機械・人の双方が言語で概念を説明可能な Few-shot 画像分類

西田光甫　　西田京介　　西岡秀一
日本電信電話株式会社　NTT 人間情報研究所
{kosuke.nishida.ap,kyosuke.nishida.rx,shuichi.nishioka.gd}@hco.ntt.co.jp

概要

人は自然言語による説明から新しい概念を学び，さらに概念を自然言語によって説明できる．この能力を機械学習モデルで実現することは，人と人工知能が共生する社会の実現のために重要である．本研究は few-shot 画像分類を用いて，モデルが新しい概念に関する説明を理解・生成することに挑戦した．本研究はテキストデコーダによる説明の生成と，テキストエンコーダによる説明の入力により few-shot 画像分類の性能を改善する初の研究である．評価実験により，提案手法は既存研究の分類精度を上回り，かつ人の説明からの学習および内部表現の説明を精度良く行うことを確認した．

1　はじめに

人はテキストを読むことで新たな概念を効率的に学ぶことができる [1]．機械学習のタスクあるいはデータに自然言語による説明を与えて学習に利用することは，人のような効率と精度を両立する人工知能の実現に向けて重要な課題である [2]．また，深層学習モデルは black box であるため内部表現を自然言語によって説明することも重要である [3]．

本研究では few-shot 画像分類問題を題材に，人が説明として与えた言語情報から新たな概念を獲得し，さらにモデル内で特徴量として表現されている概念を機械が言語によって説明することに取り組む．本研究の設定は，モデルが画像キャプションを追加情報として利用可能とする．この設定は，人が少ないデータを用いて他者に知識を教える際，演繹的に言語で説明することに相当する．

本研究では，図 1 に示す *LIDE (Learning from Image and DEscription)* を提案する．LIDE は画像エンコーダ，テキストデコーダ，テキストエンコーダ，画像分類器によって構成される．画像エンコーダが出力する画像特徴量に基づいてデコーダがテキストを生成することで，LIDE はモデルの内部表現を自然言語によって説明できる．テキストエンコーダは生成したキャプションあるいは人が与えたキャプションをテキスト特徴量に変換し，画像分類器は画像とテキスト特徴量の双方に基づいた分類を行う．

評価実験により LIDE は既存手法を上回る画像分類精度を達成した．また，人の説明の入力により性能がさらに向上すること，生成した説明の品質が画像分類の正否と相関を持つことを確認した．本研究の貢献は，マルチモーダル特徴量を few-shot 画像分類問題で利用することで，1) 言語によって説明された概念の学習可能性，2) モデルの内部表現の言語による説明可能性をそれぞれ検証した点にある．

2　問題設定および関連研究

N-way K-shot 画像分類　本問題では，多数のクラスが含まれるデータセットが，クラスの重複が無いように $\mathcal{T}_{train}$，$\mathcal{T}_{dev}$，$\mathcal{T}_{test}$ に分割されている．各サブセットから，少数の N クラス，学習サンプル（サポート，各クラス K 個），評価サンプル（クエリ，各クラス M 個）をサンプリングしてタスクとする．この問題はメタ学習の一種であり，訓練時に $\mathcal{T}_{train}$ からサンプリングした複数個のタスクを用いてモデルを訓練する．その後，テスト時に新規の概念を $\mathcal{T}_{dev}$ あるいは $\mathcal{T}_{test}$ のサポートから学習し，クエリに対する性能を評価する．

代表的手法は Prototypical Network（ProtoNet） [4] である．ProtoNet は訓練・テスト時の個別タスクに対する学習をクラスプロトタイプの計算で代替する．h_k^c をクラス c の k 番目のサポートの特徴量，クラス c のプロトタイプを $z^c = \frac{1}{K}\sum_k W_{proto} h_k^c$ とする．訓練時は $\mathcal{T}_{train}$ のクエリに関する損失 L_{class} を最小化することで，モデルパラメータを更新する

$$L_{class} = -\frac{1}{KM}\sum_i \sum_c y_i^c \log \frac{\exp\left[s(z^c, h_i)\right]}{\sum_{c'} \exp\left[s(z^{c'}, h_i)\right]}.$$

W_{proto} は学習可能な重み，$y_i^c \in \{0,1\}$，h_i は i 番目のクエリの真のラベルおよび特徴量である．スコア関数は $s(z^c, h_i) = z^{c\top} W_{proto} h_i$ と定義する．テスト時

