FigureQA タスクにおける抽象画像を考慮したアプローチ

坂本　凜†　酒井　哲也†

† 早稲田大学基幹理工情報理工学科　〒 169–8555 東京都新宿区大久保 3-4-1

E-mail: †ringring@suou.waseda.jp，†tetsuyasakai@acm.org

あらまし　画像とその画像についての質問が与えられた時，正しい回答を求める Visual Question Answering (VQA) というタスクがある．VQA の一種である FigureQA は棒グラフや円グラフ等のグラフの画像と，Yes/No で回答できる問題文が与えられる推論タスクである．棒グラフや円グラフ等の抽象画像 (イラストレーション) は自然画像と異なる特徴を持ち，自然画像を用いたタスクで優れた結果を出したモデルを適用しても，同程度の正解率は出ない．本研究では抽象画像であることを考慮した手法の提案を行い，FigureQA タスクに適用し評価行ったところ，FiLM がもっとも高い正解率となった．各モデルの判断が上手くいかない例をあげ，そのモデルの構造的な要因を分析した．

キーワード　FigureQA，VQA

1 はじめに

この先，多くの人間がウェアラブルデバイスなどから自身のライフログを取得し，そのデータを元にしたグラフを見ることが考えられる．その際，グラフの画像から特徴を読み取ることになる．グラフの画像キャプション生成技術があれば誰でもグラフから得られる特徴を読み取ることができる．グラフは数量の情報を効率よくまとめた表現であり，ライフログ表示の他にも教科書，研究論文，インターネットの記事等のメディアからも取得できる．しかし，必ずしもそのグラフを作成するための元のデータを入手できるかというと，そうとは限らない．たとえ元のデータがあったとしても，グラフの画像キャプション生成技術には，データそのものを解析するアプローチとは異なる利点があるかもしれない．

画像キャプションの生成では画像，言語の 2 つのモーダルを扱う．同じように，画像，言語の 2 種類の情報を取り扱うマルチモーダルなタスクとして Visual Question Answering (VQA) [1] が存在する．VQA はある画像と，その画像に関する質問が与えられた時，その質問の正しい回答を得るためのタスクである．

FigureQA [12] は VQA において関係推論に注目したデータセットである．図 1 に示すように画像として折れ線グラフ，点線グラフ，水平棒グラフ，垂直棒グラフ，円グラフの 5 種類のグラフと，それに対するプロットと全体，プロット同士の関係性に関する質問文が与えられている．トレーニングセットとして 100,000 個の画像と 1,327,368 個の質問と答えのペアが与えられているデータセットである．質問文は *Is X the maximum ?* や *Does X intersect Y?* といった Yes/No で答えられるものである．グラフの画像が与えられて，その画像に関する質問に回答することができるようになれば，先に述べたようなグラフ画像のキャプション生成につなげることができる．

FigureQA は関係推論の VQA データセットであるが，これに似たデータセットに CLEVR [7] がある．図 2 に示すように，CLEVR の画像は球や立方体などの物体が複数配置された画像である．この 2 つのデータセットはどちらも人工的に作られた画像が用いられているが，CLEVR の画像が自然画像なのに対して，FigureQA の画像は抽象画像である．CLEVR において人間の正解率を上回った Relation Network [18] を使用しても，十分な正解率は得られない [12]．これは FigureQA の画像が抽象画像であり，自然画像と特徴が異なるためと考えられる．

そこで本研究では，与えられるデータが抽象画像であることを考慮したアプローチを提案する．

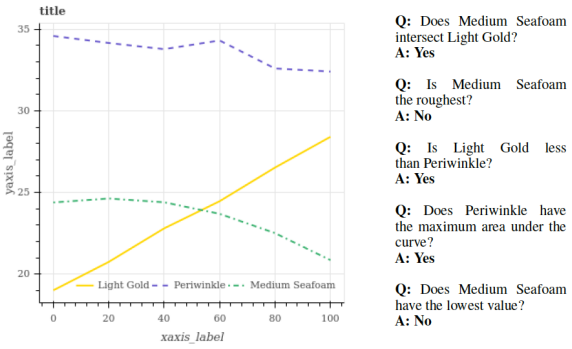


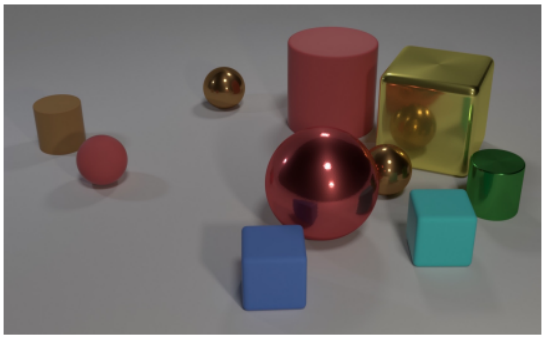
図 1　FigureQA データセットの画像と質問文のサンプル ([12] より転載)

2 関連研究

2.1 Relation Network

FigureQA のベースラインとして使われている Relation Network は，関係推論のための構造を持つニューラルネットワークのモジュールである．式 1 に示すように入力はオブジェクトの集合 $O = \{o_1, o_2, \cdots, o_n\}$ である．o_i は i 番目のオブジェクトである．f_ϕ と g_θ はそれぞれパラメータが ϕ と θ の関数である．

$$\mathrm{RN}(O) = f_\phi \left(\sum_{i,j} g_\theta \left(o_i, o_j \right) \right) \tag{1}$$



Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or metal things?

