私のブックマーク

視覚と自然言語の融合研究†1

牛久 祥孝（東京大学）

1. はじめに

視覚（もう少し具体的にいえば画像や動画など）を対象とした研究と，自然言語を対象とした研究は，それぞれComputer VisionおよびNatural Language Processingの領域において，お互い少しずつ影響し合いながら発展を遂げてきた．特に近年，深層学習の一種であるConvolutional Neural Network（CNN）やRecurrent Neural Network（RNN）といった共通の機械学習手法が台頭し，それぞれの領域への参入障壁が低くなった．結果として，視覚と自然言語を融合する研究がさまざまな広がりを見せつつある（深層学習によって，初めてこれらの研究が可能になった，というわけではないことに注意されたい）．このブックマークでは，画像と自然言語の融合という括りの中で，具体的に以下の研究分野を紹介する．

- 画像からのキャプション生成
- 動画からのキャプション生成
- 画像も含めた質問応答システム
- キャプションからの画像生成
- 画像を伴う言語横断検索・キャプション翻訳

まずはそれぞれの分野について，「主要な論文」，「概要がわかるチュートリアル・サーベイ論文・コンペティション」をあげる．サーベイ論文などが存在しない場合であっても，コンペティションの結果発表に用いられた資料が手法の概要や精度の理解に役立つ場合がある．合わせて入力→出力の概要図も掲載するが，これはあくまで本ブックマークのために作成したイメージ図である．

次に，こうした分野で用いられるデータセットやツールについて紹介する．また最後に，関連会議と国内の主要研究グループについても触れる．論文を紹介する際に会議名のみをあげているものは，この関連会議のセクションで扱っているものである．国内の主要研究グループについては，筆者の見聞し得る中から，独断と偏見が入ったセレクションを行っている可能性があることをご容赦願いたい．

2. 画像からのキャプション生成

入力した画像の内容を自然な文として記述する問題である（図1）．画像の主要な内容を，平均10単語程度の英語キャプションとして生成するのがオーソドックスな問題設定となっている（イメージ画像は日本語キャプション）．

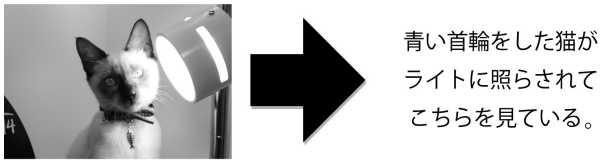

図1

2·1 主要な論文

- Every Picture Tells a Story：Generating Sentences from Images [1]

  Ali Farhadiらによる2010年の論文．入力画像のみからキャプションを生成する手法の研究としては世界初のもの．便宜上生成といっているが，基本的にはデータベース内のキャプションで一番最適なものを丸ごと再利用するアプローチをとっている．

†1 http://www.ai-gakkai.or.jp/my-bookmark_vol32-no1

・Show and Tell：A Neural Image Caption Generator [2]

Google の Oriol Vinyals らによる，CNN + RNN を用いたキャプション生成手法．実際には，この研究より前から深層学習によるキャプション生成が提案されている．発表されたのは 2015 年の CVPR であるが，この年の CVPR では CNN と RNN を組み合わせたキャプション生成が複数の研究機関から同時に提案されている．

2·2 チュートリアル・サーベイ論文・コンペティション

・Automatic description generation from images：A survey of models，datasets，and evaluation measures [3]

サーベイ論文．いろいろと漏れと誤りもあるが，キャプション生成のサーベイ論文としては一番網羅性が高い．

・画像キャプションの自動生成 [4]

このブックマークの著者自身の資料で恐縮ではあるが，国内会議である「画像の理解・認識シンポジウム（MIRU）」のチュートリアルでも用いた資料なので，チュートリアル資料ということでご容赦願いたい．画像キャプション生成の歴史的な経緯から，キャプション生成の評価方法，このブックマークで紹介するような視覚と自然言語の融合研究までを広く紹介している．

・COCO Captioning Challenge [5]

これはチュートリアルやサーベイ論文ではないが，2015 年の CVPR でワークショップとして開かれた国際コンペティションである．Microsoft が，自身の公開している MS COCO という画像キャプションデータセットによる画像キャプション生成精度を競わせたものである．ここで優勝したのが上にある Show and Tell を提案した Google のチームであり，次点が Microsoft Research のチームであった．なお，人間が人手で付与したキャプションは Google のチームの 2 倍の精度をたたき出した．画像の主要な物体を識別する問題については，100 万枚の画像を学習して 1 000 種類の物体を識別・識別する ImageNet Large Scale Visual Recognition Challenge（ILSVRC）が主なコンペティションとなっているが，ここでは人間自身のエラー率（5%程度）より低いエラー率でしのぎを削っており，対して画像キャプション生成には精度的にまだまだ改善の余地があることを示唆している．優勝した Google のチームによるキャプション生成結果 [6] も公開されているが，しばしばリンクが切れているので注意されたい．

