エッジプロービングを用いた事前学習済みの視覚と言語に基づくモデルにおける言語知識の分析

白井尚登 [1]　上垣外英剛 [1]　渡辺太郎 [1]
[1] 奈良先端科学技術大学院大学
{shirai.naoto.sq5,kamigaito.h,taro}@is.naist.jp

概要

近年 Transformer ベースの Vision-and-Language (V&L) モデルが次々に提案され, 画像情報に対する質問応答などのマルチモーダルタスクで成功を収めている. 一方で，V&L モデルが適するタスクを把握するためには，V&L モデルが有する言語処理能力を知ることが必要である．本研究ではエッジプロービングというフレームワークを採用し, 事前学習済み (V&L) モデルである VisualBERT と LXMERT が有する言語知識の分析を行う. このプロービング手法を用いることで品詞タグ付けなど 8 つの分類タスクのスコアの算出と，分類に寄与する隠れ層の定量化を実現する. 実験の結果,VisualBERT は全体的に BERT とタスクごとの精度や分類に寄与する層が近しい一方,LXMERT は全体的に精度が低く, より低層の情報が寄与していることが判明した.

1　はじめに

近年，Vision-and-Language (V&L) モデルとして，言語と画像を横断した汎用的なマルチモーダル表現の獲得のために,Transfomer に基づく事前学習済みモデルが次々と提案されている [1].V&L モデルには BERT [2] を模したエンコーダを用いるものが多く，言語と画像を含む入力に対する扱いの差異によって，Single-stream モデルと Dual-stream モデルに大別される (図 1)．Single-stream モデルは言語と画像を一つのエンコーダで処理し，Dual-stream モデルは言語と画像を別々のエンコーダで処理した後, その結果を一つのエンコーダで統合する階層的なモデル構造である. これらのモデルは画像情報を踏まえた質問応答 [3] などマルチモーダルタスクの精度向上に貢献している [1].

現在，これらの V&L モデルが視覚情報を学習することで，テキストのみを学習したモデルよりも良い文脈表現を獲得できているのかの調査が行われている. 先行研究 [4, 5] では，事前学習済み V&L モデルの言語理解能力をベンチマーク GLUE [6] で評価し，Single-stream モデルは Dual-stream モデルよりも僅かに高い言語理解能力を有しているが，単に画像情報を学習するだけでは言語理解能力が向上しないことを主張している. また, これらの研究ではテキストのみの評価データに対として入力する視覚情報の差異による性能変化についても議論されている.

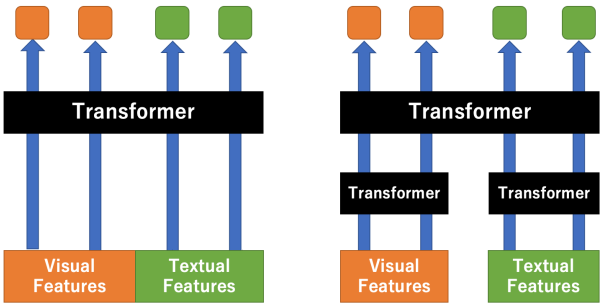

図 1　Single-stream(左) と Dual-stream(右) モデルの構成. 入力した視覚情報, 言語情報に対する処理方法が異なる.

このように V&L モデルにおける検証ではベンチマークスコアに基づいた方法が行われている一方，自然言語処理分野では事前学習済みモデルに対するより詳細な分析手法として，**エッジプロービング** [7] と呼ばれるフレームワークが提案されている. エッジプロービングは事前学習済みモデルにトークン列からなるスパンを入力し，出力された表現を利用した分類器が対応するラベルをタスクごとに予測することで行われる. この際，分類器に入力するスパン表現はパラメータを固定したモデルの各層に重み付けをする**スカラーミキシング**により得られる. エッジプロービングとスカラーミキシングを組み合わせた手法は学習した各層への重みを参照することで，分類タスクごとに寄与する層を定量的に評価できる利点がある.

自然言語処理分野ではこのアプローチで BERT を

分析することにより,BERT の隠れ層は品詞タグ付け, 構文解析, 固有表現抽出といった深層学習導入以前の自然言語処理で使用されていた逐次的な言語処理と対応する関係を持つという画期的な知見が得られている [8].

