百聞は一見に如かず？
視覚情報は言語モデルに文の階層構造を教示するか

栗林 樹生
東北大学　Langsmith 株式会社
kuribayashi@tohoku.ac.jp

概要

巨大言語モデルの進展が著しい反面，人間の言語獲得と対照すると言語モデルのデータ効率の悪さが強調される．本研究では，なぜ人間の言語獲得が効率的なのかという問いを，人間と言語モデルの言語獲得シナリオのどのような差異を埋めると両者の乖離が縮むかという問いに読み替え，特に人間のみが言語獲得時に活用し得る視覚情報の寄与を調査する．少なくとも本実験の範囲では，言語的な事前知識のない学習者（ニューラルモデル）がリアルな画像・文対から言語の適切な汎化規則を支持する手がかりを見出すことは容易でなく，単なる視覚情報の有無のみでは，人間と言語モデルの言語獲得効率の差異を説明できないことが示唆された．

1　はじめに

巨大言語モデルが進展を遂げた一方，人間の言語獲得効率の良さが対照的に強調されている．例えば GPT-3 [1] は，人間が 10 歳までに享受する言語刺激のおよそ 2000 倍のデータで学習されており [2]，非常に単純な計算では人間の 20,000 年分の学習に匹敵する．それではなぜ人間は言語獲得のデータ効率が良いのだろうか？本研究では，この問いを**言語モデルと人間のどのような差異を埋めることで両者の言語獲得効率の乖離が縮まるか**という問いに読み替え，効率的な言語獲得が実現される十分条件について計算機上での概念実証を目指す．

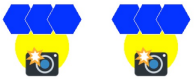
本研究では特に，記号接地 [4] や身体性 [5] の問題に関連する**視覚情報の影響**に焦点を当てる．通常人間は，言語モデルと異なり，言語獲得中に視覚情報を利用でき，効率的な言語汎化に貢献している可能性がある．実験では主語・動詞の数の一致を手がかりに，言語の階層的な汎化の達成に視覚情報が寄与するかを検証する（図 1）．なお近年の視覚・言語処理では大規模モデルの開発が盛んであるものの [6, 7]，言語と視覚の相互作用に関する言語科学的な分析や学習過程の分析は限られている [8, 9]．

図 1　実験の概要．文に対応する視覚情報が，汎化規則の曖昧なデータの下での言語の階層的規則獲得を促進するか調査する．なお図中の画像は，例示のために DALL-E [3] で生成したものである．

リアルなキャプションデータを用いた実験では，**文に対応する静止画像を単にモデルに入力するだけでは，言語の階層的汎化は促されない**ことが観察された．一方，問題を抽象化・簡略化した**人工データを用いた実験では，言語の階層的な汎化に対する視覚情報の著しい寄与が確認された**．この設定ではデータの偏りやキャプションに関係のない視覚情報は捨象されており，閉じた世界の中で学習者が視覚情報の適切な抽象化をできた状況を想定している．

少なくともリアルなデータを用いた今回の実験からは，言語的な事前知識を持たない学習者（ニューラルネット）が，静止画像と文の対から自然言語の

表 1　画像キャプションペアの例．AMBIG. データでは，動詞に対応する主語と動詞に最も近い名詞の数が同じであり，DISAMBIG. データでは両者の数は異なる．AMBIG. データが大部分を占める学習データでモデルを訓練し，そこで学習された規則を DISAMBIG. データを用いて峻別する．表中の画像は作成したデータセットから抽出したものである．

種類	自然画像キャプション	人工画像キャプション
AMBIG.		*girl aged stands with a hand on a tree alone*		*a lime rectangle with a red rectangle waves its hand*
		*young boys with school uniforms and backpacks prepare for school on an early morning*		*two yellow circles with three blue hexagons take a photo*
DISAMBIG.		*young girls dressed in colonial gear tie their shoes at farm*		*two red rectangles with a black circle play soccer*