3. 動画からのキャプション生成

入力した動画の内容を自然な文として記述する問題である．画像キャプション生成が静止画 1 枚に対してキャプションを生成するのに対して，動画キャプション生成では（データセットにもよるが）10 秒前後の clip と呼ばれる短い動画単位に対してキャプションを生成する．

図 2

3·1　主要な論文

・Grounded Language Learning from Video Described with Sentences [7]

Haonan Yu と Jeffrey Mark Siskind による論文で，動画とその動画を説明する文のみから，言語と動画内の物体とのグラウンディングを学習し，新規動画にもキャプションを生成できる．対象物体が 4 種類と少なく，かなりコントロールされた小規模データセットでの実験ではあるが，2013 年の ACL でベストペーパに選ばれている黎明期の論文である．

・Long-Term Recurrent Convolutional Networks for Visual Recognition and Description [8]

Jeff Donahue らによる論文で，先ほど画像キャプション生成として紹介した Google の論文と同じく 2015 年の CVPR で発表された．画像のキャプション生成にも動画のキャプション生成にも取り組んだ論文で，CNN + RNN を用いている．RNN によるニューラル統計翻訳（NMT：Neural Machine Translation）はつい最近 Google の日英翻訳にも採用されて話題となったが，基本的には日本語（単語の系列データ）を英語（単語の系列データ）に変換するものである．そのため，CNN によって動画フレームからの特徴量抽出を，RNN によって動画（フレーム画像の系列データ）を言語（単語の系列データ）に翻訳するのも自然な流れである．

3·2 チュートリアル・サーベイ論文・コンペティション

・ Microsoft Multimedia Challenge - Video to Language Challenge [9]

残念ながら，筆者の知る限りにおいて体系だった動画キャプション生成の資料は日本語でも英語でも存在していない．ここでは，国際コンペティションとして開かれた Microsoft Multimedia Challenge について紹介する．ACM Multimedia では，本会議中に Multimedia Grand Challenge というコンペティションが開かれ，世界中の企業からの提案課題に世界中の研究機関がトライし，その結果をプレゼンして賞を競う．北京の Microsoft Research Asia が 2016 年にオーガナイズしたのが本チャレンジで，後述する MSR-VTT という大規模動画キャプションデータセットによる動画キャプション生成精度を競った．やや専門的ではあるが，Leaderboard を参照すれば今の最先端の手法の精度を知ることもできる．

4. 画像を含めた質問応答システム

質問応答システム（QA：Question Answering）として今一番話題なのは IBM の Watson であろう．言語のみで入力された問いに答えるのが QA システムなのに対し，画像を含めた質問応答システム（VQA：Visual Question Answering）は，お題となる画像が存在していて，その画像に関連した問題に答える．

図 3

4·1 主要な論文

・ A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input [10]

Mateusz Malinowski と Mario Fritz による論文で，画像を含めた QA システムを初めて提案している．2014 年の NIPS で発表されたものであり，いかにこの問題が新しいものかがわかる．

・ VQA：Visual Question Answering [11]

Stanislaw Antol らによる ICCV 2015 で発表された研究で，ここで収集・公開されたデータセットが現在最も多く用いられるベンチマークとなっている．この流れに合わせて，問題名も Visual Question Answering という名前に定着してきた．

4·2 チュートリアル・サーベイ論文・コンペティション

・ Visual Question Answering：A Survey of Methods and Datasets [12]

・ Visual Question Answering：Datasets，Algorithms，and Future Challenges [13]

執筆時ではまだ arXiv にしか上がっていないが，VQA に関してはサーベイ論文が 2 本出ている．どちらも実際に VQA を研究しているグループによるもので，特に後者を手掛けた Anton van den Hengel の研究室は，画像キャプションの自動生成に関しても精力的に研究を進めている．

・ VQA Challenge [14]