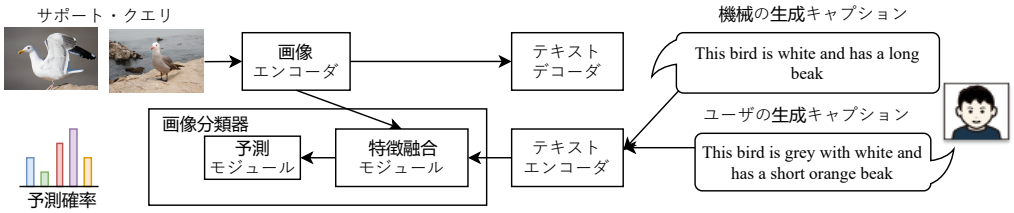


図 1 LIDE の全体像．画像エンコーダが出力する特徴量を基に，テキストデコーダがキャプションを生成する．テキストエンコーダは機械の生成キャプションとユーザの生成キャプションどちらかを受け付けテキスト特徴量に変換する．両モーダルの特徴量を特徴融合モジュールが統合し，予測モジュールが画像分類の予測確率を出力する．

は $\mathcal{T}_{dev,test}$ のサポートを用いてプロトタイプのみ更新することで個別のタスクを学習し，クエリを用いて個別のタスクにおける性能を評価する．

言語情報を用いた few-shot 画像分類モデル ProtoNet をベースに言語情報を用いて画像分類の性能を高める試みが取り組まれてきた．ProtoNet は図 1 における画像エンコーダと予測モジュールのみを用いるモデルに相当する．LSL [5] は，画像エンコーダの過学習を防ぐため，エンコーダの出力特徴量に基づいて画像キャプションを生成するテキストデコーダを導入し，生成テキストの損失を訓練の正則化項として利用した．RS-FSL [6] は，デコーダを GRU [7] から双方向 Transformer [8] に変更した．

L3 [9] は中間表現を自然言語によって表現することを目的としており，ProtoNet にテキストデコーダ・エンコーダを導入した．L3 は画像表現からデコードした自然言語記述をエンコードした表現のみを使って画像分類を行うことで，デコーダの出力を画像分類の根拠として解釈できる．しかし，説明性を備えた一方で精度が低下する問題が残った．

3 提案手法

LIDE の構成および訓練・テスト手法について述べる．なお，従来研究と同様に，未知クラスの画像分類に影響が出るため大量の画像データによる画像エンコーダ・分類器の事前学習は行わない．

3.1 モデル

画像エンコーダ 従来研究と同様に 4 層の CNN（事前学習無し）を用いる．画像を入力として受け取り，画像特徴ベクトルとして h_{img} を出力する．

テキストデコーダ 3 層の単方向 Transformer を用いる．まず，h_{img} をテキスト特徴空間に写像する

$$f_{I2T}(h_{img}) = \mathrm{Linear}(\mathrm{LayerNorm}(h_{img})).$$

LayerNorm は Layer Normalization [10] を示す．得られた表現をテキストデコーダに長さ 1 の系列として渡し，テキストデコーダでは自己回帰的に j 番目のトークン t_j を以下の確率 p_j に従い生成する

$$p_j = \Pr(t_j; f_{I2T}(h_{img}), t_{0:j-1}).$$

テキストエンコーダ BERT [11] を用いる．テキストを WordPiece [12] でトークナイズした系列を入力として受け取り，最終層の隠れ状態の系列 H_{BERT} を出力する．言語特徴ベクトル h_{text} を取得するため，テキストデコーダのトークン生成確率 p_j を用いた重み付きプーリングを行う

$$h_{text} = \frac{1}{\sum p_j w_j} \sum p_j w_j h_{BERT,j}.$$

トークンがストップワードの場合 $w_j = 0$，それ以外は 1 とする．また，デコーダの代わりに人がキャプションを与えた際は全トークンで $p_j = 1$ とする．

重み付きプーリングはテキストデコーダの確信度の低い出力の影響を抑えることで，生成文の品質の低さによる画像分類精度の低下を防ぐことを目的とする．さらに，テキスト生成の離散的操作のために通常はテキストエンコーダの勾配をテキストデコーダに渡すことができないが，重み付きプーリングでは画像分類から得られる損失を重み p_j を経由してテキストデコーダに伝播させることが可能となる．

