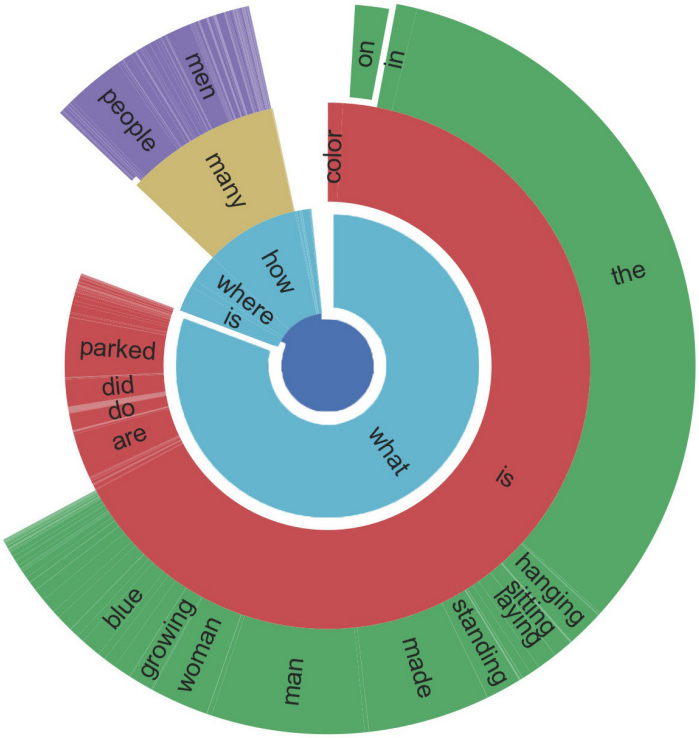

図 3.8: Densecap[4] を用いた手法によって生成された合成データセットの Word N-grams. 具体的な物体名を答えさせる質問と数を答えさせる質問が大半であることがわかる．実際の VQA データセットと比較すると質問の種類が乏しく，特に Yes-No Question がほとんど含まれない．

第 4 章

Transformer を用いたデータ拡張

本論文では，Visual Question Answering において半教師あり学習を用いてラベルなし画像の活用の知見を得ることと，VQA のためのデータ拡張を行い，精度を向上させることを目的とする．このチャプターでは，VQA モデルにおける自然言語のエンコーダとして Transformer を取り入れ，学習時に質問文内の単語をランダムに予測させることでデータ拡張を行う手法を提案する．Vaswani らによって提案された Transformer はエンコーダ・デコーダ型のアーキテクチャだが，本手法では Transformer のエンコーダ部分を発展させた BERT アーキテクチャを自然言語のエンコーダとして用いる [14]．既存の VQA 手法では学習の際に質問文と画像を入力してそれぞれエンコードし，回答を予測することでモデルの最適化を行うが，本手法では質問と回答を結合したものをモデルの入力とし，事前学習の際にランダムに文章中の単語をマスクし，それらを予測させることでデータ拡張を行う．自然言語のエンコーダに BERT を用い，ランダムマスキングによる事前学習によってデータ拡張を行う我々の手法は，VQA 以外の様々な Vision Language Task のモデルに適用可能である．

4.1 アーキテクチャ

本手法では，ベースとなる VQA モデルは Kim らの Bilinear Attention Networks (BAN) とした [5]．BAN では質問文のエンコーダとして GRU を用いているが，この部分に GRU ではなく Transformer のエンコーダを連結し発展させた BERT を組み込む．BAN のアーキテクチャ図を図 4.1 に示すが，これはオリジナルの図であり質問文エンコーダが BERT でない点に注意されたい．トークナイズされた質問文を GRU に入力する際，学習済みの GloVe 単語ベクトル [42] による変換が行われているが，BERT は大規

