抽象度を制御可能な エンティティレベルの画像キャプション生成

加藤 駿弥　Chenhui Chu　黒橋 禎夫
京都大学
{s-kato, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp

概要

画像キャプション生成は近年盛んに研究されてきたが, 先行研究は正確性, 多様性, 差別化を重視し, 抽象度は軽視してきた. 本研究では, Inverse Document Frequency (IDF) とエンティティレベルの含意関係認識を用いて, 抽象度を制御可能なエンティティレベルの画像キャプション生成手法を提案する. 含意関係認識を用いて抽象度を評価したところ, 先行研究と比べ, 提案手法は抽象度をより制御可能になり, 具体的なキャプションの生成において 15.7%の精度向上が確認された.

1　はじめに

画像キャプション生成は与えられた画像を正しく説明する分を生成するタスクであり, 物体, 物体間の関係などを正しく理解する必要がある. また, 画像全体のキャプションを生成する研究 [1] だけでなく, Dense Captioning のようにエンティティに対してキャプションを生成する研究 [2, 3, 4] も行われている.

従来のキャプション生成に関する研究は, より正確もしくは多様なキャプションを生成することに焦点を当てており, 抽象度の制御を軽視してきた. しかし, 人間は視覚情報を言語化するとき, 物体認識を行うだけではなく, 常識などをもとに認識した物体や物体間の関係を具体的な概念と結びつけて形容する. 例えば, 図 1 のようにスーツを着てネクタイを締め, 革靴を履いている男性は, 常識をもとに 「男性」だけではなく「ビジネスマン」という概念と結びつけられる.

本研究では, エンティティレベルにおける物体認識のようなより抽象的なキャプションと常識推論がいるより具体的なキャプションの生成を制御することを目指す. そのために, IDF と含意関係認識を用いた抽象度の制御するためのデータセットの構築手法を提案する. 提案手法の概要を図 2 に示す. 具体的には, まず, エンティティには複数のキャプションが付与されており, これを 1 つの文書とみなしてキャプション生成に含まれる uni-gram の IDF を計算する. 次に, あるエンティティにつけられた複数のキャプションの中で, IDF と その他のキャプションに対して含意関係である割合の和が高いものを Fine(具体的), そうでないものを Coarse(抽象的) であるとラベル付けすることで, 各エンティティにおける Coarse/Fine なキャプションを生成するためのデータセットを構築する.

図 1　抽象的, 具体的なキャプションの例

実験の結果, 先行研究と比べてより抽象度を制御可能になったことを示した. そして, 先行研究では単一エンティティのみで学習していたが, 本研究では複数エンティティも含めて学習させ, 単一エンティティと同様に抽象度を制御可能になったことを示した.

2　前提知識

エンティティレベルの画像キャプション生成には, Li ら [5] の研究がある. まず, Li ら [5] はエンティティレベルの多様なキャプション生成というタスクを提案した. このタスクはエンティティのキャプションを生成するタスクで, 複数の正解キャプションがある点で Dense Captioning とは違う.