適切な規則性を見出すことは容易でないこと，人間と言語モデルの学習効率の違いは単なる視覚情報の有無のみでは説明できないことが示唆された．一方人工データの結果を踏まえると，(i) 視覚情報を適切に抽象化する特別な視覚・言語能力を学習者に仮定できれば，視覚情報が言語獲得に良い影響を及ぼす可能性があることや，(ii) 実験で用いた事前学習済み画像エンコーダでは，リアルな画像から言語獲得に影響を及ぼすような視覚情報の抽出・抽象化ができていない可能性なども強調された．

2　背景

一般に，有限のデータから未知のデータへ汎化する際に汎化規則は一意に定まらず，規則の選択はモデルの帰納バイアスに左右される [10]．言語獲得においても，幼児が限られた言語刺激から適切な（階層的な）汎化を達成する現象を説明するには，学習者の強力な帰納バイアスが必要であると主張されてきた [11, 12]．一方，通常のニューラルモデルでは，表層的手がかりの利用や言語的に妥当な汎化を完全に達成できないなどの人間らしくない学習傾向があり [13, 14, 15, 16, 17]，分野全体の問題として認知されている．またこのバイアスを覆すために大量のデータが必要であることも指摘されている [14, 15]．

汎化を導く帰納バイアスとしては，学習者の性質に由来するもの（生得的要因）と，学習データなどの学習環境に由来するもの（環境的要因）が想定され，本研究では視覚情報の有無という環境的要因の影響を，計算機シミュレーションのもと調査する．言語理解における視覚情報の重要性は長らく説かれているものの [18]，（人工知能の）言語獲得の視点では，適切な言語知識なしに視覚情報を与えても，むしろ表層的な疑似相関などの可能性が増え問題が過度に複雑になるといった批判もあり [19, 20]，視覚情報が言語学習にどのような影響を与えるかは自明でない．なお図 1 のような線形・階層的な規則の対立において，人間が階層的汎化を好むことは心理実験などで検証されてきた [21, 22]．また，ニューラルモデルを用いた計算言語学的な検証は条件を統制しやすいなどの利点をもち，人間を直接観察する方法論と相補的な役割を果たしてきた [23, 12, 24]．

3　問題設定

概要を図 1 に示す．主語と動詞の数の一致の観点から，線形的・階層的規則両方に整合するキャプションデータのもとでキャプション生成モデル（視覚情報が利用可能な言語モデル）を訓練し，視覚情報の有無で言語の汎化傾向が変わるかを調べる．視覚情報のもとでは，階層的規則に整合する動詞の数を選ぶ問題は，典型的にはその動作をしているものが 1 つかそれ以上かを数える問題となり（表 1），**視覚情報により階層的規則に紐づく特徴量がより顕在的になる可能性**がある[1)]．

3.1　データ

既存の画像キャプションデータを収集したリアルなデータと，人工的に統制されたデータを用い，相補的な検証を行う．既存研究では人工コーパスでの実験が主流であり [16]，**視覚モダリティを追加したとともにリアルなデータと人工的なデータで結果を比較したこと**も本論文の貢献である．

自然画像キャプション: Conceptual captions コーパス [27] から，以下を満たすデータを抽出した：

1)　言語学的な数が，いわゆるものの数に必ずしも一致しないことは指摘されており，議論の余地はある [25, 26]．

- キャプションに主語・述語が存在する[2)]
- キャプションの主語が *family*, *pair of* など単複判断の難しい集合名詞・表現でない
- キャプションに文法誤りがない[3)]

これらのうち，主動詞の数と直前の名詞の数が同一である，線形的・階層的規則を峻別できないデータ (AMBIG.) と，主動詞の直前の名詞と同士の数が異なる，階層的規則を支持するデータ (DISAMBIG.) に分けた (表 1)．DISAMBIG. のような階層性を明示するデータは実際の言語刺激において非常に少ないという議論（刺激の貧困）[28, 21] をもとに，本研究では AMBIG. データと DISAMBIG. データが 100:1 になるように学習データ（計 352,359 画像キャプションペア）を作成した[4)]．残りの DISAMBIG. データ（1,269 ペア）を評価データとした．