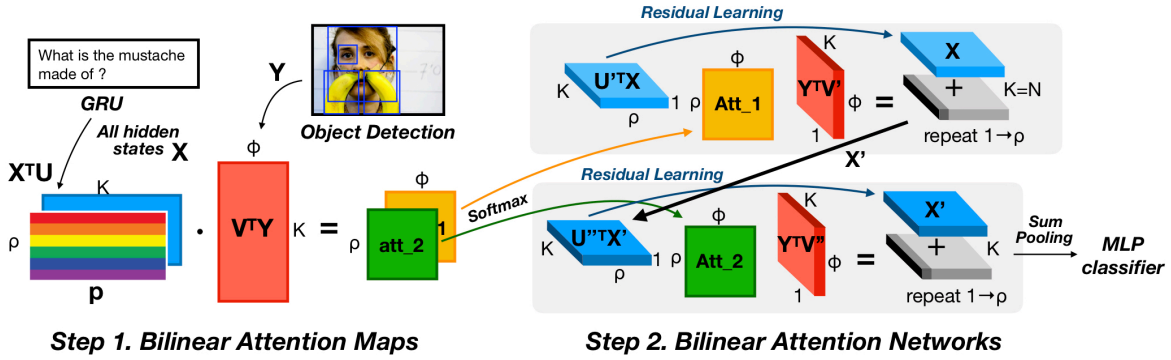

図 4.1: Bilinear Attention Networks(BAN) のアーキテクチャ図 [5]．GRU を用いた言語特徴量と，Bottom-Up Attention による画像特徴量から Bilinear Attention Map を得る．残差学習を用いた MRN によって，最大 8 つのアテンションマップを学習することができる．

模コーパスによる事前学習の際に Embeddings レイヤーもまとめて学習を行っているため，我々の手法では GloVe は使用しない．また，画像の特徴量としては事前学習された Anderson らの Bottom-Up Attention による特徴量を利用している．入力文章が BERT によってエンコードされた特徴量のサイズは，バッチサイズを B，最大文章長を ρ，BERT の隠れ層の数を N として，$B \times \rho \times N$ となる．質問文がエンコードされた後はオリジナルの BAN と同様に，言語特徴量と画像特徴量を結合し，g 個の Bilinear Attention Map を用いて学習を行う．回答の予測部分は VQA データセットに含まれる回答の候補数分の多値分類となっている．実験時の各パラメータの詳細な条件については，実験の章で述べる．

4.2 マスキングを用いた事前学習と VQA データの学習

本手法では，VQA の本学習を行う前に，ランダムマスキングによって BERT 部分の事前学習を行っている．本節では，具体的な学習の方法と処理について述べる．図 4.3 にランダムマスキングによる事前学習の概要図を示す．従来の VQA 手法では，学習の際に自然言語エンコーダに VQA データの質問文のみを入力している．しかし我々の手法ではモデルの学習の際に入力とするのは質問文だけでなく，前節までに述べたように固有トークンで区切られ結合された質問回答ペアの文章である．BERT 部分の事前学習では，入力となる質問解答文書の固有トークン以外の部分をイテレーションごとにランダムにマスク

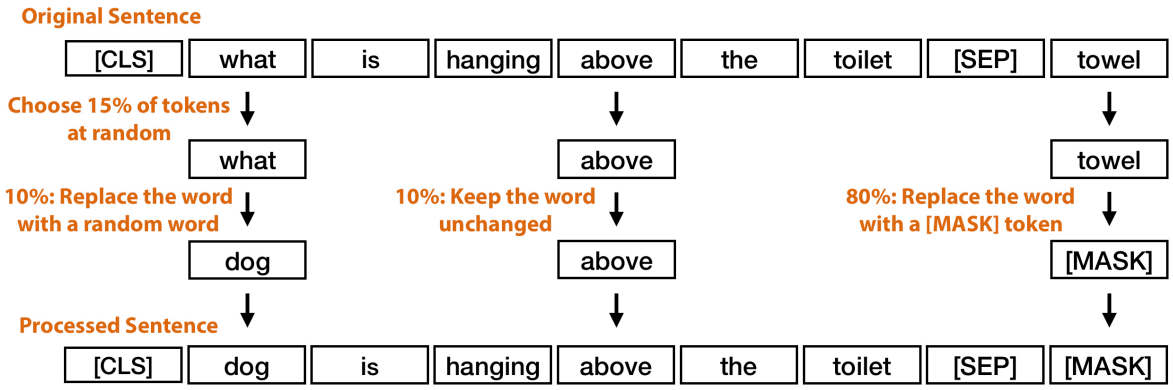

図 4.2: 具体的なランダムマスキングの処理．まず特殊トークン以外から 15% のトークンをランダムに選び出す．選び出されたそれぞれについて，80% で $[MASK]$ トークンに置き換え，10% でランダムな別の単語に置き換え，10% で置き換えずそのままにする．

し，マスクされた部分の単語予測を行う．具体的な条件と処理を以下に示す．これらの処理は，主に Devlin らの論文中の記述を参考にしている．また，この処理を図解したものを図 4.2 に示す．