図 2　CLEVR データセットの画像と質問文のサンプル ([7] より転載)

オブジェクト数を N とするとオブジェクトのペアは N^2 個あり，その回数 g_θ の演算を行う．これによってオブジェクト同士の関係性を学習させる．ここでオブジェクトとは画像を Convolutional Neural Network (CNN) [14] 層に入力して出力した特徴マップの一つ一つの要素を指す．たとえば FigureQA のベースラインでは，CNN 層は 8×8 の特徴マップを出力するので 64^2 回 g_θ の計算を行う．f_ϕ と g_θ は FigureQA のベースラインでは多層パーセプトロン [17] が採用されている．Santoro ら [18] の研究や FigureQA のベースラインでは，質問文の特徴抽出に Long Short-Term Memomry [4] を利用している．

2.2　FiLM モデル

FiLM : Feature-wise Linear Modulation [16] は CLEVR で CNN+LSTM+Relation Network モデルよりも高い正解率を出したモデルである．FiLM モデルは後述する提案手法で利用している．Perez ら [16] は CLEVR の質問および画像を FiLM モデルの入力とするために，Johnson ら [8] が行なった前処理を利用している．質問文を Gated Recurrent Unit (GRU) [2] 層に入力し出力した特徴マップを全結合層の入力とする．更に，全結合層の出力を用いて FiLM 層の出力を式 2 で得る．

$$\mathrm{FiLM}(F_{i,c}|\gamma_{i,c},\beta_{i,c}) = \gamma_{i,c}F_{i,c} + \beta_{i,c} \qquad (2)$$

ここで，$(\gamma_{i,c},\beta_{i,c})$ は全結合層の出力で，$F_{i,c}$ は i 番目の入力の c 番目の特徴または特徴マップを指す．画像は CNN 層に通したあと N 個の FiLM-ed residual block (ResBlock) というモジュールに通す．ResBlock は 1×1 の畳み込み層，活性化関数 ReLU，3×3 の畳み込み層，Batch Normalization，FiLM 層，活性化関数 ReLU から成り立つ．1×1 の畳み込み層の後の活性化関数 ReLU の出力と FiLM 層の後の活性化関数 ReLU の出力を連結する．連結された出力が ResBlock の出力となる．N 個目の ResBlock の出力を分類器の入力として質問文の回答を推測する．

2.3　SENet

Squeeze-and-Excitation (SE) Networks [5] は，ILSVRC 2017 で画像分類のトップとなった手法である．SENet は，SE Block という Tensor のチャンネル間の関係性に注目したアーキテクチャを導入している．SE Block は Global Average Pooling (GAP) 層，全結合層，ReLU，全結合層，Sigmoid 関数から成り立ち，畳み込み層の後に接続される．まず式 3 で表される GAP 層を通すことでチャンネルごとの統計情報を生成する．高さ H，幅 W，チャンネル数 C のテンソル $\boldsymbol{U}$ の i 番目のチャンネル $\boldsymbol{u}_{c_i}$ を関数 F の入力として z_{c_i} を出力する演算である．すなわち $\boldsymbol{U} \in \mathbb{R}^{H \times W \times C}$ を $\boldsymbol{Z} \in \mathbb{R}^{1 \times 1 \times C}$ に圧縮する．生成されたチャンネルごとの統計情報を利用してチャンネルごとの依存性を完全に捉えるために，Sigmoid 関数を用いる．モデルが一般性を得るために GAP 層と Sigmoid 関数の間に，2 つの全結合層と ReLU でボトルネックを形成する．最後に畳み込み層とその後に接続される SE Block の出力の要素積をとる．提案手法では，この SE Block を使用している．

$$z_{c_i} = \mathbf{F}(\mathbf{u}_{c_i}) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_{c_i}(i,j) \qquad (3)$$

2.4　類 似 研 究

FigureQA 以外にもグラフ画像理解への研究が存在する．Jung ら [10] の研究では，グラフの種類を分類した後，データの抽出を行う手法を提案している．Cliché ら [3] の研究では散布図からデータを抽出する手法を提案している．また FigureQA のように図に関する QA では Kembhavi ら [13] が Textbook Question Answering (TQA) データセットを導入している．TQA データセットは，文章とダイアグラムと画像とそれに対する質問文が与えられているデータセットである．Kafle ら [11] が導入した DVQA (Data Visualization Question Answering) データセットは FigureQA と同じように画像としてグラフ画像が用いられているが，質問文に Yes/No 疑問文に限らず $5W1H$ の疑問文も採用されている．類似したデータセットに，Siegel ら [20] が導入した FigureSeer データセットがある．FigureQA と違い研究論文に載っている図を用いている．実際のデータのプロットである利点があるが，データセットのサイズが小さいという欠点を持つ．

2.5　画像キャプション

画像キャプションは，CNN を用いて画像の特徴を抽出し，LSTM によってキャプションを生成するのが標準的な手法である．Shin ら [19] の研究では，VQA のタスクを用いて，画像のキャプション生成を行なっている．Visual Question Generation (VQG) モジュールによって生成された質問文を VQA モジュールの入力とする．これで得られた回答の文章を再構成することで画像のキャプションを生成する．MS COCO [15] データベースで学習した従来のキャプショニング，SIND データセット [6] で学習したキャプション，DenseCap [9] によって生成されたキャプションとの比較を行なった． Amazon Mechanical Turk で雇用した働き手が，画像 1 枚につき 2 人，計 5000 枚の画像を，Shin らが画像キャプションの評価に必要と考えられ

る 5 つの指標 (Diversity, Interesting, Accuracy, Naturalness, Expressivity) に基づいて評価した．その結果 VQA を用いた手法が他の手法より評価が上回った．このことからも本研究が，グラフ画像のキャプション生成に繋がるといえる．