人工画像キャプション: `NUM1 COLOR1 SHAPE1 with NUM2 COLOR2 SHAPE2 VP` というテンプレートから文と画像をルールで生成した（付録 A）．表 1 に例を示す．`VP` の主語は `SHAPE1` であり，画像中では `VP` と `SHAPE1` オブジェクトが近接し，文中では数（屈折）が一致している．この設定では少なくとも，自然なデータがもつ以下の性質を捨象している：

- 単語や画像の特徴に関するクラス不均衡性・多クラス性・クラスの開放性
- 着目すべき対象以外の視覚情報（背景など）
- ものや動作の呼び名と視覚特徴間の多対多関係
- 動作主と動作の視覚的に自然な構成（「人が走る」画像では，本来人の姿勢が変わるべきだが，本データでは「走る」を示す画像を「人」の画像に重ねるようなアプローチをとっている）

上で述べた手続き同様，数の一致から AMBIG. データと DISAMBIG. データに分け，100:1 で混ぜたものを学習データ（16,500 ペア），残りの DISAMBIG. データを評価データ（11,300 ペア）とした．

3.2 評価

評価データにおいて，動詞の数のみが異なる 2 つのキャプションを用意し (例えば，*a red circle with three black circles* ***plays/play*** *a soccer*)，対応する画像を入力したもと，どちらのキャプションに高い生成確率を付与するかを観察する．階層的規則に整合するキャプションを正解とし，二値分類問題としてモデルの選好を Macro F1 値で評価した．キャプション生成能力の目安として，開発用データ[5)]で ROUGE-L F1 値 [29][6)] も計算した．また言語獲得効率に関心があるため，学習初期ステップでの値も報告する．

3.3 モデル

Transformer 系列変換モデルを用いる [30]．ただし入力が画像，出力が文となるキャプション生成モデルであり，エンコーダには Vit [31] などの事前学習済みエンコーダを，デコーダには GPT-2 small（124M） [32] アーキテクチャに交差アテンションを追加したものを採用する[7)]．直観的には，交差アテンションを通して視覚情報にアクセス可能な（文レベル）言語モデルである．視覚情報を与えない設定では学習・推論時共に無情報なノイズ画像を入力する．ゼロからの言語獲得シナリオを想定し，デコーダは初期化して学習を始めた．また，交差エントロピー誤差でモデルを訓練する．

エンコーダの初期値・構造として，大きさの異なる 3 つの Vit (base/large/huge) [31] と，Beit (base) [33], Deit (base) [34], Swin (base) [35] の計 6 種類を試し，モデル横断的な結論の一般性を検証する（付録 B）．参考として，画像エンコーダを Vit-base にした上でパラメータ初期化した設定 (Scratch) と，デコーダに学習済み GPT-2 [32] を用いる設定 (Vit-GPT2) も試した．ハイパーパラメータは付録 C に記載する．

4 実験・結果

各設定において異なる 2 つのシードでモデルを訓練し，いずれの指標においても平均値を報告する．自然/人工画像キャプション実験における 10,000/500 ステップはおよそ 2 エポック目終了時点である．

自然画像キャプション: 画像情報の有無による階層的汎化傾向の違いを表 2 中央に示す．**学習初期において，視覚情報は階層的汎化を促す方向にはあるものの，変化は限定的**であった．画像を入力とする設定では ROUGE-L F1 値は 3–40 程度に達しており，画像情報が活用されていることは確認できる．また，例えば 1000 ステップ目における階層的汎化への寄与（階層的汎化における🐵 – 🙈行）に着目すると，画像エンコーダの性能指標として用いられる ImageNet

2) Spacy で主動詞 `ROOT` が主語 `nsubj` を持つかを解析した．
3) language-tool-python 2.7.1 を用いた．
4) Conceptual captions 全体において DISAMBIG. データはおよそ 2%程度存在したため，混入の程度の参考とした．
5) 学習データとは重複のない AMBIGUOUS データ 5,000 件
6) https://huggingface.co/spaces/evaluate-metric/rouge
7) https://huggingface.co/docs/transformers/model_doc/vision-encoder-decoder の実装を用いた．