- まず，入力となるトークナイズされた文章のうち，15% のトークンをランダムに選び出す．
- 選ばれたトークンそれぞれについて，以下の処理を行う．
  - 80% の確率で，そのトークンを $[MASK]$ トークンに置き換える．
  - 10% の確率で，そのトークンをランダムな別の単語のトークンに置き換える．
  - 10% の確率で，そのトークンを変更せずに元の単語のまま保持する．

事前学習の際，BERT の入力長と出力長は同じである．すなわち，マスクされていない部分の単語はそのまま出力し，マスクされている部分は周囲の単語情報を手がかりに予測する必要がある．また，詳細は実験の節で述べるが，損失にはマスクされた単語の予測のみ考慮される．

従来の VQA 手法では VQA データセットの質問文を 1 ステップずつ LSTM または GRU で読み込み，特徴量を抽出して回答を予測していたが，このような方法を取ることによって，質問回答文を学習データとしてより活用できると考える．大規模コーパスで事前学習された BERT は既に自然言語のよい特徴抽出器としてはたらくが，この事前学習は BERT をより VQA に特化した特徴抽出器としてチューニングさせる，という意味合いを持つ．VQA の本学習では，BERT 部分だけでなく BAN のパラメータ全体を学習さ

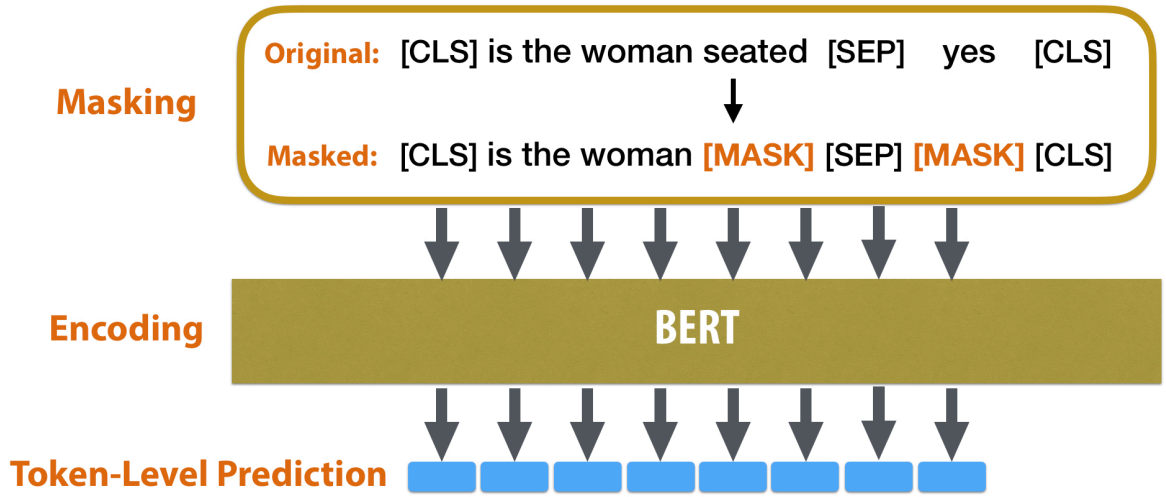

図 4.3: ランダムマスキングによる事前学習の概要図．BERT モデルは一部をマスクされたトークン群を入力とし，各単語ごとの予測を行う．入力トークン長と出力トークン長は同じである．この際，損失として学習に影響するのはマスクされた部分の予測のみとする．

せる．本学習では質問文のランダムマスキングは行わないが，入力となる結合された質問回答文の回答部分のみを $[MASK]$ トークンに置き換えることで，VQA の問題設定と同様にする．事前学習と本学習での詳細な条件や初期パラメータについては，実験の章で述べる．

4.3 実験

4.3.1 前処理

我々の手法では大規模コーパスで事前学習済みの BERT モデルを用いている．その際に用いているコーパスは，BookCorpus（約 8 億語）[46] と英語版 Wikipedia から生成したコーパス（約 25 億語）である．入力文章のトークナイズに用いる辞書は，配布されているサイズ 30522 のものとした．前節の実験と同様に，質問文と回答文中の単語はそれぞれ，上記の辞書のインデックスに置き換えた．そして，トークナイズされた質問文と回答文を固有トークンで結合し，モデルの入力とした．BERT では明示的に BOS トークンと EOS トークンを入力に含めており，それぞれ $[CLS]$ と $[SEP]$ である．この実験では $[SEP]$ トークンを質問と回答を区切るトークンとしても使用している．BERT の学習の