CVPR 2015 のワークショップとして開催された国際コンペティションで，実画像部門とアニメ画像部門それぞれの VQA 正答率を競った．

5. キャプションからの画像生成

10 単語前後のキャプションを入力して，それにふさわしい画像を出力する．

広義でのキャプション生成が「既存キャプションの再利用」と「新規キャプションの生成」のどちらをも含んでいることのアナロジーを考えると，画像生成でも「既存画像の再利用」と「新規画像の生成」両方があり得る．入力画像を既存キャプションで説明する手法の多くは，そのまま入力キャプションにふさわしい既存画像を検索する手法としても使える．実際，既存キャプションによる画像キャプション生成のなかでも初期の論文 [15] から，入力文にふさわしい画像が検索できたかどうかを評価している（リンク先文献表 3 参照）．

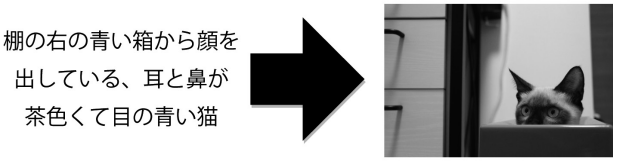

図 4

5·1　主要な論文

ここでは特に新規画像の生成について紹介する．あらかじめ，新規のキャプション・画像を生成する難しさについて粗く比較してみたい．キャプションであれば，10 万の語彙で 10 単語の文を書くとして，活用などを全く考えなければ 10^{50} 通りの文が考えられる．画像であれば，Web 上での視聴に耐え得る程度の解像度として 320×240 ピクセルのカラー画像を生成するとして，RGB の各画素が 256 段階とすれば大体 $10^{555\,000}$ の画像が考えられる．単語の選択を一つ誤るのと画素値の値を一つ間違えるのが等価とは考えにくいが，画像のほうが探索空間がより巨大であることについては賛同を得られるのではないだろうか．

・ Text to 3D Scene Generation with Rich Lexical Grounding [16]

Angel Chang らによる IJCNLP 2015 の論文で，三次元の物体モデルを組み合わせて文から画像を生成している．モデルを用いる利点は，上述のような巨大な探索空間を小さくし得る点にある．キャプションからの画像生成にモデルを用いる研究は，このブックマークの筆者が所属する研究室でもいくつか試みている [17]．

・ Generating Images from Captions with Attention [18]

Elman Mansimov らによる ICLR 2016 の論文で，世界で初めて入力キャプションに対応する画像をピクセルレベルで新規生成しようとした研究である．RNN によってキャプション解析と画像生成モジュールが別々にデザインされており，ニューラル機械翻訳のように一度入力文を特徴量ベクトルにエンコードしてからデコードする流れになっている．文生成のデコード時は単語が一つずつ増えていくのに対して，この論文ではデコードを進めるにつれて最初はぼやけた画像が徐々にシャープになるように画像が上塗りされていくような過程をたどっている．

・ Generative Adversarial Text to Image Synthesis [19]

Scott Reed らによる ICML 2016 の論文で，敵対的生成ネットワーク（GAN：Generative Adversarial Network）の深層畳込みバージョン（DCGAN）を用いてキャプションからの画像生成を試みている．DCGAN では，乱数から画像を生成する Generator と，既存の実画像とシステムが生成した画像とを見分けようとする Discriminator がお互いに勝とうとする学習（敵対的学習）を行う．Discriminator をうまく騙せるようになった Generator は，解像度の高い画像を生成できるようになっている．この研究では Generator に入力する乱数と，Discriminator が真贋を見分ける特徴量に入力キャプションの特徴量が統合されており，生成する画像をキャプションによってコントロールしようとしている．

5·2　チュートリアル・サーベイ論文・コンペティション

既存画像の再利用という意味であれば，画像キャプション生成のサーベイなどが役に立つ．しかしながら，入力文から新規画像を生成する研究は極めて少ない．パーツレベルでの新規画像生成については，アニメ調の画像と割り切ったパーツで生成するならともかく，実画像の切り貼りで実画像に見えるような画像を再構成するのは別の難しさをはらんでいる．キャプションから実画像をピクセルレベルで生成するものは上記の論文のみであり，うまく生成できているように見えるのは「花もしくは鳥のみの画像を 1 万枚ほど集めて，その説明文をキャプションとして添えたデータセット」のみで学習させた場合などである．一般的な画像を生成させるにはより一層のブレークスルーが必要な状況にあり，このブックマークで紹介する各研究分野の中でもかなり難しい問題といえる．