一方で, 事前学習済みの V&L モデルに対しては, エッジプロービングを用いてテキストとともに画像情報を入力する VisualBERT [9], 動画情報を入力する VideoBERT [10] の調査がなされている [11]. この研究でも視覚情報の学習が言語処理能力の大幅な改善に繋がっていないという見解が示された. なお, Dual-stream モデルが有する言語知識に関しては言語エンコーダに限定し, 文埋め込みに基づくベンチマーク SentEval [12] によって評価されている [13].

このように V&L モデルに対して様々な検証が行われている一方，既存研究では V&L モデルで各隠れ層がどの分類タスクに寄与しているのかの十分な分析がなされておらず，自然言語処理データで獲得された層間の関係がどれほどモデルに残されるのかも明らかにはなっていない．同様に Single-stream モデルや Dual-stream モデルといった構造の違いが各層に与える影響についても明らかではない．

本研究ではこれらの点を明らかにするために，事前学習済みの VisualBERT と LXMERT [14] の有する言語知識に焦点を当て，エッジプロービングとスカラーミキシングの手法からタスクごとの精度や自然言語タスクの解決に寄与する層が何かを定量的に示す．実験の結果，VisualBERT はタスクに応じた層を利用することで視覚情報を学習しながらも BERT に劣らない精度を維持する一方，LXMERT は全体的に精度が低く，より低層の情報が自然言語処理に寄与していることが判明した．

2 V&L モデル

本研究では異なる構造 (Stream) のモデルが有する言語知識を評価するために Single-stream モデルである VisualBERT と Dual-stream モデルである LXMERT を比較する．以下では両モデルの概要を紹介する．

2.1 VisualBERT

VisualBERT [9] は最初期の事前学習済み V&L モデルとして知られており [1]，画像ベクトル表現とテキストの埋め込み表現を連結して入力する Single-stream モデルである．初期化は事前学習済みの $\text{BERT}_{\text{BASE}}$ [2] の重みを引き継ぐことで行われ，

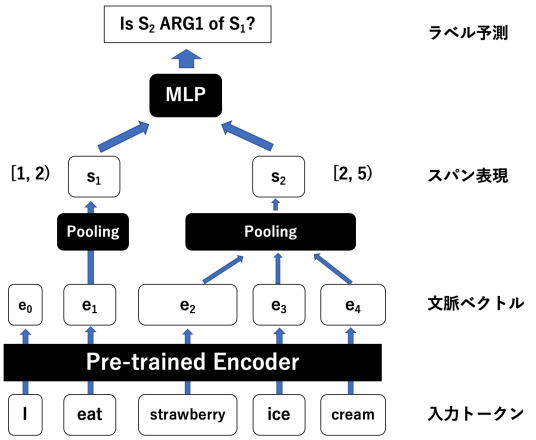

図 2 意味役割 (SRL) タスクのプロービングモデルの例. 与えられたスパンから”strawberry ice cream” (s_2) が”eat” (s_1) の対象となるか予測する. スパン表現は固定した事前学習済みモデルから得たトークンの表現ベクトル $\mathbf{e} = [e_0, e_1, \cdots, e_n]$ をプーリングすることで獲得する.

MS COCO [15] によって追加の事前学習が行われる．また，VQA [16] などの下流タスクへの適用はファインチューニングで行う. 画像の特徴量は Faster R-CNN [17] でオブジェクト領域のバウンディングボックスを検出し，エンコードすることで獲得される．

2.2 LXMERT

LXMERT [14] は画像と言語を別々のエンコーダで学習し，両モダリティの関係性を上層のクロスモダリティエンコーダで学習する Dual-stream のモデルである．事前学習時にはデータセットとして MS COCO,VQA,Visual Genorm [18],GQA [19],VG-QA [20] を使用する．入力する画像表現には VisualBERT と同様に Faster R-CNN で検出した特徴量を用いる．