また, ターゲットエンティティと同じ画像に含まれる他のエンティティ (neighbors) の情報を考慮す

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

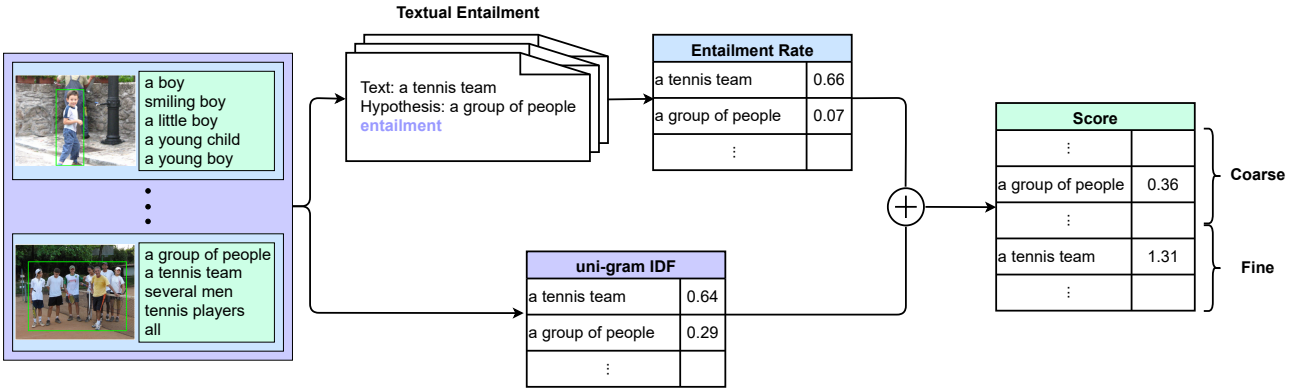


図 2　提案手法の概要

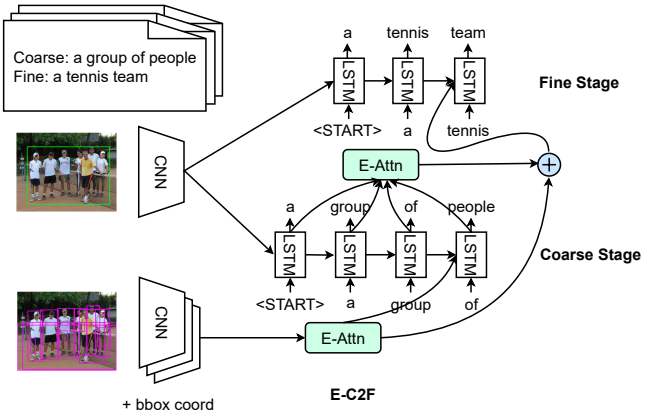


図 3　E-C2F のアーキテクチャ

るための E-Attn という新たな Attention 機構を提案し, それを組み込んだ E-C2F というモデルを提案した. E-C2F のアーキテクチャを図 3 に示す. E-C2F は 2 階層構造のモデルであり, Coarse と Fine という 2 つの抽象度の違うキャプションを生成できる. この 2 階層構造により, データセットのキャプションを Coarse と Fine に分割する必要があるが, Li ら [5] はキャプションごとの bi-gram の IDF の和を用いて分割した. E-C2F は提案されたエンティティレベルの多様なキャプション生成タスクにおいて SOTA を達成した. 注意点として Li ら [5] は複数エンティティはデータセットから除いた.

3　提案手法

3.1　IDF と含意関係認識を用いた分割

先行研究には以下の 3 つの問題点がある.

- Li ら [5] は, 同一エンティティに付与されているキャプションからそれらのペアを全通り作り, ペアの中で bi-gram の IDF の和が小さいキャプションを Coarse, 大きいキャプションを Fine とした. しかしこの手法は, あるペアでは Coarse となったキャプションが, 別のペアでは Fine となってしまうことがあり, データセットに一貫性がない.
- IDF の和で計算しているため, キャプションの語数が長いほどスコアが高くなる.
- bi-gram の IDF を計算しているため, キャプションが 1 単語だとスコアが 0 となる.

これらを改善するため, uni-gram の IDF と含意関係認識を用いた Coarse, Fine の分割手法を提案する.

3.1.1　IDF

先行研究では, エンティティに付与されているキャプションを 1 つの文書とみなし, bi-gram の和を用いて Coarse, Fine を分割したが, 本研究では uni-gram の IDF を用いる.

3.1.2　含意関係認識

キャプションの抽象度の高低を予測するため, 含意関係認識を用いる. 既存の含意関係認識モデルは文で事前学習されているため, エンティティのデータセットでの fine-tuning が必要である. 本研究では, エンティティの含意関係認識データセットとして Chu ら [6] のデータセットを用いる. このデータセットは 2 つのキャプション (pharse1, phrase2) を alternation, forward entailment, reverse entailment, equivalence, independence の 5 クラスに分類したが, NLI は neutral, entailment, contradiction の 3 クラスであるのでそのまま fine-tuning できない. ここでは, 表 1 のようにマッピングを行った. 表中の Forward はデータセットの phrase1 を前提, phrase2 を仮説にしたもので, Reverse は逆である.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1　Chu ら [6] のデータセットのクラス分類

VGP Class	Forward	Reverse
alternation	contradiction	contradiction
forward entailment	entailment	neutral
reverse entailment	neutral	entailment
equivalence	neutral	neutral
independence	neutral	neutral

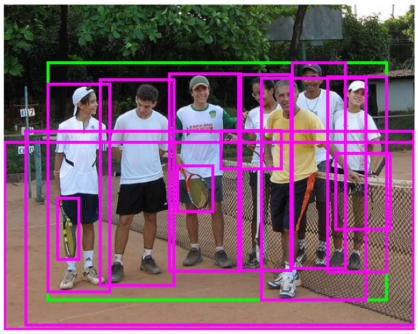


図 4　複数エンティティの定義 (緑: ターゲットエンティティ, ピンク: neighbors)

3.1.3　Coarse, Fine の分割

まず Chu ら [6] がアノテーションしたデータにラベルを付与する. アノテーションしてないデータについては, fine-tuning した RoBERTa でラベルを付与する. ここで, Chu らの手法と同様に stop words は除外する. 次に, 含意関係認識において, キャプションが前提のとき, 含意関係となる割合を計算する. さらに, uni-gram の IDF を計算し, キャプション中の単語の IDF の最大値を IDF のスコアとする. ここで含意関係認識と IDF の寄与が同一になるように IDF のスコアを正規化する. 最後に含意関係の割合と IDF のスコアの和を計算する.