6. 画像を伴う言語横断検索・キャプション翻訳

IBM モデルやフレーズに基づく統計的機械翻訳，そして近年のニューラル機械翻訳と，翻訳技術はいくつもの深化を経て発展してきた．その中で一部の研究者から提案されているのが，画像を伴う言語横断検索やキャプション翻訳である．通常の翻訳では自然言語のみを入力するのに対して，これらの分野では画像と文のペアを入力として翻訳を進める．

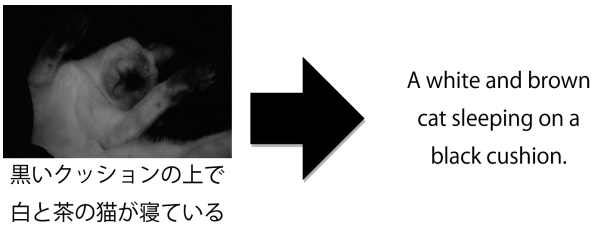

図 5

画像を伴って翻訳をする意義の簡単な例は，Images as Context in Statistical Machine Translation [20] に見ることができる．これはワークショップにて提案されたコンセプトであり，手法の実装が伴っているものではないが，ある英葡翻訳の例をあげている．英語で“seal”と書かれている場合，貼り付ける「シール」以外にも「アシカやアザ

ラシ」を指している可能性がある．文だけでなく画像も入力されれば，このような曖昧さが解消されると期待される．

6·1 主要な論文

・ Image-Mediated Learning for Zero-Shot Cross-Lingual Document Retrieval [21]

舟木類佳と中山英樹による EMNLP 2015 の論文である．英語キャプションと画像のデータセット，日本語キャプションと画像のデータセットが独立して存在するとし，日英対訳コーパスと組み合わせて 3 変量版の一般化正準相関分析を適用している．すると入力文から画像特徴量に対応するベクトルを推定し，対となっている語のキャプションがついている画像をこのベクトルから検索できる．

・ Multimodal Pivots for Image Caption Translation [22]

Julian Hitschler らによる ACL 2016 の論文で，より多くの英独キャプションつき画像データセットを構築し，画像を介した英独翻訳を提案している．

6·2 チュートリアル・サーベイ論文・コンペティション

・ Shared Task：Multimodal Machine Translation [23]

残念ながら，画像を伴う言語横断型検索や翻訳についてのチュートリアルやサーベイ論文についてもまだ存在していない．そこで，例によって，最近開催されたコンペティションを紹介する．自然言語処理の国際学会で毎年開催されてきたワークショップに Workshop on Statistical Machine Translation（WMT）というものがあるが，2016 年から一つの国際会議として開催されるようになった．この中の Multimodal Machine Translation において，Multimodal Machine Translation と題してこのタスクが課され，世界中の参加者が翻訳の精度を競った．

7. データセットやツール

7·1 データセット

まずデータセットについては，On Available Datasets for Empirical Methods in Vision & Language [24] というページが役に立つ．このページでは，画像キャプションや動画キャプションのデータセット，VQA のデータセットなどが網羅的に紹介されている．このページは EMNLP 2015 で発表されたデータセットのサーベイ論文に基づいたものである．そこでこのセクションではまず，同ページでも扱われているが，非常に有名なデータセットを二つ紹介する．次に，この論文の発表後に公開されたデータセットの中から興味深いもの二つを紹介する．

・ Microsoft COCO [25]

画像からの物体認識，領域分割，そしてキャプション生成のために作成されたデータセット．計 30 万超の画像からなるが，画像キャプションデータセットとしては半分の 16 万枚ほどの画像と，それぞれに 5 文ずつついたキャプションを用いる．もう半分は物体認識やセグメンテーションのためのデータセットであり，ECCV 2016 のワークショップとしてコンペティションが開かれている．

・ VQA Dataset [25]

バージニア大学の Devi Parikh の研究室によるもので，その名のとおり VQA 用に収集されたデータセット．MS COCO の画像 20 万枚超とアニメ調の画像 5 万枚について，それぞれ 60 万と 15 万の質問文，さらにそれぞれ 800 万弱と 200 万弱の回答が付与されている．

・ Visual Genome [26]

スタンフォード大学の Li Fei-Fei の研究室によるもので，画像枚数こそ 10 万枚程度だが，そこに付与されたラベルの質と量が圧倒的である．画像内の矩形領域それぞれに総計 500 万超の短いキャプションが付与されていたり，170 万の VQA が付与されていたりするほか，画像 1 枚 1 枚に多くの情報が付与されている．

・ MSR-VTT [27]