3 分析手法

3.1 エッジプロービング

本研究は事前学習済みモデルが有する言語知識を調査するために Tenney らが提案した**エッジプロービング** [7] というアプローチを採用する．エッジプロービングは分類器を通して事前学習済みモデルに含まれる言語構造に関する情報を抽出して調査することを目的としている．この分類器は事前学習済みモデルに与えられたトークンのスパンに対する出力を入力として受け取り，品詞情報などタスクに関するラベルを予測する．図 2 に処理手順の例を示

す．この例では分類器はスパン $s_1 = [i_1, j_1)$ とスパン $s_2 = [i_2, j_2)$ に対応する表現を受け取る．スパン表現は与えられたスパンのトークンの埋め込み表現を各層の活性度を重み付け (スカラーミキシング §4) し，プーリングしたものである．なお，事前学習時に獲得した言語知識を調査するためにエンコーダの重みは固定し, 分類器をタスクごとに学習する．

3.2 分類タスクとデータセット

本研究では言語知識に関する分類タスクとして 8 つのラベル付けタスクを対象とし，結果を micro-F1 スコアで評価する．データセットには依存関係 （Deps.）に English Web Treebank [21]，非文法的な意味役割 (SPR2) には SPR2 [22]，関係分類 （Relations）として SemEval 2010 Task 8 [23] を使用する．そして，品詞 (POS), 句構造 (Consts.), 固有表現 (Entities), 意味役割 (SRL), 共参照 (Coref.) には OntoNotes 5.0 [24] を用いる．これらタスクの一例は付録 A.1 に記載している．なお，POS, Consts., Entities は単一のスパンからラベルを予測をするため，プロービングでは図 2 のような s_2 を使用しない．

3.3 入力画像

本分析では言語知識を問うデータセットがテキストのみで構成されているため，V&L モデルに入力する視覚的特徴が問題となる. 先行研究ではエッジプロービングによって言語知識を定量化する際に事前学習済み V&L モデルには視覚情報を入力せず，テキスト入力のみで調査している [11]. 一方，GLUE で評価する際には黒画像 (224 × 224 ピクセル) を使用し，特徴量抽出のための Faster R-CNN detector [25][1]で最初に検出した 36 個のバウンディングボックスの特徴量をエンコードしている [4]. 本研究でも黒画像の入力を採用しているが，黒画像はあくまで入力する視覚的特徴の代替案の一つであることに注意されたい．黒画像の代わりにゼロベクトルや V&L タスク用データセット内の画像の特徴量平均などを視覚情報の代替として入力することが提案されている [5]．なお，VisualBERT はテキストのみの入力が可能であるが，LXMERT は言語と画像を別々のエンコーダで処理し，クロスモダリティエンコーダでその結果を統合している．したがって，テキストのみを入力する場合にはクロスモダリティエンコーダ内の視覚側のエンコーダからの残差接続を無効にする必要がある．

3.4 分析モデル

本研究では V&L モデルとして Huggingface Transformers [26] で提供され,2023 年 1 月現在一番利用されている事前学習済み VisualBERT[2] 及び LXMERT[3] のパラメータを用いてモデルの有する言語知識の調査を行った.VisualBERT は 12 層を使用し, LXMERT は 8 層の言語エンコーダと 5 層のクロスモダリティエンコーダの計 13 層を調査対象とする. また, 自然言語のみで学習されたモデルとして, 12 層の事前学習済み BERT[4] を使用する．

4 評価方法

我々はプロービングを行う際に言語知識の獲得にモデルのどの層が寄与しているのか定量化する．そのため，ELMo [27] で提案された**スカラーミキシング**という手法を使用する．スカラーミキシングは学習可能なパラメータをエンコーダの層の数だけ用意し, 分類タスクごとに寄与する層を重み付けとして学習させることで全層を通した文脈ベクトルを導出する．具体的にはタスク τ ごとに各層 $\ell = [0, 1, \cdots, L]$ に対応した学習可能なパラメータ $a_\tau = [a_\tau^{(0)}, a_\tau^{(1)}, \cdots, a_\tau^{(L)}]$ を用意し，それらをソフトマックス関数により正規化した重み s_τ や調整パラメータ γ_τ からあるトークン h_i のタスクに応じた表現ベクトル $h_{i,\tau}$ を導出する．まとめると以下の計算式 1, 2 となる．

$$s_\tau = softmax(a_\tau) \tag{1}$$

$$h_{i,\tau} = \gamma_\tau \sum_{\ell=0}^{L} s_\tau^{(\ell)} h_i^{(\ell)} \tag{2}$$