際に文章長は固定する必要があるため，特殊トークンを含めて 19 トークンを最大文章長とした．BERT の入力として，入力トークンの他にインプットマスクとセグメントマスクという二つのマスクを用意している．インプットマスクは入力文章長が最大文章長よりも短い場合に，パディングされた部分を学習に影響させないためのマスクである．セグメントマスクは一つ目の文章と二つ目の文章を明示するためのマスクである．今回の設定では質問文が一つ目，回答文が二つ目の文章となるため，これらを分けるためにセグメントマスクを用いる．

4.3.2 事前学習

事前学習では BAN モデル全体の学習は行わず，エンコーダである BERT 部分のみパラメータの最適化を行う．毎イテレーションごとに，手法の節で示したようにランダムマスキング処理を入力文章に対して行い，BERT にマスクされた部分の単語を予測させる．この際，損失関数としては Negative Log Likelihood(NLL) を用いた．これは，単純に正解単語のインデックスのみが 1 となる One-Hot ベクトルを教師データとし，モデルの出力となる単語の確率分布との交差エントロピーを損失に反映しているものである．損失として考慮されるのは，マスクされた単語の予測部分のみであり，他単語の復元については考慮されない．オプティマイザには Adam を使用し，パラメータは元論文に準拠して $\beta_1 = 0.9$，$\beta_2 = 0.999$ とした．また，初期学習率は 1×10^{-4} とした．

4.3.3 本学習

本学習では，BAN モデル全体の最適化を行う．入力文章は質問回答のペアを結合した後に，回答部分のみ $[MASK]$ トークンでマスキングを行い，BAN モデルに回答の予測を行わせる．本学習でもオプティマイザは Adam とし，初期学習率は 5×10^{-5} とした．また，学習率の Warmup と Decay も行った．BAN で用いる Co-Attention Map の数は最大の 8 個とした．オプティマイザには Adam を使用し，パラメータは事前学習と同じく $\beta_1 = 0.9$，$\beta_2 = 0.999$ とした．初期学習率は 1×10^{-4} とし，Gradient Clip の値は 1.0，隠れ層の数は 768 とした．BERT モデルは学習率が 0.001 を超えると学習が非常に不安定となるため，幾つかの学習率で実験をし，最も学習が安定する値を選択した．

一方で，比較のために BERT ではなく GRU をエンコーダに用いた BAN では，幾つかハイパーパラメータが異なっている．今回は初期学習率を 7×10^{-4}，隠れ層の数は 1280 とした．また GRU の入力部分の Word Embedding として，大規模コーパスで事前学習

済みの GloVe を利用している．

本学習で各モデルの学習データとして用いているのは VQA データセットの Train Set であり，精度評価は同じく VQA データセットの Validation Set によって行う．また Visual Genome データセットは VQA の分野で学習に加えられることが多いが，今回は使用していない．

4.3.4 結果

本実験によって得られた，VQA データセットの Validation Set における精度での各モデルの比較を表 4.1 に示す．ベースライン手法として，Anderson らによって提案された Bottom-Up Attention を利用したモデルを挙げた．Word Embeddings として事前学習済みの GloVe を利用し，エンコーダとして GRU を用いたオリジナルの BAN は，Bottom-Up Attention モデルと比較して 2% 近く精度を改善している．BAN に事前学習済み BERT を組み込み，ランダムマスキングによる事前学習を行わない我々のモデルはオリジナルの BAN と比較すると精度を 0.61% 改善している．さらに，上記に加えて事前学習でのランダムマスキングによるデータ拡張を行なった我々のモデルは，オリジナルの BAN と比較して 1.24% 精度を改善した．

オリジナルの BAN モデルと，BERT を組み込んだ我々のモデルの学習曲線を比較したものを図 4.4 に示す．オリジナルの BAN は初期学習率が高いため，前半は高いスコアを示している．一方で提案手法は学習が進み Train Loss がある程度下がってからも，オリジナルの BAN と比較して Validation Score が少しずつ上昇している．40 エポック目に学習率が減衰し，全モデルとも Score が同程度上昇している．結果として提案手法はオリジナルの BAN よりも高い精度を示していることがわかる．