3.2　複数エンティティのキャプション生成

次に複数エンティティに焦点を当てる. Li ら [5] は複数エンティティはデータセットから除いている. 本研究では, 図 4 のように, ターゲットエンティティを複数の bounding box を結合したものと定義し, neighbors はその他のエンティティにターゲットの複数エンティティを加えたものと定義する.

4　実験

4.1　実験設定

本研究では, 画像からのエンティティの特徴ベクトル抽出は ResNet-152[7] を用いた. データセットは Flickr30k Entities[8] を用いた. Flickr30k Entities に

表 2　データセットの統計と fine-tuning の結果

split	entailment	neutral	contradiction	size	acc (%)
Train	10,796	36,994	2,194	49,984	96.6
Validation	2,722	8,146	600	11,468	92.4
Test	3,051	8,261	810	12,122	91.6

は約 3 万枚の画像があり, それぞれの画像に 5 つのキャプションが付与されている. そしてキャプション中のフレーズと画像中のエンティティが紐づけられている. エポック数は 40 で統一し, 他の設定は Li ら [5] に従った. 本研究では正確性, 多様性, 抽象度の評価をした. 正確性の評価指標として, BLEU, METEOR, ROUGE_L, CIDEr を用いた. 多様性の評価指標として, LSA, self-CIDEr, n-gram, DIV-n を用いた. 抽象度の評価指標として, 含意関係の割合を用いた. モデルには 4.2 で fine-tuning した RoBERTa を用いた.

4.2　含意関係認識のための fine-tuning

含意関係認識モデルには SNLI[9], MNLI[10], FEVER-NLI[11], ANLI (R1, R2, R3)[12] で事前学習済み RoBERTa-Large[1)] を用いた. データセットの統計と fine-tuning した結果が表 2 である. fine-tuning 前の RoBERTa の精度は 67%であったが, fine-tuning により, テストデータにおいて 91%の精度を達成した.

4.3　単一エンティティのキャプション生成

Coarse, Fine の数をおおよそ等しくするため, 本研究では提案手法の閾値を 0.55 とした. まず正確性, 多様性の評価を行った. 単一エンティティのキャプション生成の正確性と多様性の実験結果を表 3 に示す. データセットが異なるので直接的な比較はできない. 提案手法により, 正確性は少し低下したが, 多様性は非常に向上した. 提案手法の Fine の正確性が低い原因として, より具体的なキャプションを生成するタスクの難易度が高いことが考えられる.

次に抽象度の評価を行った. E-C2F の生成した Fine のキャプションと, 提案手法の生成した Fine のキャプションに対して含意関係認識を適用し, それぞれのキャプションを前提, 仮定としたときに含意関係である割合を比較した. E-C2F から生成されたキャプションが前提で提案手法から生成されたキャプションが仮定のとき, 含意関係となるものは 3.5%であったのに対し, 逆は 19.2%であった. よって提案手法により, より具体的なキャプションを生成でき

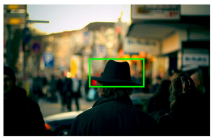
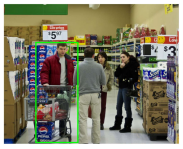
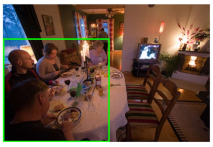
1)　https://huggingface.co/ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 3　正確性と多様性の結果

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
E-C2F - Coarse	0.47	0.31	0.22	0.13	0.19	0.47	0.38
E-C2F - Fine	0.41	0.25	0.16	0.10	0.18	0.44	0.33
ours (単一エンティティ) - Coarse	0.45	0.30	0.21	0.12	0.18	0.45	0.35
ours (単一エンティティ) - Fine	0.30	0.14	0.07	0.04	0.13	0.34	0.15
ours (複数エンティティ) - Coarse	0.27	0.19	0.16	**0.14**	0.13	0.27	0.24
ours (複数エンティティ) - Fine	0.19	0.11	0.08	0.04	0.10	0.20	0.12
	Vocabulary	Unique Bi-gram	DIV-1	DIV-2	LSA	Self-CIDEr	
E-C2F	533	1,205	0.32	0.53	0.31	0.48	
ours (単一エンティティ)	**1,078**	**2,549**	**0.61**	**0.89**	**0.58**	**0.83**	
ours (複数エンティティ)	**656**	**1,247**	**0.60**	**0.86**	**0.56**	**0.81**	

表 4　生成されたキャプションの例

	単一エンティティ				複数エンティティ
					
ours - Coarse	a man	a woman	a hat	a man	a group of people
ours - Fine	**a bagpipe player**	**a mother**	**a fedora hat**	**a customer**	**a family**
E-C2F - Coarse	a man	a little girl	a hat	a man	N/A
E-C2F - Fine	a man	a little girl	a black hat	a man	N/A

ることが示された.