北京にある Microsoft Research Asia による動画キャプションデータセットで，これまでのデータセットの中では規模が最大である．2013 年の ACL ベストペーパをとった動画キャプション生成（上記）では，3 〜 5 秒の動画が 61 本だった．それに対して MSR-VTT は総じて 41.2 時間の長さの動画 1 万本に対してキャプションが

ついており，3 年で規模が急激に増えたことがわかる．

紹介したデータセットは，すべて人手でラベル（キャプションや QA）を付与したものである．それ以外の収集方法として，Google 検索や Flickr など特定の Web サービスから収集するデータセットも複数存在する．人手でラベルを付与すると，誤りが少ないために学習しやすい．かつこのラベルとシステム出力との差をうまく比較できれば定量評価も自動で計算できるのがメリットである．一方で Web からの自動収集はラベルの信頼性に問題がある反面，容易に大規模なデータセットを収集できるというメリットがある．Web から収集したデータセットについては上記データセットサーベイ論文を参照されたい．

7·2　ツール

視覚と言語の統合研究と題して紹介しているものは，いずれも長くてまだ 5 年ほどの歴史しかもっておらず，デファクトスタンダードとなるようなツールは存在していない．スクリプト言語として Python や MATLAB が多かったり，深層学習を使う場合が多いので TensorFlow や Theano，MatConvNet，Chainer といったフレームワークを使うというような基本的な話はあるが，ここではもっとお試しに近いものをツールとして三つ紹介する．

- CaptionBot [28]

  Microsoft が提供する Cognitive Service と呼ばれる API 群を利用し，Microsoft Research による画像キャプション生成 [29] を Web サービスとして実装したもの．ブラウザでこの Web ページを開き，手元にある画像をアップロードするだけでよい．画像キャプション生成が大体どのようなレベルにあるのかを知るためには，一番手っ取り早い手段だといえる．

- Visual QA Demo [30]

  上述の VQA Challenge 2016 において実画像部門で一番精度が高かった手法を用いた公式デモページ．この手法はのちに EMNLP 2016 でも発表されている [31] が，筆頭著者はソニーの Akira Fukui で，UC バークレーの Trevor Darrell のもとで研究されたものである．手元の画像をアップロードし，英語で質問を入力すれば，回答の候補を提示してくれる．より深く知りたい場合は，GitHub にソースコード [32] も公開されている．

- NeuralTalk2 [33]

  スタンフォード大学の Andrej Karpathy による画像キャプション生成のソースコード．画像キャプション生成のコードを GitHub で探すといろいろと見つかるのだが，データの前処理を自前で実装するなどの手間がかからないものは意外と限られる．このソースコードでは Karpathy 自身のキャプション生成や Google の Show and Tell を模したキャプション生成を試せる．先日，ようやく Google 自身から Shown and Tell の公式実装 [34] が公開され，Karpathy 自身も Google の公式実装のほうが精度が高いと報告しているが，ユーザからの情報の多さもあいまって，NeuralTalk2 のほうがお試しには適しているといえよう．

8. 関 連 会 議

ここでは，上述の視覚と言語の融合研究が発表される主要な国際会議について，それぞれの分野ごとに紹介する．複数の論文が観測される主要会議に絞った結果として，IJCAI や INLG など，各分野である程度名の知れた会議でも省略されているものがある．

- Artificial Intelligence，Machine Learning

  AAAI：Conference on Artificial Intelligence
  ICLR：International Conference on Learning Representations
  ICML：International Conference on Machine Learning
  NIPS：Neural Information Processing Systems

- Natural Language Processing

  ACL：Annual Meeting of the Association for Computational Linguistics
  COLING：International Conference on Computational Linguistics
  CoNLL：Conference on Natural Language Learning

IJCNLP：International Joint Conference on Natural Language Processing
EMNLP：Conference on Empirical Methods on Natural Language Processing
NAACL：Annual Conference of the North American Chapter of the Association for Computational Linguistics

・Computer Vision
BMVC：British Machine Vision Conference
CVPR：Conference on Computer Vision and Pattern Recognition
ECCV：European Conference on Computer Vision
ICCV：International Conference on Computer Vision
ICIP：International Conference on Image Processing

・Multimedia
ACM MM：ACM Multimedia
ICME：International Conference on Multimedia and Expo
ICMR：International Conference in Multimedia Retrieval

9. 研究室・研究者・研究プロジェクト

・Perception and Language Understanding Group（PLU）[35]（産業技術総合研究所）
非言語な知覚と言語との融合を産学官で試みるグループである．国立情報学研究所の宮尾祐介准教授のもと，各分野の研究者たちが集って「画像・動画理解」，「データ言語化」，「意味論的構文解析」に取り組んでいる．