3 提 案 手 法

提案手法の概要を図 3 に示す．FigureQA はグラフの画像と，それに関する質問文が与えられ，その質問文の Yes,No の答えを得るタスクである．提案手法は GRU 層と CNN 層，後述する 4 つの SE ResBlock, 分類器から成り，GRU 層に質問文を，CNN 層に画像を入力する．SE ResBlock で GRU 層の出力である質問文の特徴と CNN 層の出力である画像の特徴を結合し，その出力を分類器に入力する．分類器では最終的に 2 クラスに分類され，それぞれ質問文の解答の Yes，No に対応する．

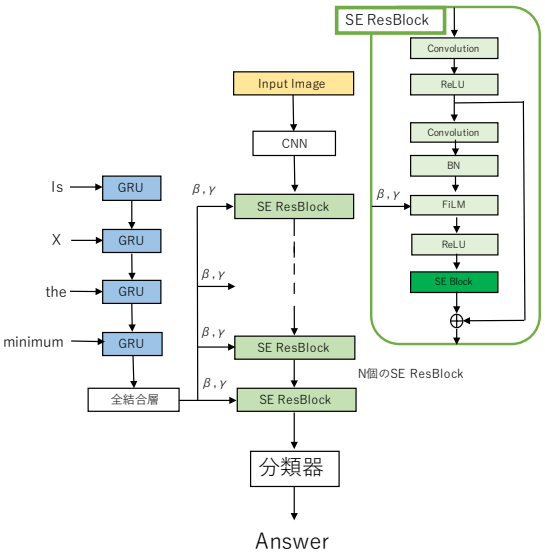


図 3　提案手法　：SE FiLM Network

3.1　質問文の特徴抽出

質問文は単語の系列データなので，特徴抽出にはゲート付き Recurrent Neural Network である LSTM や GRU などの層を使用することが考えられる．学習時間を短くすることを考慮し，ここでは GRU を採用する．GRU のユニット数は 256 で各系列の最後の出力を利用する．式 2 で定義される FiLM 層を生成するために全結合層への入力とする．全結合層の出力のサイズは ($ResBlock$ の数 N) $*2*$ (画像の特徴のチャンネル数 C) である．本研究では，4 つの ResBlock を用い，画像のチャンネル数は 64 とするので，512 である．全結合層の出力から式 2 の $(\gamma_{i,c}, \beta_{i,c})$ を求め FiLM 層を作る．

3.2　画像の特徴抽出

画像の特徴と質問文の情報を結合する前段階として，CNN 層に 256×256 のサイズの画像を入力として，特徴を抽出する．ここで CNN 層は 5 つの畳み込み層から成る．各畳み込み層では 3×3 サイズのフィルターを用い，ストライド 2 でゼロパディングし，チャンネル数 64 の特徴を出力する．それぞれの層では ReLU で活性化し，Batch Normalization を行う．

3.3　SE ResBlock

SE ResBlock 内の FiLM 層で画像の特徴と質問文の特徴を結合する．4 つの連続する SE ResBlock の最初の SE ResBlock への入力は，CNN に通した画像の特徴である．SE ResBlock は 1×1 の畳み込み層，活性化関数 ReLU，3×3 の畳み込み層，Batch Normalization (BN)，FiLM 層，活性化関数 ReLU，SE Block から成り立つ．SE Block は 3.5 節で後述する．1 つ目の活性化関数 ReLU の出力と SE Block の出力の要素和をとることで残差接続し，勾配消失を防ぐ．

3.4　FiLM　層

FiLM 層では以下に記述するプロセスで質問文の特徴と，画像の特徴の結合を行う．質問文の特徴抽出に用いた，GRU と全結合層の出力 $\boldsymbol{x} \in \mathbb{R}^{N*2*C}$ を変形して，$\boldsymbol{x} \in \mathbb{R}^{N\times(2*C)}$ の形にする．変形させた $\boldsymbol{x}$ の i 行目の要素 $\boldsymbol{x}_{i,\bullet} \in \mathbb{R}^{(2*C)}$ を取り出し，2 分割し，$\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i \in \mathbb{R}^C$ を得る．i 番目の SE Resblock 内の FiLM 層の入力 $\boldsymbol{F}_i$ の形は，高さを H, 幅を W, チャンネル数を C とすると，$\boldsymbol{F}_i \in \mathbb{R}^{H\times W\times C}$ である．$\boldsymbol{x}$ を変形させて得た $\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i$ が，$\boldsymbol{F}_i$ の形になるようにそれぞれタイルして形を合わせ $\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i \in \mathbb{R}^{H\times W\times C}$ とする．そこで $\boldsymbol{\gamma}_i \circ F_i + \boldsymbol{\beta}_i$ の演算を行い FiLM 層の出力とする．ただし $\circ$ は 2 つのテンソルの要素積の演算子である．

3.5　SE Block

SE Block の概略を図 4 に示す．SE Block では入力のチャンネル間の依存関係を抽出する．SE Block は，Global Average Pooling 層，全結合層，活性化関数 ReLU，全結合層，sigmoid 関数から成り立つ．高さ H，幅 W，チャンネル数を C とすると，入力 $\boldsymbol{s}$ は $\boldsymbol{s} \in \mathbb{R}^{H\times W\times C}$ である．Global Average Pooling 層では，入力の各チャンネル毎の $H \times W$ 個の要素を，その要素の平均値に圧縮する．つまり $\boldsymbol{s'} \in \mathbb{R}^{1\times 1\times C}$ の形のテンソルを出力する．続いて，全結合層に入力し，チャンネル数を C から C/r に圧縮する．r は圧縮率で, [5] では $r = 16$ が採用されていたが，本研究では，圧縮する対象のチャンネル数が 64 と少ないので，$r = 4$ とした．その後，活性化関数 ReLU を用いて非線形化し，再び全結合層を通して，チャンネル数を C に戻す．Sigmoid 関数を通すことでチャンネルごとの関係性を抽出し，最後に SE Block の入力のテンソルと要素積をとることで，入力にチャンネル間の関係を反映させる．