表 2 学習過程におけるモデルの階層的規則への選好（階層的汎化 F1 値）とキャプション生成能力 (ROUGE-L F1 値)．1,000，5,000，10,000 といった数字はパラメータ更新ステップ数を示す．🐵行は視覚情報ありのモデルの性能を示す．🐵 – 🙈行は視覚情報を用いるモデルの性能から用いないモデルの性能を引いた値を示し，値が高いほど視覚情報が好影響を与えたことを意味する．また参考として，各画像エンコーダのパラメータ数を一列目括弧内に記載し，エンコーダの ImageNet top-1 正解率の報告値も載せる．

		自然画像キャプション						人工画像キャプション				ImageNet
		階層的汎化 F1 ↑			ROUGE-L F1 ↑			階層的汎化 F1 ↑		ROUGE-L F1 ↑		正解率
モデル	視覚情報	1,000	5,000	10,000	1,000	5,000	10,000	100	500	100	500	@1 ↑
Vit-base (86M)	🐵	52.8	72.0	81.9	32.0	35.5	37.8	90.6	99.7	80.5	100	84.0
	🐵 – 🙈	+0.41	−2.38	−0.94	+17.3	+20.2	+22.8	**+57.4**	−0.31	+45.1	+64.5	
Vit-large (307M)	🐵	52.9	74.9	83.1	30.8	35.1	37.9	52.6	92.2	76.3	100	85.2
	🐵 – 🙈	+0.93	−1.13	+0.65	+16.1	+20.2	+22.6	**+19.4**	−7.76	+40.7	+64.5	
Vit-huge (632M)	🐵	52.6	73.9	82.6	29.2	34.1	35.8	42.6	100	59.1	100	85.1
	🐵 – 🙈	+1.98	−2.07	+0.10	+14.9	+18.8	+20.5	+9.21	0.00	+23.8	+63.9	
Beit-base (86M)	🐵	46.7	59.0	66.4	31.7	34.5	37.4	45.8	74.8	51.5	100	85.2
	🐵 – 🙈	+2.99	+5.68	−1.50	+15.9	+19.2	+22.1	**+11.7**	−25.0	+16.5	+64.6	
Deit-base (86M)	🐵	54.9	72.5	81.2	32.2	35.6	38.2	67.4	99.9	98.5	100	83.4
	🐵 – 🙈	+4.23	−1.77	−1.35	+18.5	+20.4	+22.9	**+32.9**	+0.08	+63.0	+64.4	
Swin-base (88M)	🐵	53.0	73.0	81.8	34.3	37.6	40.7	80.5	100	99.3	100	85.2
	🐵 – 🙈	+0.92	−2.61	−1.05	+19.6	+22.3	+25.4	**+33.2**	0.00	+64.0	+64.3	
Scratch (86M)	🐵	49.3	72.6	81.0	13.94	23.7	24.5	50.7	100	37.3	65.6	-
	🐵 – 🙈	+1.75	−3.22	−1.62	+0.16	+8.78	+8.93	+5.10	0.00	+1.88	30.3	
Vit-GPT2 (86M)	🐵	95.6	97.0	96.6	32.4	35.3	37.4	90.8	100	93.3	100	84.0
	🐵 – 🙈	+0.04	+0.18	−0.11	+17.7	+20.4	+22.1	−9.21	0.00	+57.7	+64.2	

正解率の高いエンコーダ（Vit-large, Swin-base）やパラメータ数の大きいエンコーダ（Vit-huge）が必ずしも好影響を与えているわけではなく，**画像処理的な指標で良い画像エンコーダや巨大なエンコーダが必ずしも言語モデリングに適切なバイアスを与えるとは限らない**可能性も予想した（付録 D）．なお Vit-GPT2 では，学習初期からほぼ完全な階層的汎化を達成しており，大規模な言語データの観察のもとでは対処可能な問題設定であった．また今回はコーパスを基に学習データへの DisAmbig. データの 1% の混入を認めたが，この程度の混入でも最終的に階層的汎化に選好が寄ることは確認された．また DisAmbig. データが 0% の場合でも予備実験をしており，この場合は線形的汎化に選好が寄り，一貫して視覚情報の好影響は観察されなかった．