結果から読み取れることとして，まず大規模コーパスで事前学習済みの BERT モデルは，GloVe を Word Embedding として用いた GRU モデルと比較してよい特徴を抽出できていると考える．さらに，ランダムマスキングによる我々のデータ拡張手法は，BERT を VQA タスクにより特化した特徴抽出器としてチューニングできており，オリジナルの BAN と比較して 1% 以上の精度向上をもたらした．

表 4.1: VQA2.0 データセットの Validation Set における各手法の精度．ランダムマスキングによる事前学習を行なった BERT をエンコーダとして用いたモデルは，オリジナルの BAN と比較して精度が 1.24% 向上している．

Model	VQA Score
Bottom-Up Attention[8]	63.37
Bilinear Attention Networks(BAN)[5]	65.26
BAN+Transformer-Encoder(ours)	65.87
BAN+Transformer-Encoder+Data Augmentation(ours)	**66.50**

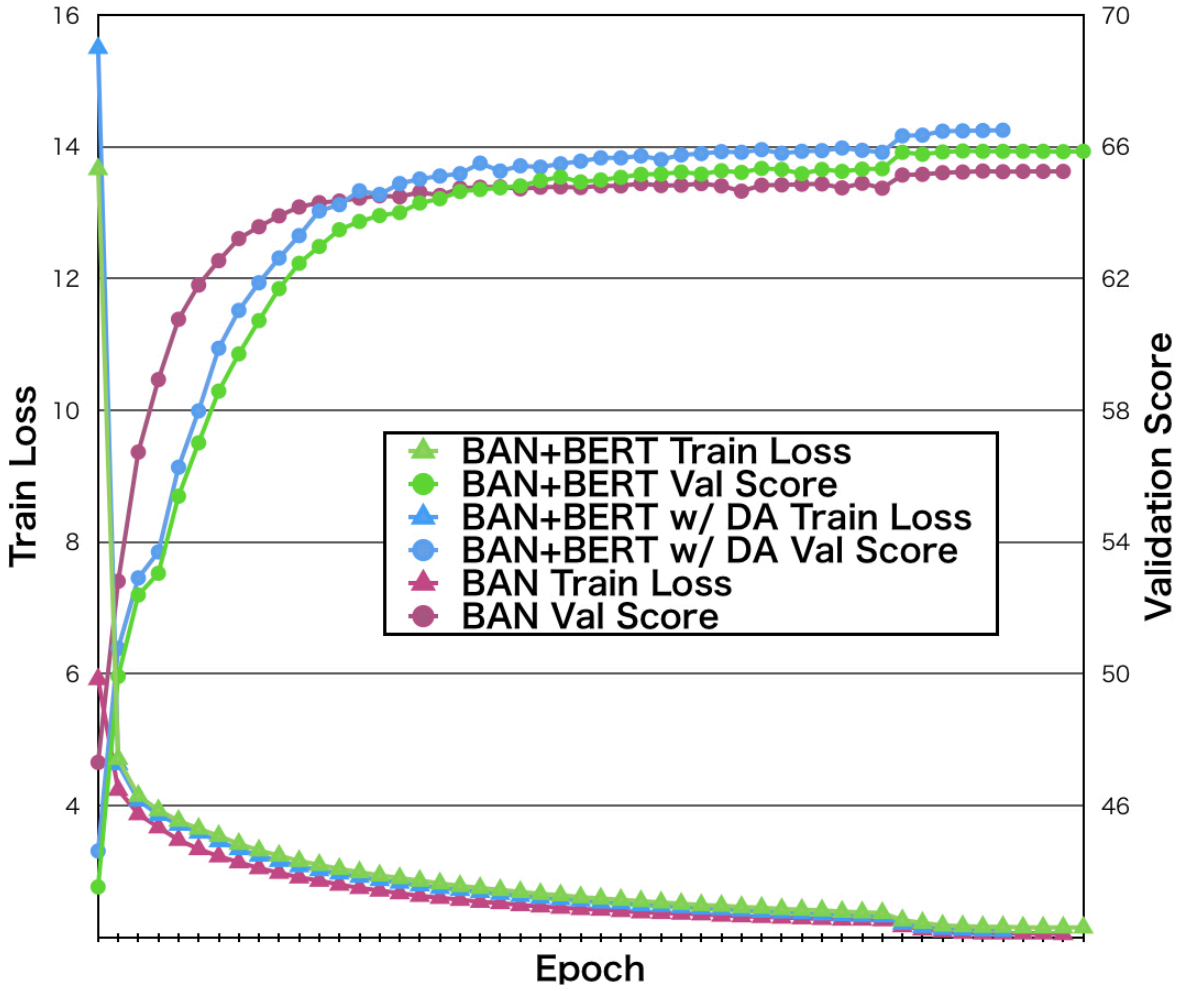

図 4.4: 提案手法とオリジナルの BAN との学習曲線の比較．提案手法では Train Loss がある程度下がってからも Validation Score が少しずつ上昇している．学習率減衰時に全モデルとも Score が同程度上昇し，結果として提案手法はオリジナルの BAN よりも高い精度を示した．

第 5 章

結論

5.1 まとめ

本論文では，Visual Question Answering における現在の課題として，ラベル無し画像データの活用，Data Augmentation（データ拡張)，自然言語のエンコーダを挙げた．これらの課題を踏まえて，我々は半教師あり学習を用いてラベルなし画像の活用の知見を得ることと，VQA に特化したデータ拡張を行い，精度を改善することを目的として，本論文で大きく分けて二つの手法を提案した．

一つは 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回答ペアを生成するモデルの学習を行う手法である．ラベルなし画像に対して様々な質問回答ペアを生成し，それを合成データとして用いることで，半教師あり学習を行うフレームワークを構築した．我々はこの半教師あり学習の精度を質問回答ペア生成モデルの性能として見なし，手法の比較に用いている．本論文ではより詳細な比較を行うため，4 つの生成方法を提案している．具体的には，敵対的学習を用いて生成モデルを作る方法，キャプション生成の手法を組み合わせて質問回答ペアを作る方法，事前学習されたキャプション生成モデルを VQA ドメインに適応させる方法，そしてテンプレートベースで質問回答ペアを得る方法である．我々の手法の新規性は，画像に対応する質問だけでなく質問と回答のペアを同時に生成する点と，質問回答生成手法の評価尺度として半教師あり精度を用いる点である．