3.6　分　類　器

分類器は最後の SE ResBlock の出力を入力とし，最終的に 2 クラスに分類する．分類器は畳み込み層と 3 つの全結合層からなる．畳み込み層のフィルターサイズは 3×3 で，ストライド 1，出力するチャンネル数は 128 である．ReLU で活性化した後 Max Pooling を行う．3 つの全結合層の出力のサイズは，そ

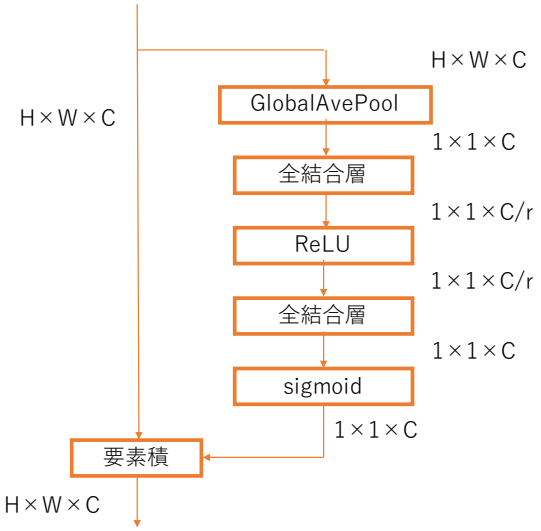


図 4　SE　Block

れぞれ 512, 512, 2 である．1 つ目の全結合層の後ろに活性化関数 ReLU，2 つ目の全結合層では dropout と ReLU を用いる．

4　評 価 実 験

4.1　データセット

データセットには Maluuba が公開している FigureQA を用いる [1]．FigureQA は人工的なデータセットで，数量データを表現する折れ線グラフ，点線グラフ，水平棒グラフ，垂直棒グラフ，円グラフの 5 種類のグラフと，それに対するプロットと全体，プロット同士の関係性に関する Yes/No で答えられる質問文が与えられている．図 5 に 5 種類のグラフの画像の例を示す．各画像は 256 × 256 のサイズで，RGB の 3 チャンネルであり，PNG 形式でエンコードされている．表 1 に示すように，質問文は 15 種類のテンプレートのうちのいづれかであり，垂直棒グラフ，水平棒グラフ，円グラフには question_id が 0 – 5 の質問文が，折れ線グラフと点線グラフには 6 – 14 の質問文が与えられている．

FigureQA のデータセットの統計情報を表 2 に示す．トレーニングセットは 100, 000 個の画像に対し 1, 327, 368 個の質問が与えられている．本研究では，20, 000 個の画像と，それに対する 265, 798 個の質問文が与えられている Validation2 をバリデーションデータに用いた．Test1,Test2 の正解データは現在公開されていない．Github で公開されている学習モデルの評価プログラム [2] を使用して，学習したモデルの Test1,Test2 に対する解の予測値を csv ファイルに出力し，Maluuba に出力されたファイルを送って評価していただくことで，正解率を得ることができる．本研究ではテストセットに Test2 と，検証データに使用しておらず，なおかつ正解のアノテーションがされている Validation1 を使用した．Test2 は 20, 000 個の画像に対し，265, 402 個の質問文が与えられていて，Validation1 は 20, 000 個の画像に対し，265, 106 個の質問文が与えられている．

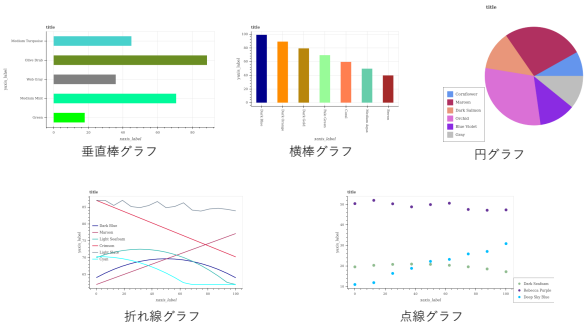


図 5　FigureQA の 5 種類のグラフのサンプル [3]

表 1　FigureQA の質問文のテンプレート

question_id	質問文のテンプレート
0	Is X the minimum?
1	Is X the maximum?
2	Is X the low median?
3	Is X the high median?
4	Is X less than Y?
5	Is X greater than Y?
6	Does X have the minimum area under the curve?
7	Does X have the maximum area under the curve?
8	Is X the smoothest?
9	Is X the roughest?
10	Does X have the lowest value?
11	Does X have the highest value?
12	Is X less than Y?
13	Is X greater than Y?
14	Does X intersect Y?