人工画像キャプション: 結果を表 2 右部に示す．学習初期（100 ステップ）において，視覚情報が階層的汎化を顕著に促しており，**言語の階層的規則の効率的な獲得が達成された**．学習初期において，事前学習済みエンコーダを用いた際の階層的汎化の向上幅は Scratch（+5.10）よりも大きく，視覚的事前知識が言語の階層的汎化を促していた．なお視覚情報のない設定でもモデルごとに性能が異なり，この設定ではエンコーダが画像処理以外の用途（入力非依存なメモリなど [36]）で活用された可能性もある．

5 考察・おわりに

ニューラルモデルがリアルな画像・文対から自然言語の階層性を学ぶ手がかりを得ることは困難であることが示唆された．人工画像キャプションでの対照的な結果を踏まえると，自然な画像・言語特有の性質が視覚情報からの言語獲得を困難にしている可能性が示唆され，言語獲得の視点からは，(i) 視覚情報は言語獲得に寄与し得るが，リアルな視覚情報を適切に抽象化し言語と対応付ける能力を先立って獲得しておく必要がある，あるいは (ii) 視覚情報は言語の階層的な汎化に直接影響を与えないという仮説が導かれる．自然・人工データのどのような差異が結論の違いに紐づくのか人工データを制御して調査する方向や，データを幼児の言語獲得シナリオに近づける（絵本など）方向，韻律情報など他のモダリティを取り入れる方向などは今後検討している．また本研究は自然言語処理，画像処理，計算心理言語学の 3 領域をまたぐ学際研究として意義がある．

謝辞

日頃から助言いただいた乾健太郎先生，鈴木潤先生，原稿に助言いただいた小林悟郎氏をはじめ Tohoku NLP グループの皆様に感謝申し上げます．YANS 2022 にてコメントをくださった皆様にも感謝いたします．本研究は，JST，CREST，JPMJCR20D2 の支援を受けたものである．

参考文献

[1] Tom B Brown, et al. Language Models are Few-Shot Learners. In **Proceedings of NeurIPS**.

[2] Alex Warstadt and Samuel R Bowman. What artificial neural networks can tell us about human language acquisition. **Algebraic Structures in Natural Language**, p. 17, 2022.

[3] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image generation. Technical report, February 2021.

[4] Deb Roy and Ehud Reiter. Connecting language to the world. **Artificial Intelligence**, Vol. 167, No. 1-2, pp. 1–12, September 2005.

[5] Lawrence W Barsalou. Grounded cognition. **Annual Review of Psychology**, Vol. 59, pp. 617–645, 2008.

[6] Jean-Baptiste Alayrac, et al. Flamingo: a visual language model for Few-Shot learning. **arXiv:2204.14198**, April 2022.

[7] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional image generation with CLIP latents. **arXiv:2204.06125**, April 2022.

[8] Alexander Ororbia, Ankur Mali, Matthew Kelly, and David Reitter. Like a baby: Visually situated neural language acquisition. In **Proceedings of ACL**.

[9] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In **Proceedings of CVPR**, pp. 5238–5248, 2022.

[10] Tom M Mitchell. **The need for biases in learning generalizations**.

[11] Noam Chomsky. Rules and representations. **Behavioral and Brain Sciences**, Vol. 3, No. 1, pp. 1–15, March 1980.

[12] R Thomas McCoy, Robert Frank, and Tal Linzen. Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. **Proceedings of Cogsci**, 2018.

[13] Morten H Christiansen and Nick Chater. Toward a connectionist model of recursion in human linguistic performance. **Cognitive Science**, Vol. 23, No. 2, pp. 157–205, 1999.

[14] Alex Warstadt and Samuel R Bowman. CAN NEURAL NETWORKS ACQUIRE a STRUCTURAL BIAS FROM RAW LINGUISTIC DATA? In **Proceedings of Cogsci**, 2020.

[15] Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R Bowman. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In **Proceedings of EMNLP**.

[16] R Thomas McCoy, Robert Frank, and Tal Linzen. Does syntax need to grow on trees? sources of hierarchical inductive bias in Sequence-to-Sequence networks. **TACL**, Vol. 8, pp. 125–140, 2020.