実験では，各手法で生成された質問回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，事前学習に VQA データを使ったかキャプションデータを使ったかにかかわらず，ベースラインを上回ったものの，正ラベルのみを用いた学習結果を下回る精度となった．ラベルなし画像に対して生成された質問回

答ペアを分析すると，回答のみが妥当でないなどノイズデータが混じっており，これが精度の低下を招いていた．一方で，キャプションデータセット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．Densecap を用いた手法で生成される質問回答ペアは比較的簡単なものだが，VQA データセットに含まれていない新規な質問解答文をほぼノイズなく生成できていたため，精度が向上したと考えられる．テンプレートベースの手法は，正ラベルのみの精度をわずかに下回った．

もう一つの手法は，VQA モデルにおけるエンコーダとして Transformer を発展させたBERT アーキテクチャを取り入れ，学習時に質問文内の単語をランダムに予測させることでデータ拡張を行う手法である．我々の手法では，学習の際に質問文を 1 単語ずつ RNN に入力し，回答の予測を行う既存手法群とは異なり，質問と回答を結合したものをモデルの入力とし，事前学習の際にランダムに文章中の単語をマスクし，それらを予測させている．この方法によって VQA データをより有効活用し，データ拡張を行っている．具体的には，質問文のみを RNN に入力して一方向の学習を行うのではなく，質問文と回答文をどちらも BERT に入力し，マスク部分を周辺情報から予測（双方向学習）することで，より VQA に向いたよい特徴量を得ることができる．この手法の新規性は，入力文章のエンコーダとして Transformer を発展させた BERT アーキテクチャを採用した点と，学習時に入力文章をランダムにマスクし，画像だけでなく自然言語側でもデータ拡張を行う点が挙げられる．また我々のデータ拡張手法は，VQA 以外の様々な Vision Language Task のモデルにも適用可能であり，汎用性が高い．

Transformer を用いたデータ拡張の実験では，大規模コーパスで学習された BERT モデルは VQA においてもよりよい言語特徴量の抽出に役立つことがわかり，マスキングによる事前学習を組み合わせることでさらに VQA 精度を高めることができた．従来は GloVe 等の学習済み単語ベクトルを Embedding レイヤーの初期値とした LSTM や GRU が VQA モデルにおけるエンコーダの主流であったが，今回の結果によって BERT 等の Transformer 系言語モデルが VQA において高い性能を示すことがわかった．