表 2　FigureQA の統計情報

データ	画像	質問文	回答の有無	Color Scheme
Train	100,000	1,327,368	Yes	Scheme 1
Validation1	20,000	265,106	Yes	Scheme 1
Validation2	20,000	265,798	Yes	Scheme 2
Test1	20,000	265,024	No	Scheme 1
Test2	20,000	265,402	No	Scheme 2

1：https://datasets.maluuba.com/FigureQA
2：https://github.com/vmichals/FigureQA-baseline
3：本論文の図 5, 7, 8, 9 は全て https://datasets.maluuba.com/FigureQA のデータセットから抽出した．

4.2 モデルの実装と実行環境

Github で公開されている FigureQA のベースラインのプログラム [2] が Tensorflow[4] で実装されているので，Python3.6 と Temsorflow1.6 で実装した．実行環境は Amazon Web Service[5] のサーバー上で，CUDA9.0，cuDNN7.0.4 の環境下で Tesla V100 を 1 個を使用している．

4.3 比 較 手 法

比較手法は Relation Network, FiLM と提案手法である SE FiLM の 3 つでそれぞれ最適化関数には Adam，損失関数には Softmax-Cross-Entropy を用いていて学習率は 0.00025 である．Relation Network は，Github で公開されている FigureQA のベースラインのプログラム [2] を使用した．FiLM は図 6 に示すように，SE FiLM との相違点は SE Block の有無である．バッチサイズは 160 で統一した．トレーニングセットは $1,327,368$ 個の画像と質問文のペアがあるので，1epoc は $8,296.05$ バッチ必要となる．

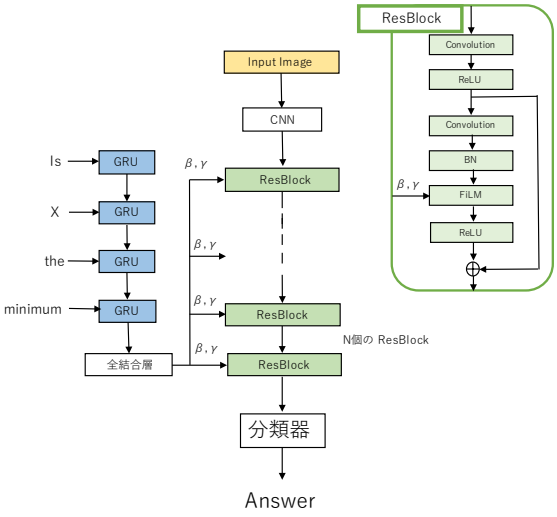


図 6　比較手法：FiLM Network

4.4 評 価 方 法

評価方法は式 4 で表される正解率を用いる．T は正解のデータのラベルと，モデルの出力のラベルが等しい質問の個数，N は質問の総数を意味する．

$$accuracy = \frac{T}{N} * 100 \tag{4}$$

5 結　　果

5.1 テストセット：Test2 における正解率

まず正解データが公開されていない Test2 を用いてテストしたところ，表 3 に示す結果が得られた．Relation Network (RN)，FiLM，SE FiLM それぞれに対し，各カラムは，水平棒グラフ (vbar)，垂直棒グラフ (hbar)，円グラフ (pie)，折れ線グラフ (line)，点線グラフ (dot_line) に関する問題ごとの正解率，全ての問題 (overall) に対する正解率を示す．ここで，各カラムにおいて，もっとも正解率の高い結果を太字で示した．また 2 番目に正解率の高い結果を下線で示した．この表をみてわかるように，FiLM は全てのカラムに置いてもっとも高い正解率となっている．提案手法である SE FiLM はベースラインである RN に比べ，2 種類の棒グラフではそれぞれ 6.27%, 9.63% 高い正解率が出ているものの，円グラフ，折れ線グラフ，点線グラフに関しては 1.19%, 3.30%, 3.01% 低い正解率となっている．

表 3　Test2 における正解率

	vbar	hbar	pie	line	dot_line	overall
RN	79.80	75.13	76.76	63.03	63.85	70.77
FiLM	**88.56**	**87.74**	**77.74**	**69.58**	**73.21**	**78.50**
SE FiLM	86.07	84.76	75.57	59.73	60.84	71.90

5.2 テストセット：Validation1 における正解率

Test1，Test2 は正解データが公開されていないので，より詳細な分析を行うことができない．正解がアノテーションされていて，かつ検証セットとして使用していない Validation1 をテストセットに使用することで，RN，FiLM，SE FiLM の予測値と正解の値との詳細な比較を行う．

5.2.1 SE FiLM と他手法の比較

SE FiLM と RN の予測値の相違を表 4 に示す．表の C_*, C_R と C_R は，それぞれ正解のラベル，RN の予測値，SE FiLM の予測値を意味する．RN だけが正解しているデータは $37,100$ 個，SE FiLM だけが正解しているデータは $37,521$ 個あることがわかる．また，SE FiLM と FiLM の予測値の相違を表 5 に示す．表の C_*, C_F, C_R は，それぞれ正解のラベル，FiLM の予測値，SE FiLM の予測値を意味する．FiLM だけが正解しているデータは $45,487$ 個，SE FiLM だけが正解しているデータは $28,213$ 個あることがわかる．なお，3 つのモデルのうち，少なくとも 1 つのモデルは正解した質問は 250,331 個であり，これは Validation1 の質問のうち 94.43% となる．

表 4　正解と SE FiLM と RN の比較

	$C_R=0$		$C_R=1$		合計
	$C_S=0$	$C_S=1$	$C_S=0$	$C_S=1$	
$C_*=0$	77,458	19,711	18,259	17,125	132,553
$C_*=1$	16,473	19,262	17,389	79,429	132,553
合計	95,214	38,973	35,648	96,554	265,106

5.2.2 質問の種類ごとの正解率

テストセット Validation1 について 3 つのモデルの各質問の種類ごとの正解率を表 6 に示す．正解率がもっとも高い結果を太字，2 番目に高い結果を下線で示した．FiLM は question_id が $1, 8, 9$ 以外の質問ではもっとも正解率が高く，$1, 8, 9$ の質問