生成されたキャプションの例を表 4 の左に示す. 先行研究と比較して, より抽象的, 具体的なキャプションを生成することを達成した.

4.4　複数エンティティのキャプション生成

Coarse, Fine の数をおおよそ等しくするため, 本研究では提案手法の閾値を 0.55 とした. まず正確性, 多様性の評価を行った. 複数エンティティのキャプション生成の正確性と多様性の実験結果を表 3 に示す. 単一エンティティに比べ, 多様性は近い値を得ているが, 正確性は低下した. 原因として複数エンティティのキャプション生成のほうが考慮する情報が多く, より難易度が高いことが考えられる. また単一エンティティに比べ, データが少ないので語彙数が少なくなっている.

次に抽象度の評価を行った. E-C2F は複数エンティティをデータセットから除いており, 単一エンティティのような比較ができないので, 生成された Coarse, Fine のキャプションで比較実験を行った. Coarse が前提で Fine が仮定のとき, 含意関係となるものは 3.9%であったのに対し, 逆は 47.2%であった. 単一エンティティではあるが, E-C2F で同様の実験を行ったところ, Coarse が前提で Fine が仮定のとき, 含意関係となるものは 7.1%であったのに対し, 逆は 23.6%であった. よって提案手法により, E-C2F に比べ, 抽象度を制御可能であることが示された.

生成されたキャプションの例を表 4 の右に示す. 先行研究と比較はできないが, 抽象的, 具体的なキャプションを生成することを達成した.

5　おわりに

本研究では, IDF と含意関係認識を用いて抽象度を制御可能なエンティティレベルのキャプション生成手法を提案し, 従来の手法より多様性を向上させ, より具体的なキャプションを生成可能にし, 抽象度の制御を可能にした. またエンティティの抽象度を制御可能にすることで, 常識推論の導入, そして画像全体の深い理解に繋がると考える. 今後の展望として, 本研究の手法を画像全体に拡張することが挙げられる.

謝辞

本研究は富士通株式会社の助成を受けたものである.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. In **arXiv**, 2021.

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

[3] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. Dense captioning with joint inference and visual context. In **CVPR**, 2017.

[4] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, and Jing Shao. Context and attribute grounded dense captioning. In **CVPR**, 2019.

[5] Linwei Li, Xiangyu Lin, Chenhui Chu, Noa Garcia, Yuta Nakashima, and Teruko Mitamura. Achieving entity-level diverse caption generation via attention over neighbors. **Computer Vision and Image Understanding (under review)**.

[6] Chenhui Chu, Vinicius Oliveira, Felix Giovanni Virgo, Mayu Otani, Noa Garcia, and Yuta Nakashima. The semantic typology of visually grounded paraphrases. **Computer Vision and Image Understanding**, Vol. 215, No. 103333, 2022.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In **CVPR**, 2016.

[8] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In **ICCV**, 2015.

[9] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In **EMNLP**, 2015.

[10] Samuel R. Bowman Adina Williams, Nikita Nangia. A broad-coverage challenge corpus for sentence understanding through inference. In **NAACL**, 2018.

[11] Yixin Nie, Haonan Chen, and Mohit Bansal. Combining fact extraction and verification with neural semantic matching networks. In **AAAI**, 2019.

[12] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Proceedings of the 58th annual meeting of the association for computational linguistics. In **ACL**, 2020.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A　提案手法のスコアの例

提案手法のスコアの例を表 5 に示す.

表 5　提案手法のスコアの例

Caption	Entailment	IDF	Sum
a woman	0.02	0.08	0.10
a young boy	0.06	0.20	0.27
a group of people	0.07	0.29	0.36
two women	0.07	0.32	0.40
a side walk	0.12	0.46	0.59
a motorcyclist	0.00	0.89	0.89
a family	0.26	0.63	0.90
a tennis ball	0.60	0.56	1.16
chess	0.50	0.85	1.35
a cigar	1.00	0.89	1.89

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).