特徴融合モジュール LIDE の画像分類器は特徴融合と予測モジュールから成る．特徴融合モジュールは，単一モーダル特徴のみを予測に用いる既存研究と異なり，h_{img} と h_{text} をマルチモーダル特徴 h_{mm} に融合する．f_{T2I} をテキスト特徴から画像特徴空間への写像として，h_{mm} を下記の様に求める

$$[w_{img}; w_{text}] = \mathrm{softmax}(\mathrm{Linear}([h_{img}; h_{text}])) \in \mathbb{R}^2,$$

$$h_{mm} = w_{img} h_{img} + w_{text} f_{T2I}(h_{text}).$$

$[;]$ はベクトル連結を表す．f_{T2I} には活性化関数を ReLU とした 3 層の FFNN を用いた．

予測モジュール ProtoNet [4] を用いる．特徴量 h はマルチモーダル特徴量 h_{mm} に置き換える．

3.2 訓練時およびテスト時のアルゴリズム

損失関数 画像分類のために 2 種類の損失関数を用いる．$L_{class,gold}$ は画像特徴量と真のテキストから計算した損失である．$L_{class,gen}$ は画像特徴量と生成したテキストから計算した損失である．テキスト生成の損失として，teacher-forcing とクロスエントロピー損失で計算した L_{text} を用いる．

また，多様な言語表現を獲得するために対照学習を行う．クラス c のサポートサンプルの真および生成キャプションの表現を v^c_{gold}，v^c_{gen} とする．cosine 類似度で計算した対照学習の損失を L_{cntr} とする

$$L_{cntr} = \frac{1}{2N}\sum_{c}\log\frac{\exp\left[\cos({v^c_{gold}}^\top v^c_{gen})/\tau\right]}{\sum_{c'}\exp\left[\cos({v^c_{gold}}^\top v^{c'}_{gen})/\tau\right]} + \frac{1}{2N}\sum_{c'}\log\frac{\exp\left[\cos({v^{c'}_{gold}}^\top v^{c'}_{gen})/\tau\right]}{\sum_{c}\exp\left[\cos({v^{c}_{gold}}^\top v^{c'}_{gen})/\tau\right]}.$$

以上を合計した L を損失関数とする

$$L = L_{class,gold} + L_{class,gen} + \lambda_{text}L_{text} + \lambda_{cntr}L_{cntr}.$$

τ，λ_{text}，λ_{cntr} はハイパーパラメータである．

事前学習 [6, 13] と同様，訓練データのみを用いた事前学習を行う．事前学習は few-shot ではない通常の画像分類として解く．損失は $L_{class,gold} + L_{text}$ である．

言語生成 訓練時は貪欲法とランダムサンプリングによってキャプションを生成する．ランダムサンプリング時は各時刻で top 20 の単語をサンプリングして生成する．テスト時，人からキャプションが与えられない場合はテキストデコーダの生成をテキストエンコーダに渡す．この際，length penalty を 0.5，ビーム幅を 5 としたビームサーチを行ってスコアが最大となる生成文をテキストエンコーダに渡す．

4 評価実験

4.1 実験設定

データセット Caltech-UCSD Birds (CUB) [14, 15] を 5-way 1-shot 分類問題として用いた．本データは 200 種の鳥の品種をクラスとしており，1 種に対して 40-60 枚の画像がある．200 種の画像は 100 種が訓練用，50 種が開発用，50 種がテスト用である，

比較手法 画像のみを用いる ProtoNet [4]，画像特徴量からデコードしたテキストの特徴量のみを基に分類を行う L3 [9]，テキスト情報を正則化項としてのみ用いる LSL [5]，RS-FSL [6] を利用した．

表 1 各モデルの分類精度．

Model	Accuracy	Img. Enc.	Text Enc.	Fusion	Text Dec.
ProtoNet	57.97 ±0.96	✓			
L3	53.96 ±1.06		✓		✓
LSL	61.24 ±0.96	✓			✓
RS-FSL	65.66 ±0.90	✓			✓
LIDE	**67.53** ±0.91	✓	✓	✓	✓