4：https://www.tensorflow.org/?hl=ja
5：https://aws.amazon.com/jp/

表 5　正解と SE FiLM と FiLM の比較

	$C_F=0$		$C_F=1$		合計
	$C_S=0$	$C_S=1$	$C_S=0$	$C_S=1$	
$C_*=0$	82,505	24,334	13,212	12,502	132,553
$C_*=1$	12,709	15,001	21,153	83,690	132,553
合計	95,214	39,335	34,365	96,192	265,106

も 2 番目に高い．その $1, 8, 9$ は RN がもっとも高く，$6, 7, 10, 11$ は RN が 2 番目に高い．SE FiLM は $0, 2, 3, 4, 5, 12, 13, 14$ の質問で 2 番目に正解率が高い．

表 6　各問いの種類毎の SE FiLM の結果

question_id	RN	FiLM	SE FiLM
0	82.16	**86.72**	84.79
1	**91.80**	91.54	90.78
2	86.60	**92.37**	91.04
3	86.82	**91.91**	91.09
4	71.46	**76.52**	74.84
5	72.16	**77.65**	75.23
6	61.64	**80.05**	57.94
7	68.35	**78.00**	59.44
8	**60.36**	59.87	58.78
9	**60.10**	60.09	58.67
10	65.58	**77.30**	60.88
11	69.37	**79.68**	61.32
12	61.10	**73.61**	62.81
13	63.00	**73.71**	63.28
14	61.44	**74.06**	63.34

6　考　　察

6.1　RN と SE FiLM の比較

表 3 に示したようにテストセットが Test2 のとき，水平棒グラフと垂直棒グラフの正解率は，SE FiLM が $6.27\%, 9.63\%$ 高いのに対し，円グラフと折れ線グラフ，点線グラフの正解率は RN が $2.49\%, 2.98\%, 2.17\%$ 高くなっている．全体の正解率は SE FiLM は 1.13% 高くなっている．表 4 に示したようにテストセットが Validation1 のとき，SE FiLM だけが正解している問題が $37,521$ 個あり，RN だけが正解している質問は $37,100$ 個ある．これは Validation1 の全 $265,206$ 問のうち，14.15%と 13.99%にあたり，どちらの場合も無視できない大きさであるといえる．一見，グラフの種類によってモデルの優劣が分かれるように考えられるが，表 6 に示したように，RN は question_id が 1 の問題で SE FiLM の正解率を上回っている．これは円グラフと FiLM の方が正解率が高かった棒グラフの質問である．また， FiLM も折れ線グラフと点線グラフの質問である question_id$12, 13, 14$ の質問で RN の正解率を上回っている．

6.2　FiLM と SE FiLM の比較

表 3 に示したようにテストセットが Test2 のとき，5 種類全ての種類のグラフにおいて FiLM の方が正解率が高くなっている．2 種類の棒グラフと円グラフの正解率の差は $2.49\%, 2.98\%, 2.17\%$ となっているが，折れ線グラフと点線グラフに至っては $9.85\%, 12.37\%$ と大きな差が生じてしまっている．一見これだけ見ると，SE FiLM モデルは FiLM モデルよりも優っている点が無いように見受けられる．しかし表 5 を見ると，テストセットが Validation1 のとき，SE FiLM だけが正解している質問は $28,213$ 個あり，これは質問全体の 10.64% に値する．SE FiLM のみ予測ができている質問が 10.64% あるということは，SE FiLM にもモデルの利点があるといえる．

SE FiLM と FiLM の違いは，SE Block の有無である．ここでなぜ棒グラフや円グラフと異なり，折れ線グラフと点線グラフに大きな正解率の差が生じてしまったか考察する．SE Block はチャンネル間の依存関係を抽出し，元の特徴と要素積をとることで，チャンネル間の関係を反映させるというものであった．モデルに入力する画像は 3 チャンネルであるがこれは，RGB を表現している．よって画像の特徴は，CNN 層と SE ResBlock で処理されている時，$H \times W \times C$ の 3 次元の表現を持つが，チャンネル C は色に関する情報を持つ軸であると考えられる．図 5 に示した 5 種類のグラフの画像のサンプルをみると，棒グラフと円グラフでは，1 つの項目を表す表現の画像全体を占める割合が比較的大きい．一方，折れ線グラフと点線グラフは，1 つの項目を表す表現の画像全体を占める割合が比較的小さい．Global Average Pooling は $H \times W$ の情報を平均して 1 つの情報に圧縮する．画像の余白が占める割合が大きいということは，平均した際，余白の影響が大きくなることを意味する．これによって，項目を表す表現の持つ色の情報が消されてしまっている可能性が考えられる．

6.3　個別の結果の分析

3 つのモデルのうち，1 つのモデルの回答だけが間違っている質問が多いグラフの画像を図 7,8,9 に示した．それぞれの図に与えられた質問と解答を表 7,8,9 に示す．各表の C_F, C_S, C_R, C_* はそれぞれ FiLM，SE FiLM，RN の予測値と正解の値を示す．1 つのモデルの予測値だけが間違っている箇所を太字にした．各グラフの画像における特徴を分析することで，3 つのモデルが上手く認識できないケースを考察する．

RN だけが上手く判断できない例は，図 7 のように，差が大きくない複数の項目があるときが挙げられる．FiLM と SE FiLM では 4 つの ResBlock，SE ResBlock によって画像と質問文の結合を行なった上で畳み込みなどの演算を行なっているのに対し，RN では式 1 の g_θ のみで結合を行う．つまり FiLM と SE FiLM ではより強く画像の特徴に，質問文の特徴の影響が反映され，それによって質問に答えることができていると考えられる．