重み a_τ はプロービング分類器とともにタスクごとに学習され，ソフトマックス関数により正規化されることで分類タスクに対する各層の寄与度を理解しやすくなる．ここで，各層の重み $s_\tau^{(\ell)}$ はどの層の情報をどの程度で使用するかという係数として捉えることができるため，より重みが高いほどその層がその特定のタスクに関連する情報をより多く含んでいることの根拠と解釈することができる．

重心 モデルのどの層が主にタスクごとの分類に寄与しているのか分かりやすくするため，本実験で

1) 本研究の物体検知には https://github.com/peteanderson80/bottom-up-attention#demo で提供されている事前学習済みモデルを使用している．

2) https://huggingface.co/uclanlp/visualbert-vqa-coco-pre

3) https://huggingface.co/unc-nlp/lxmert-base-uncased

4) https://huggingface.co/bert-base-uncased

表 1　3 回の実験から得たタスクごとの micro-F1 スコアの平均. 括弧内は学習した重みの重心の平均.

Model	Stream	POS	Consts.	Deps.	Entities	SRL	Coref.	SPR2	Relations
BERT	-	96.59 (5.7)	87.00 (6.0)	95.02 (6.5)	96.02 (6.7)	91.09 (6.6)	90.42 (7.8)	83.90 (6.1)	81.91 (6.7)
VisualBERT	Single	95.93 (5.7)	85.79 (6.2)	94.34 (6.6)	94.90 (5.8)	90.09 (6.6)	88.12 (7.4)	83.49 (6.1)	80.81 (6.6)
LXMERT	Dual	87.09 (4.7)	73.07 (5.4)	86.55 (5.5)	87.13 (3.0)	78.70 (5.0)	78.65 (5.0)	81.38 (6.4)	62.33 (6.7)

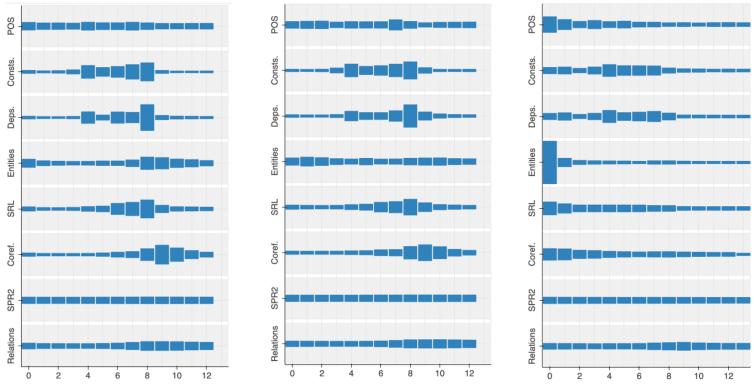

図 3　各層に対応する重み付けを視覚化した一例. 左から BERT,VisualBERT,LXMERT. タスクは表 1 の左右が上下に対応.

は学習した重みの重心を以下のように求める 3.

$$\bar{E}_s[\ell] = \sum_{\ell=0}^{L} \ell \cdot s_\tau^{(\ell)} \tag{3}$$

これは各タスクに寄与した平均的な層を反映しており，この重心の値が分類タスクを解く際に重要な層を示していると解釈できる [8].

5　実験結果

5.1　分類スコアの比較

表 1 に事前学習済みモデルを使用した分類タスクごとの micro-F1 スコアを示す．VisualBERT は先行研究 [11] と同じく大きな精度低下は見られない．これは引き継いだ BERT のパラメータが画像を含めた追加的な事前学習を経ても維持されているためだと考えられる．一方，LXMERT は BERT に対して全体的にスコアが低い．特に低下した Relations タスクの詳細は付録 A.2 に示す．この結果は Single-stream よりも Dual-stream モデルの方が言語処理能力が低いという先行研究の結果 [4, 13] と一貫する．