・岡谷研究室 [36]（東北大学）
コンピュータビジョンの研究室であり，主宰の岡谷貴之教授は深層学習分野でも日本を代表する研究者の一人である．服飾画像認識をはじめとした視覚属性認識に従事する山口光太助教と共同で「画像と言語を用いた質感情報表現のディープラーニング」[37] と題したプロジェクトを進めている．

・原田・牛久研究室 [38]（東京大学）
原田達也教授が主宰し，このブックマークの筆者も2016年から講師をつとめている．手前味噌で恐縮であるが，2011年に開始した画像キャプション生成をはじめとして，画像を含めた質問応答や動画キャプション生成，入力文からの画像生成など，広範囲にわたる視覚・言語統合の研究を進めている．

なお，PLUに所属する研究者の研究室に関しては，PLUの紹介をもって割愛させていただいている．

10. お わ り に

このブックマークでは，視覚と言語の融合研究について広く紹介した．冒頭でも述べたとおり，画像と自然言語それぞれについての先端的な技術がコモディティ化し，お互いに参入障壁が下がっている状況にある．いずれも研究課題としてはまだまだやりつくされたとは言い難く，画像キャプション生成ですら人間が作成するキャプションには遠く及ばない（ImageNet上での画像の識別精度や，囲碁のような完全情報ゲーム，クイズのようなオントロジー・質問応答システムでは，それぞれ人間を超えるような性能を発揮しつつあるのは周知のとおりである）．

海外での広がりに比べると，日本でこれらの分野に取り組んだ例や企業・研究室はまだまだ少ない．同時に，こういった技術を産業へ持ち込んだ例というのは寡聞にして存じ上げない．産学それぞれにおいて，新規参入を心よりお待ち申し上げるしだいである．

[1] http://link.springer.com/chapter/10.1007/978-3-642-15561-1_2
[2] http://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Vinyals_Show_and_Tell_2015_CVPR_paper.html
[3] https://www.jair.org/media/4900/live-4900-9139-jair.pdf
[4] http://www.slideshare.net/YoshitakaUshiku/ss-57148161
[5] http://mscoco.org/dataset/#captions-challenge2015
[6] http://nic.droppages.com/

[7] https://aclweb.org/anthology/P/P13/P13-1006.pdf
[8] http://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Donahue_Long-Term_Recurrent_Convolutional_2015_CVPR_paper.html
[9] http://ms-multimedia-challenge.com/challenge
[10] https://papers.nips.cc/paper/5411-a-multi-world-approach-to-question-answering-about-real-world-scenes-based-on-uncertain-input
[11] http://www.cv-foundation.org/openaccess/content_iccv_2015/html/Antol_VQA_Visual_Question_ICCV_2015_paper.html
[12] https://arxiv.org/abs/1607.05910
[13] https://arxiv.org/abs/1610.01465
[14] http://visualqa.org/
[15] http://www.jair.org/media/3994/live-3994-7274-jair.pdf
[16] http://aclweb.org/anthology/D15-1070
[17] http://dl.acm.org/citation.cfm?doid=2964284.2967193
[18] https://arxiv.org/abs/1511.02793
[19] http://jmlr.org/proceedings/papers/v48/reed16.html
[20] http://staffwww.dcs.shef.ac.uk/people/L.Specia/projects/vlnet.html
[21] http://aclweb.org/anthology/D15-1070
[22] https://www.aclweb.org/anthology/P/P16/P16-1227.pdf
[23] http://www.statmt.org/wmt16/multimodal-task.html
[24] http://visionandlanguage.net/
[25] http://mscoco.org/
[26] https://visualgenome.org/
[27] http://ms-multimedia-challenge.com/dataset
[28] https://www.captionbot.ai/
[29] https://www.microsoft.com/en-us/research/publication/rich-image-captioning-in-the-wild/
[30] http://vqa.daylen.com/
[31] https://www.aclweb.org/anthology/D/D16/D16-1044.pdf
[32] https://github.com/akirafukui/vqa-mcb
[33] https://github.com/karpathy/neuraltalk2
[34] https://github.com/tensorflow/models/tree/master/im2txt/im2txt
[35] https://aistairc.github.io/plu/
[36] http://www.vision.is.tohoku.ac.jp/jp/home/
[37] http://shitsukan.jp/ISST/researches/
[38] http://www.mi.t.u-tokyo.ac.jp/