表 2 LIDE で導入した各手法と BERT の事前学習の効果．

LIDE	**67.53** ±0.91
対照学習なし	64.60 ±0.88
平均プーリングの重みなし	66.16 ±0.93
言語生成のランダムサンプリングなし	66.40 ±0.93
BERT の事前学習なし	66.42 ±0.94

評価指標（画像分類） 従来研究と同様に 5-way 1-shot 画像分類を 600 回行った際の平均分類精度を示す．また，画像表現 h_{img} とテキスト表現 h_{text} が分布する多様体について，局所内在次元 [16, 17] および主成分分析の累積寄与率が 90%を達成する次元数を測り，両モーダルの内部表現を分析する．

評価指標（説明可能性） デコーダが生成したキャプションについて BLEU_4 [18]，METEOR [19]，および ROUGE_L [20] を用いて正確さを評価する．さらに，生成キャプションが予測の根拠として機能しているか，キャプションの正確さと予測の正確さの相関を Spearman の順位相関係数を用いて評価する．評価指標の詳細については補足資料に示す．

4.2 言語情報を用いた画像分類の評価

LIDE は比較手法の画像分類精度を上回るか？ 表 1 に結果を示す．LIDE は全比較手法，特に言語表現を正則化項としてのみ利用する LSL や RS-FSL の性能を上回った．L3 と異なり，LIDE はテキスト表現の利用と画像分類性能の向上を両立した．

LIDE で導入した各手法は有効か？ 表 2 に LIDE の各手法を除いた結果を示す．提案した対照学習，重み付きプーリング，貪欲法とランダムサンプリングの組合せは全て性能向上に貢献した．

BERT の事前学習は有効か？ 表 2 に示すように，我々の予想と反し，LIDE は BERT の事前学習を除いた場合でも既存研究の分類精度（表 1）を上回った．CUB データセットは鳥の品種に限定されるため，訓練データのみでテキスト空間を十分に学習できるためと考える．より広範囲なドメインでの検証は今後の課題とする．

高品質のキャプションを人が与えることで予測は改善するか？ モデルが生成したキャプションの代わりに真のキャプションを入力する実験をした．

表 3　各キャプション入力における分類精度.

生成キャプションを入力	67.53 ±0.91
真のキャプションを入力	73.08 ±0.88

表 4　特徴空間の次元数.

	局所内在次元	主成分分析
画像表現	25.7	522
テキスト表現	3.71	19

CUB では 1 画像に複数の真のキャプションがあるため，生成キャプションに対する bi-gram precision が最も高いものをユーザの入力キャプションとした．表 3 に示す通り，機械が生成したキャプションを入力する場合に比べて，真のキャプションを入力することでさらに性能が改善した．

なぜテキスト表現は画像分類に貢献するのか？　画像・テキスト表現のそれぞれが分布する多様体について，局所内在次元と主成分分析の累積寄与率 90%を達成する次元数を調査した．表 4 に示す通り，テキスト表現は画像表現に比べて潜在的に小さな次元の多様体に分布していることがわかる．画像には多くの情報が含まれるが，キャプションは興味のある情報にのみ言及する．今回はキャプションが鳥の記述であるため，キャプションからは鳥の分類のために重要な情報のみを持つ特徴量を抽出できると考えられる．なお，局所内在次元が低いほど汎化された表現であり，few-shot 分類の性能が高くなるという報告 [21] があり，本研究の結果と一致する．

4.3　説明可能な機械学習としての評価

生成したキャプションは予測の説明として機能するか？　生成したキャプションが正確でなければ説明として不十分である．表 5 に生成と真のキャプションの文字列の重なりによる各指標値を示す．上限値は MSCOCO [22] を用いた教師あり事前学習を行なった CNN-LSTM モデルである [23]. LIDE は小規模な CUB データセットのみを訓練データに用いているが，上限値と比較しても METEOR で 2.0 ポイント，ROUGE で 3.3 ポイントの下落に留まっており，品質の高いキャプションが生成できている．

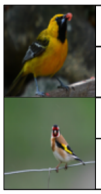
次に，生成したキャプションが正しい・誤っているときは分類結果も正しい・誤っていることがモデルの説明と分類結果の一貫性の観点からは望ましい．そこで，キャプションの評価値と画像分類の正否の関係を調査したところ，正の相関があった．LIDE からテキストエンコーダを除き画像表現のみで分類させた場合（LSL 相当）の相関値は低下することから，LIDE におけるテキストエンコーダおよび特徴融合モジュールの導入が説明可能性に大きく貢献していることが分かる．