5.2　各層の重みと重心

続いて表 1 で示した重心とスカラーミキシングにより学習されたタスクごとの各層の重みを示した図 3 に基づいた議論を進める．VisualBERT は BERT と同様の傾向を示し，品詞情報などの語彙に強く依存する情報は下層の，固有表現などの文脈理解を求める情報はより高層の埋め込み表現が分類タスクに寄与している．対する LXMERT は 2 つのモデルとは異なり，タスクの種類に関係なく，低層の情報を使用する傾向がある．特に Entities のタスクでは BERT や VisualBERT よりも低い層を重視している．また,LXMERT は 9 層以降の情報をあまり使用していないことからクロスモダリティエンコーダは言語処理タスクにあまり寄与していないと考えられる．以上の結果より,VisualBERT の言語処理能力の高さは BERT で学習された各層の関係が維持されているためであると考えられる．

6　おわりに

本研究では VisualBERT や LXMERT の有する言語知識をエッジプロービングとスカラーミキシングによって調査した．その結果，VisualBERT が示す高い性能は BERT により獲得された言語知識が引き継がれているためである可能性が示唆された．また，LXMERT はタスクにあまり関係なく低層の情報が分類タスクに寄与している傾向が判明した．このことより，言語と画像の情報を統合するクロスモダリティエンコーダで学習される表現は言語処理タスクに対してはあまり寄与しないことも明らかにした．

謝辞

本研究は JSPS 科研費 JP21K17801 の助成を受けたものです.

参考文献

[1] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. **arXiv preprint arXiv:2202.10936**, 2022.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. **CoRR**, Vol. abs/1505.00468, , 2015.

[4] Taichi Iki and Akiko Aizawa. Effect of visual extensions on natural language understanding in vision-and-language models. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 2189–2196, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[5] Lovisa Hagström and Richard Johansson. How to adapt pre-trained vision-and-language models to a text-only input? In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 5582–5596, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[6] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. **CoRR**, Vol. abs/1804.07461, , 2018.

[7] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In **International Conference on Learning Representations**, 2019.

[8] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics.

[9] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. **arXiv preprint arXiv:1908.03557**, 2019.

[10] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In **Proceedings of the IEEE/CVF International Conference on Computer Vision**, pp. 7464–7473, 2019.

[11] Tian Yun, Chen Sun, and Ellie Pavlick. Does vision-and-language pretraining improve lexical grounding? **CoRR**, Vol. abs/2109.10246, , 2021.

[12] Alexis Conneau and Douwe Kiela. Senteval: An evaluation toolkit for universal sentence representations. **CoRR**, Vol. abs/1803.05449, , 2018.

[13] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. **CoRR**, Vol. abs/2005.07310, , 2020.

[14] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. **CoRR**, Vol. abs/1908.07490, , 2019.

[15] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. **CoRR**, Vol. abs/1405.0312, , 2014.

[16] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. **CoRR**, Vol. abs/1612.00837, , 2016.

[17] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. **CoRR**, Vol. abs/1506.01497, , 2015.

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. **CoRR**, Vol. abs/1602.07332, , 2016.

[19] Drew A. Hudson and Christopher D. Manning. GQA: a new dataset for compositional question answering over real-world images. **CoRR**, Vol. abs/1902.09506, , 2019.

[20] Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. **CoRR**, Vol. abs/1511.03416, , 2015.

[21] Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. A gold standard dependency corpus for English. In **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)**, pp. 2897–2904, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[22] Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. Universal decompositional semantics on Universal Dependencies. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 1713–1723, Austin, Texas, November 2016. Association for Computational Linguistics.

[23] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In **Proceedings of the 5th International Workshop on Semantic Evaluation**, pp. 33–38, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[24] Mitchell Marcus Eduard Hovy Sameer Pradhan Lance Ramshaw Nianwen Xue Ann Taylor Jeff Kaufman Michelle Franchini Mohammed El-Bachouti Robert Belvin Ann Houston Ralph Weischedel, Martha Palmer. Ontonotes release 5.0 ldc2013t19. Web Download. Philadelphia: Linguistic Data Consortium, 2013.

[25] 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 visual question answering. In **2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 6077–6086, 2018.