結果として，本論文の貢献は以下のようにまとめられる．

- VQA タスクにおいて，半教師あり学習の設定でラベルなし画像データを活用するフレームワークを提案した．
- VQA の半教師あり学習精度を，質問回答生成手法の新たな評価尺度として捉えた．
- キャプションアノテーションを活用した合成質問回答を生成することで，半教師あり VQA 精度を向上させた．
- VQA の分野で用いられていなかった言語モデル (Transformer, BERT) を取り入

れ，精度を向上させた．

- VQA 以外の Vision Language Task にも適用可能な新たなデータ拡張の手法を提案し，さらに精度を向上させた．

5.2 今後の展望

本節では本論文で提案したアプローチの，今後の展望について述べる．

半教師あり学習のための質問回答生成については，Densecap を用いた手法を手がかりにキャプションデータセット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回答を生成する方法，すなわち合成データ中のノイズを減らす手法の検討が考えられる．さらに，Module Network[26] やテンプレート生成 [47] によるタスクの細分化などが考えられる．また質問解答ペア生成手法の評価尺度として，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や，従来の VQG の研究で用いられている言語スコア BLEU[48] による比較も検討したい．

BERT アーキテクチャを用いたデータ拡張については，本論文で行ったような最も一般的な VQA 精度の比較以外にも，VQG や Visual Dialog のタスクへ適用し，既存手法との比較を行いたい．また，今回は BERT の事前学習の際に Masked LM モデルのファインチューニングを行うためランダムマスキングを行い，その後に VQA データセットを用いた本学習を行なったが，直接 BERT を用いて回答部分の予測を行う End-to-End モデルの性能も比較したい．

参考文献

[1] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1080–1089. IEEE, 2017.

[2] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–10, 2018.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[4] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4565–4574, 2016.

[5] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *arXiv preprint arXiv:1805.07932*, 2018.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[7] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

[8] 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 vqa. *arXiv preprint arXiv:1707.07998*, 2017.

[9] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pp. 451–466. Springer, 2016.

[10] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, No. 99, pp. 1–13, 2018.

[11] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 1802–1813, 2016.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[13] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Gated feedback recurrent neural networks. In *International Conference on Machine Learning*, pp. 2067–2075, 2015.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pp. 15–29. Springer, 2010.

[16] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

[17] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pp. 2048–2057, 2015.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common

objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

[19] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pp. 1057–1063, 2000.

[20] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[21] 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 *The IEEE International Conference on Computer Vision (ICCV)*, Vol. 2, 2017.

[22] Alexander Mathews, Lexing Xie, and Xuming He. Semstyle: Learning to generate stylised image captions using unaligned text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8591–8600, 2018.

[23] Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[24] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. *European Conference on Computer Vision*, 2018.

[25] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard Product for Low-rank Bilinear Pooling. In *The 5th International Conference on Learning Representations*, 2017.

[26] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 39–48, 2016.

[27] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017.

[28] Unnat Jain, Ziyu Zhang, and Alexander G Schwing. Creativity: Generating diverse questions using variational autoencoders. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, pp. 6485–6494, 2017.

[29] Kohei Uehara, Antonio Tejero-De-Pablos, Yoshitaka Ushiku, and Tatsuya Harada. Visual question generation for class acquisition of unknown objects. In *European Conference on Computer Vision*, pp. 492–507. Springer, 2018.

[30] Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, Jianfeng Lu, and Anton Van Den Hengel. Goal-oriented visual question generation via intermediate rewards. In *European Conference on Computer Vision*, pp. 189–204. Springer, 2018.

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

[32] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. Visual question generation as dual task of visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6116–6124, 2018.

[33] Unnat Jain, Svetlana Lazebnik, and Alexander G Schwing. Two can play this game: visual dialog with discriminative question generation and answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5754–5763, 2018.

[34] Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton van den Hengel. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6106–6115, 2018.

[35] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *European Conference on Computer Vision*, pp. 160–178. Springer, 2018.

[36] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. Building generalizable agents with a realistic and rich 3d environment. *arXiv preprint arXiv:1801.02209*, 2018.

[37] Abhishek Das, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Neural modular control for embodied question answering. In *Conference on Robot Learning*, pp. 53–62, 2018.

[38] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations.

In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Vol. 1, pp. 2227–2237, 2018.

[39] 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.

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

[41] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 1983–1991. IEEE, 2017.

[42] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

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

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[45] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[46] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pp. 19–27, 2015.