FiLM だけが上手く判断できない例は，図 8 の Chartreuse の項のように，ある項目の値が 0 や 0 に近い値があるときが挙げられる．RN ではオブジェクトの関係を抽出するので，質問文の比較対象の項目が 0 でなければ，その差異から判断できると考えられる．SE FiLM では質問文と画像の結合を行なった後に，SE Block でチャンネルの関係を抽出しているので，SE Block が他の画像によって，入力画像の 0 の値の項目の色の表現を保持していて，それを反映させることで認識できるものと

考えられる．

SE FiLM だけが上手く判断できない例は，図 9 のように，RGB の値が近い表現がある場合があげられる．macOS に搭載されている Digital Color Meter を利用すると，図の Dark Khaki の RGB は $(189, 182, 111)$，Dark Seafoam の RGB は $(144, 187, 144)$ でコサイン類似度は 0.981 と 1 に近い．SE Block ではチャンネル間の関係の一般性を得るために全結合層を用いて情報を圧縮している．圧縮の際，Dark Khaki と Dark Seafoam の差が消失してしまったと考えることができる．

実際には上記にあげた上手く判断できていない時の条件を満たしていても，該当するモデルで正しい答えが導けているケースは存在し，複数の要因が関係するものと考えられる．

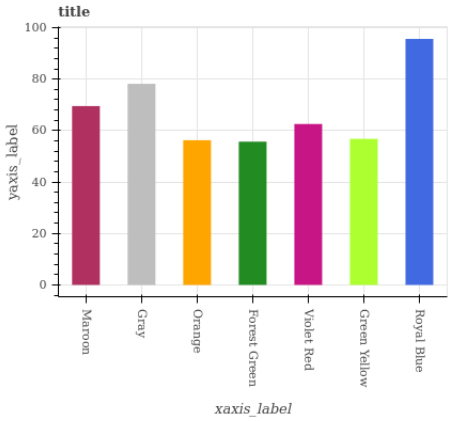


図 7　RN が上手く判断できない例 [3]

表 7　図 7 の質問と出力

Question	C_F	C_S	C_R	C_*
Is Forest Green the minimum?	0	1	0	1
Is Royal Blue the maximum?	1	1	1	1
Is Gray the minimum?	0	0	**1**	0
Is Gray the maximum?	0	0	0	0
Is Gray greater than Maroon?	1	1	**0**	1
Is Maroon less than Gray?	1	1	**0**	1
Is Maroon greater than Gray?	0	0	**1**	0
Is Gray less than Maroon?	0	0	**1**	0
Is Violet Red the high median?	1	1	**0**	1
Is Violet Red the low median?	1	1	**0**	1
Is Royal Blue the high median?	0	**1**	0	0
Is Royal Blue the low median?	0	0	0	0

7　結　　論

FigureQA において Relation Network，FiLM，SE FiLM の 3 つのモデルのうち，もっとも高い正解率を出したモデルは FiLM であった．また，結果の詳細をみると，片方のモデルしか正解しない質問が存在することから各モデルにそれぞれ利点があることがわかった．個別に画像，質問と結果の組をみると，各モデルがそれぞれ上手く認識できない要素があるといえる．

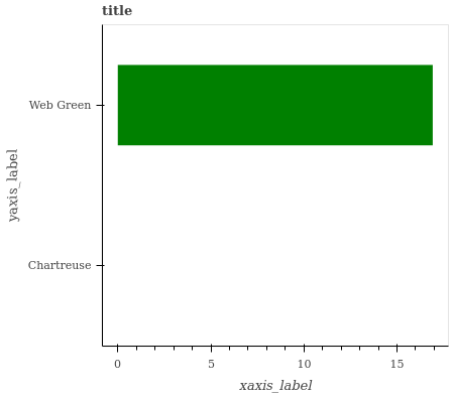


図 8　FiLM が上手く判断できない例 [3]

表 8　図 8 の質問と出力

Question	C_F	C_S	C_R	C_*
Is Chartreuse the minimum?	**0**	1	1	1
Is Web Green the maximum?	1	1	1	1
Is Web Green the minimum?	0	0	0	0
Is Web Green greater than Chartreuse?	**0**	1	1	1
Is Chartreuse less than Web Green?	**0**	1	1	1
Is Chartreuse greater than Web Green?	**1**	0	0	0
Is Web Green less than Chartreuse?	0	0	0	0
Is Web Green the high median?	1	1	1	1
Is Chartreuse the low median?	**0**	1	1	1
Is Chartreuse the high median?	**1**	0	0	0
Is Web Green the low median?	0	0	0	0

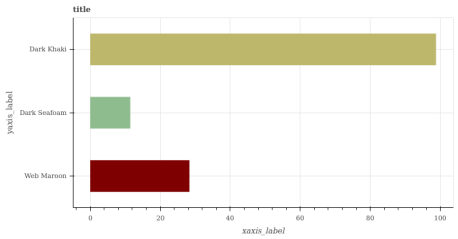


図 9　SE FiLM が上手く判断できない例 [3]

表 9　図 9 の質問と出力

Question	C_F	C_S	C_R	C_*
Is Dark Seafoam the minimum?	1	**0**	1	1
Is Dark Khaki the maximum?	1	1	1	1
Is Dark Khaki the minimum?	0	0	0	0
Is Dark Seafoam the maximum?	0	**1**	0	0
Is Dark Khaki greater than Dark Seafoam?	1	**0**	1	0
Is Dark Seafoam less than Dark Khaki?	1	**0**	1	0
Is Dark Seafoam greater than Dark Khaki?	0	**1**	0	0
Is Dark Khaki less than Dark Seafoam?	0	**1**	0	0
Is Web Maroon the high median?	1	1	1	1
Is Web Maroon the low median?	1	1	1	1
Is Dark Seafoam the high median?	0	**1**	0	0
Is Dark Seafoam the low median?	0	0	0	0