[26] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. **CoRR**, Vol. abs/1910.03771, , 2019.

[27] Matthew E. 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)**, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

A 付録

A.1 タスクの例

表 2 タスクごとの文章, スパン, ターゲットラベルの例.

タスク	例
品詞 (POS)	$[\text{Look}]_1$ at my hands. → VB (Verb)
句構造 (Consts.)	$[\text{The king}]_1$ said to him, "That's great !." → NP (Noun Phrase)
依存関係 (Deps.)	I $[\text{love}]_2$ $[\text{them}]_1$. → obj (object)
固有表現 (Entities)	This is $[\text{New York City}]_1$. → GPE (Geo-Political Entity)
意味役割 (SRL)	$[\text{The meeting}]_2$ $[\text{was}]_1$ very confused. → Arg1 (Agent)
共参照 (Coref.)	When $[\text{the men}]_2$ heard this, $[\text{they}]_1$ became very angry. → True
非文法な意味役割 (SPR2)	$[\text{He}]_1$ $[\text{deserved}]_2$ respect. → {awareness, . . . }
関係分類 (Relations)	The $[\text{light}]_1$ in the background is from $[\text{the sunrise}]_2$. → Cause-Effect(e_2, e_1)

A.2 関係分類 (Relations) タスクの詳細

表 3 関係抽出 (Relations) タスクのラベルと文章例. 各ラベルは因果 (Cause-Effect), 道具と使用者 (Instrument-Agency), 生産物と生産者 (Product-Producer), 物体と入れ物 (Content-Container), エンティティと起源 (Entity-Origin), エンティティと目的地 (Entity-Destination), 構成物と全体 (Component-Whole), 一部と集合したもの (Member-Collection), メッセージとトピック (Message-Topic) との関係性を表す.

ラベル	文章例
Cause-Effect (CE)	The $[\text{light}]_1$ in the background is from $[\text{the sunrise}]_2$. → Cause-Effect(e_2, e_1)
Instrument-Agency (IA)	The $[\text{shaman}]_1$ cured him with $[\text{herbs}]_2$. → Instrument-Agency(e_2, e_1)
Product-Producer (PP)	The $[\text{company}]_1$ fabricates plastic $[\text{chairs.}]_2$ → Product-Producer(e_2, e_1)
Content-Container (CC)	I lost a $[\text{suitcase}]_1$ with $[\text{money}]_2$ in it last week. → Content-Container(e_2, e_1)
Entity-Origin (EO)	The $[\text{woman}]_1$ was born in the $[\text{village.}]_2$ → Entity-Origin(e_1, e_2)
Entity-Destination (ED)	$[\text{Dioxide}]_1$ has been released into the $[\text{atmosphere.}]_2$ → Entity-Destination(e_1, e_2)
Component-Whole (CW)	A $[\text{bird}]_1$ is touching the sun with his $[\text{wing.}]_2$ → Component-Whole(e_2, e_1)
Member-Collection (MC)	The $[\text{lawyer}]_1$ was a member of the $[\text{team.}]_2$ → Member-Collection(e_1, e_2)
Message-Topic (MT)	The $[\text{section}]_1$ concludes by giving a summary of $[\text{findings.}]_2$ → Message-Topic(e_1, e_2)

表 4 3 回の実験から得た Relations タスクの micro-F1 スコアの平均. VisualBERT は BERT と比べ, 関係分類タスク全体 (ALL) だけではなくラベルレベルでも精度に大きな差は見られない. 一方, LXMERT は BERT に対してタスク全体もラベルレベルでもスコアが低い傾向がある. 特に生産物と生産者との関係性を問う PP の精度が大幅に低下している.

Model	CE	IA	PP	CC	EO	ED	CW	MC	MT	Other	ALL
BERT	88.29	72.97	78.18	79.95	75.12	83.92	78.96	86.83	86.16	33.42	81.91
VisualBERT	86.94	72.64	76.69	77.03	74.62	85.80	77.57	84.43	83.96	28.54	80.81
LXMERT	62.21	51.07	39.60	72.43	51.18	70.66	68.51	71.51	60.21	11.19	62.33