[47] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. *arXiv preprint arXiv:1803.09845*, 2018.

[48] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

発表文献

国内会議

[1] 築山将央, 伊神大貴, 入江豪, 相澤清晴. “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” 第 17 回 情報科学技術フォーラム (FIT), 2018.

[2] 築山将央, 伊神大貴, 入江豪, 相澤清晴. “視覚的質問応答における多種質問回答生成と Transformer を用いたデータ拡張” 電子情報通信学会 画像工学研究会 (IE), 2019 (発表予定) .

受賞

[3] 築山将央, 伊神大貴, 入江豪, 相澤清晴. 第 17 回 情報科学技術フォーラム (FIT), FIT 論文賞, 2018.

謝辞

指導教員である相澤清晴教授には，二年間 Open World 班のミーティングや全体ミーティングに加え，発表原稿の執筆時なども手厚い指導を賜りました．とても出来が良い学生とはいえない僕に対して，辛抱強く指導してくださり，大変感謝しています．

山﨑俊彦准教授には，全体ミーティングでの進捗発表やリハーサルの場において非常に的確なアドバイスを頂きました．また研究室合宿や懇親会の際は気さくに話しかけて頂きました．ありがとうございました．

学術支援職員の松林さんと江川さんには，主に事務手続の面で何度もお世話になりました．研究室ではお二人ともとても暖かく接して頂き，本当にありがとうございました．

共同研究を行いました NTT コミュニケーション科学基礎研究所の入江豪氏には，共同研究ミーティングの際に非常に専門的なアドバイスを幾つも頂きました．

Open World 班の先輩である伊神大貴さんには，本当にお世話になりました．僕の研究における指針を何度も示してくれ，手法の細かい部分についてのどんな相談にも乗っていただきました．研究以外でも，色々な活動に付き合って頂き，楽しく過ごせました．本当に感謝しています．

Open World 班の先輩であった竹木章人さん，また後輩である田中大揮君，郁青君，浅井明里さんは，毎回の班ミーティングで非常にレベルの高い進捗報告を聞き，研究の面で大変刺激を受けました．

僕が修士一年の時に修士二年の先輩であった小川徹さんには，研究のアイデアを頂いたり，実装についての様々な質問にいつも答えて頂きました．

博士課程の先輩である古田諒佑さん，井上直人さんは，研究室内での論文の輪読会を主催してくださいました．参加できない回もありましたが，大変ためになりました．ありがとうございました．

同じく博士課程の先輩である橋本侑樹さんには，発表資料やプレゼンテーションスライドなどを添削して頂き，毎度的確なアドバイスを頂きました．

同期の皆は，各々個性が強いながらも全体として非常に仲がよく，研究室での居心地が大変良かったです．同期仲が良かったことは，僕の修士生活の心の支えでもありました．感謝を込めて，一人ひとりに謝辞を述べたいと思います．大渕友暉君は，同期での懇親会の幹事を毎度務めてくれ，飲み会の場を盛り上げてくれました．小川将範君は，同じくキャンパス付近の一人暮らしだったこともあり，いつも雑談や相談に付き合ってくれました．高田祐樹君は，学部二年の頃から付き合いがあり，実験や学生間の交流の時にもお世話になりました．中村遵介君は，内定先の部署まで同じということもあり，よく研究の相談に乗ってくれました．成田嶺君は，趣味が近いこともありいつも雑談に付き合ってくれ，色々と教えてくれました．張軼威さんは，特に個性が強くマンネリ化しがちな研究生活にいつも刺激をくれました．合田悠治君は，普段寡黙ながらもユーモアセンスがあり，懇親会の際などに笑わせてくれました．大坪篤志君は，一緒に修了することはできませんでしたが，中退後も同期間の懇親会に何度も来てくれました．

研究室に行った際，いつも夕飯に付き合って頂いた坪田亘記君，小川将範君，張軼威さん，小川徹さん，石見和也さん，宮田真里さんに感謝いたします．

その他の相澤研究室，山﨑研究室の全ての後輩，先輩，研究員の方々にも感謝いたします．

最後に，大学入学時は学部で卒業するはずだった僕が理転し大学院まで進むことになっても，一切渋らずに喜んで援助をしてくれ，またいつも暖かい言葉をかけてくれた両親や家族に感謝を述べたいです．

本当にありがとうございました．

2019 年 1 月 31 日

築山 将央