[17] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In **Proceedings of ACL**.

[18] Emily M Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In **Proceedings of ACL**.

[19] Lila R Gleitman and Henry Gleitman. A picture is worth a thousand words, but that's the problem: The role of syntax in vocabulary acquisition. **Current Directions in Psychological Science**, Vol. 1, No. 1, pp. 31–35, February 1992.

[20] Emmanuel Dupoux. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. **Cognition**, Vol. 173, pp. 43–59, 2018.

[21] Julie Anne Legate and Charles D Yang. Empirical re-assessment of stimulus poverty arguments. **The Linguistic Review**, Vol. 19, No. 1-2, pp. 151–162, June 2002.

[22] Stephen Crain and Mineharu Nakayama. Structure dependence in grammar formation. **Language**, Vol. 63, No. 3, pp. 522–543, 1987.

[23] M W Crocker. Computational psycholinguistics. **Computational Linguistics and Natural Language**, 2010.

[24] J D Lewis and J L Elman. Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. **Proceedings of the 26th annual Boston University**, 2001.

[25] Benjamin Spector. Aspects of the pragmatics of plural morphology: On Higher-Order implicatures. In Uli Sauerland and Penka Stateva, editors, **Presupposition and Implicature in Compositional Semantics**, pp. 243–281. Palgrave Macmillan UK, London, 2007.

[26] Eytan Zweig. Number-neutral bare plurals and the multiplicity implicature. **Linguistic Philosophy**, Vol. 32, No. 4, pp. 353–407, August 2009.

[27] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In **Proceedings of ACL**.

[28] Noam Chomsky. **Problems of Knowledge and Freedom: The Russell Lectures**. New York: W.W. Norton, 1971.

[29] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In **Proceedings of text summarization branches out**, pp. 74–81, 2004.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Proceedings of NIPS**, pp. 5998–6008, 2017.

[31] Alexey Dosovitskiy, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In **Proceedings of ICLR**, 2020.

[32] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI, 2019.

[33] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In **Proceedings of ICLR**, 2021.

[34] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang, editors, **Proceedings of ICML**, Vol. 139, pp. 10347–10357. PMLR.

[35] Liu, Lin, Cao, Hu, Wei, Zhang, Lin, and Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In **Proceedings of ICCV**.

[36] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer Feed-Forward layers are Key-Value memories. In **Proceedings of EMNLP**.

[37] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In **Proceedings of CVPRW**, pp. 702–703. IEEE, June 2020.

[38] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In **Proceedings of CVPR**, pp. 12104–12113, 2022.

A　人工画像キャプションの生成

以下のようにテンプレートから文を作成した．

a　lime　circle　with　two　red　triangles runs
NUM1 COLOR1 SHAPE1 with NUM2 COLOR2 SHAPE2　VP

対応する画像を表 1 の例のように自動生成した．各記号に対応する語彙と対応する画像オブジェクトを表 3 に示す．NUM2 COLOR2 SHAPE2 オブジェクトは各 NUM1 COLOR1 SHAPE1 オブジェクトの上部に小さく配置し，VP オブジェクトは NUM1 COLOR1 SHAPE1 オブジェクトの中央下部に配置している．色の重複を防ぎ，計 3×3×5×4×4×4×10=28,800 種類の画像を生成した (16,500 学習データ，1,000 開発データ，11,300 評価データ)．

表 3　人工画像キャプション作成時の語彙と画像特徴

カテゴリ	単語	画像
NUM1/2	a	
	two	
	three	
COLOR1/2	black	
	red	
	blue	
	yellow	
	lime	
SHAPE1/2	circle(s)	
	rectangle(s)	
	triangle(s)	
	hexagon(s)	
VP	walk(s)	
	sleep(s)	
	run(s) fast	
	wave(s) its hand	
	write(s) a text	
	take(s) a bus	
	take(s) a photo	
	play(s) soccer	
	play(s) baseball	
	throw(s) an arrow at a target	

B　画像エンコーダ

各画像エンコーダの性質を簡潔に説明する．ImageNet-21k(22k) の上で訓練されたモデルを用い，解像度は 224^2，パッチサイズは 16（に近いもの）に統一している．

Vit [31]　画像を分割したパッチをトークンとみなしてエンコードする．以下の実装を用いた：https://huggingface.co/google/vit-base-patch16-224-in21k, https://huggingface.co/google/vit-large-patch16-224-in21k, https://huggingface.co/google/vit-huge-patch14-224-in21k.