8　今後の展望

FigureQA は人工的にデータが作られたデータセットだった

ので，今後の課題としては，教科書や論文など，現実に存在するグラフの画像を対象とした十分大きなサイズのデータセットで，これらのモデルを適用してみることが挙げられる．実際に，現実のデータを扱った時，Relation Network，FiLM，SE FiLM その他のモデルが FigureQA に適用した場合と異なる振る舞いをするのかどうかを調べていきたい．また FigureQA は Yes/No 疑問文のみを扱ったデータセットであったが，質問文として $5W1H$ 疑問文を使用することも今後の課題として挙げられる．Yes/No 疑問文に質問が制限されている場合，分類器の出力は 2 クラスに限定されてしまい，分類器の出力を利用したアプローチを適用しようと考えても，可能な表現の幅が小さく効果がないと考えられる．例えば，2.5 節で述べた VQA を利用した画像キャプションの研究では，質問として $5W1H$ 疑問文のみを用いている．一般に，Yes/No 疑問文よりも，$5W1H$ 疑問文の方が難易度は高いが，$5W1H$ 疑問文を扱うことで適用できるアプローチも増えることもあり今後は $5W1H$ 疑問文にも取り組んでいきたい．

文　献

[1] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L. and Parikh, D.: VQA: Visual Question Answering, *Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433 (2015).

[2] Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y.: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, pp. 1724–1734 (2014).

[3] Cliche, M., Rosenberg, D. S., Madeka, D. and Yee, C.: Scatteract: Automated Extraction of Data from Scatter Plots, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18-22, 2017, Proceedings, Part I*, pp. 135–150 (2017).

[4] Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Comput.*, Vol. 9, No. 8, pp. 1735–1780 (1997).

[5] Hu, J., Shen, L. and Sun, G.: Squeeze-and-Excitation Networks, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, pp. 7132–7141 (2018).

[6] Huang, T.-H. K., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C. L., Parikh, D., Vanderwende, L., Galley, M. and Mitchell, M.: Visual Storytelling, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, Association for Computational Linguistics, pp. 1233–1239 (2016).

[7] Johnson, J., Hariharan, B., v. d. Maaten, L., Fei-Fei, L., Zitnick, C. L. and Girshick, R.: CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1988–1997 (2017).

[8] Johnson, J., Hariharan, B., van der Maaten, L., Hoffman, J., Fei-Fei, L., Zitnick, C. L. and Girshick, R.: Inferring and Executing Programs for Visual Reasoning, *Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2989–2998 (2017).

[9] Johnson, J., Karpathy, A. and Fei-Fei, L.: DenseCap: Fully Convolutional Localization Networks for Dense Captioning, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, pp. 4565–4574 (2016).

[10] Jung, D., Kim, W., Song, H., Hwang, J.-i., Lee, B., Kim, B. and Seo, J.: ChartSense: Interactive Data Extraction from Chart Images, *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, New York, NY, USA, ACM, pp. 6706–6717 (2017).

[11] Kafle, K., Cohen, S., Price, B. and Kanan, C.: DVQA: Understanding Data Visualizations via Question Answering, *Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–16 (2018).

[12] Kahou, S. E., Atkinson, A., Michalski, V., Ákos Kádár, Trischler, A. and Bengio, Y.: FigureQA: An Annotated Figure Dataset for Visual Reasoning, *Proceedings of Sixth International Conference on Learning Representations (ICLR)*, pp. 1–20 (2018).

[13] Kembhavi, A., Seo, M. J., Schwenk, D., Choi, J., Farhadi, A. and Hajishirzi, H.: Are You Smarter Than a Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension., *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9 (2017).

[14] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, USA, Curran Associates Inc., pp. 1097–1105 (2012).

[15] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollr, P. and Zitnick, C. L.: Microsoft COCO: Common Objects in Context, *Computer Vision – ECCV 2014* (Fleet, D., Pajdla, T., Schiele, B. and Tuytelaars, T., eds.), Cham, Springer International Publishing, pp. 740–755 (2014).

[16] Perez, E., Strub, F., De Vries, H., Dumoulin, V. and Courville, A.: FiLM: Visual Reasoning with a General Conditioning Layer, *Proceedings of The Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, United States, pp. 1–13 (2018).

[17] Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: Learning Internal Representations by Error Propagation, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1* (Rumelhart, D. E., McClelland, J. L. and PDP Research Group, C., eds.), MIT Press, Cambridge, MA, USA, pp. 318–362 (1986).

[18] Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P. and Lillicrap, T.: A Simple Neural Network Module for Relational Reasoning, *Advances in Neural Information Processing Systems 30* (Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R., eds.), Curran Associates, Inc., pp. 4967–4976 (2017).

[19] Shin, A., Ushiku, Y. and Harada, T.: Customized Image Narrative Generation via Interactive Visual Question Generation and Answering, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9 (2018).

[20] Siegel, N., Horvitz, Z., Levin, R., Divvala, S. and Farhadi, A.: FigureSeer: Parsing Result-Figures in Research Papers, *Proceedings of 14th European Conference on Computer Vision (ECCV)*, pp. 1–16 (2016).

[21] Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D. and Parikh, D.: Yin and Yang: Balancing and Answering Binary Visual Questions, *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5014–5022 (2016).