Beit [33]　Vit は画像分類タスクで学習されているのに対し，Beit では画像上でマスク穴埋め形式の自己教師あり事前学習をする．また，絶対位置埋め込みを用いる Vit とは異なり，相対位置埋め込みを用いている．実装は，https://huggingface.co/microsoft/beit-base-patch16-224-pt22k-ft22k を用いた．

Deit [34]　特殊な知識蒸留法によって，モデルサイズと訓練コストを削減した．実装は，https://huggingface.co/facebook/deit-base-distilled-patch16-224 を用いた．

Swin [35]　異なる粒度で画像のパッチ分割を行い，局所的な処理を適用することで計算コストを抑えつつ，画像の階層的な特徴量を抽出する．実装は，https://huggingface.co/microsoft/swin-base-patch4-window7-224-in22k を用いた．

C　ハイパーパラメータ

全モデル共通のハイパーパラメータを表 4 に示す．エンコーダのハイパーパラメータは対応する huggingface モデルの設定に従い，ドロップアウト率のみを上書きしている．画像を提示する設定では，過学習を防ぐため学習画像に対し随時 RandAugment [37] を適用し，更に 20%の確率で実際の画像をノイズ画像に置き換えた．パラメータの固定はしていない．

表 4　共通のハイパーパラメータ

デコーダ設定	https://huggingface.co/gpt2/blob/main/config.json に従う
エンコーダ ドロップアウト率	0.1 （attention，隠れ層共に）
最適化手法	AdamW
learning rate	1e-4
betas	(0.9, 0.999)
epsilon	1e-8
学習率スケジューラ	線形減衰
max steps	10,000（自然画像キャプション） , 500（人工画像キャプション）
warm up steps	0
weight decay	0
バッチサイズ	512
ビーム幅　(生成評価時)	4

D　画像エンコーダのスケーリング

画像エンコーダに関して，ImageNet 画像認識性能についてスケーリング則が報告されている [38]．ImageNet での画像認識性能と階層的汎化の促進度に単調な関係がないことから，言語モデルに階層的なバイアスを与える画像エンコーダはスケーリングの結果得られるとは限らないことが予想された．なお本文では言及していないが，Beit-large, Swin-large, Deit-small, Deit-tiny でも実験をしており，より幅広い画像認識性能を持つエンコーダ間で傾向を比較した．傾向を図 2 に示し，画像認識性能が良いエンコーダほど，言語の階層的な汎化を促すとは限らない．

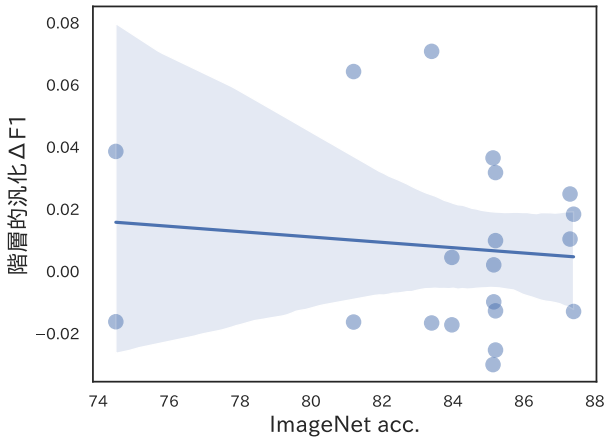


図 2　自然画像キャプションデータでの学習 1000 ステップ目における，画像エンコーダの ImageNet top-1 正解率と階層的汎化 F1 値の向上度合いの関係．　各点は各訓練試行に対応し，計 20 試行={ 画像エンコーダ 10 種 ×2 シード } の結果である．