表 5　生成キャプションの評価値と分類結果との相関.
†: $p < 0.1$, * : $p < 0.05$, ** : $p < 0.01$

	BLEU_4	METEOR	ROUGE_L
上限値: キャプション	59.0	36.1	69.7
LIDE: キャプション	48.1	34.1	66.4
相関値	$0.309^{\dagger}$	0.468^{*}	0.436^{**}
LIDE w/o Text Enc.: キャプション	50.0	34.6	67.2
相関値	0.114	0.201	0.217

なお本実験では，先行研究との比較のために画像エンコーダの構造と訓練用画像データを制限した．これらの制限を除くことで，説明可能な機械学習モデルとしての性能は向上すると考えられる．

キャプション生成の課題はなにか？　キャプションの出力例を図 2 に示す．モデルは鳥の特徴を捉えてキャプションを生成している一方で，構文が類似しており，言及する部位とその色だけが異なるような画一的なキャプションになっている．特に 2 枚目の画像は ‘red face’ のような珍しい特徴を持っているが，生成したキャプションは言及できていない．多様なキャプションの生成が今後の課題である．



Ours	this bird is yellow with black and has a very short beak.
Gold	the bird has a small black bill and a yellow crown
Ours	this bird has wings that are brown and has a white belly.
Gold	a small bird with a red face black crown white cheeks brown back and black and yellow wings

図 2　生成事例.

5　おわりに

本研究は，few-shot 画像分類問題を題材に，機械学習モデルが新規の概念を自然言語の説明から獲得し，またモデルの挙動を自然言語によって説明することに取り組んだ．本研究の貢献を以下に示す．

本研究の独自性．　提案手法の LIDE はテキストデコーダによる説明の生成と，テキストエンコーダによる説明の入力によって few-shot 画像分類の性能を改善する初めての研究である．

本研究の重要性．　自然言語の説明から概念を学習し，また概念を自然言語によって説明することは，人には可能であるが人工知能にはまだ難しい領域のひとつである．本研究は人と人工知能の共生社会の実現に向け，人と同じように学び説明する機械学習モデルの開発に貢献し得ると考える．

参考文献

[1] Sahil Chopra, Michael Henry Tessler, and Noah D Goodman. The first crank of the cultural ratchet: Learning and transmitting concepts through language. In **CogSci**, pp. 226–232, 2019.

[2] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. **arXiv preprint arXiv:2104.08773**, 2021.

[3] Hui Liu, Qingyu Yin, and William Yang Wang. Towards explainable nlp: A generative explanation framework for text classification. In **ACL**, pp. 5570–5581, 2019.

[4] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In **NIPS**, Vol. 30, pp. 4077–4087, 2017.

[5] Jesse Mu, Percy Liang, and Noah Goodman. Shaping visual representations with language for few-shot classification. In **ACL**, pp. 4823–4830, 2020.

[6] Mohamed Afham, Salman Khan, Muhammad Haris Khan, Muzammal Naseer, and Fahad Shahbaz Khan. Rich semantics improve few-shot learning. **arXiv preprint arXiv:2104.12709**, 2021.

[7] Kyunghyun Cho, Bart Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In **EMNLP**, pp. 1724–1734, 2014.

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

[9] Jacob Andreas, Dan Klein, and Sergey Levine. Learning with latent language. In **NAACL-HLT**, pp. 2166–2179, 2018.

[10] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. **arXiv preprint arXiv:1607.06450**, 2016.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **NAACL-HLT**, pp. 4171–4186, 2019.

[12] Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. In **ICASSP**, pp. 5149–5152, 2012.

[13] Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens van der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. **arXiv preprint arXiv:1911.04623**, 2019.

[14] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[15] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In **CVPR**, pp. 49–58, 2016.

[16] Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. In **NIPS**, Vol. 17, 2005.

[17] Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E. Houle, Ken-ichi Kawarabayashi, and Michael Nett. Estimating local intrinsic dimensionality. In **KDD**, p. 29–38, 2015.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **ACL**, pp. 311–318, 2002.

[19] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In **Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL**, pp. 65–72, 2005.

[20] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out@ACL**, pp. 74–81, 2004.

[21] Dong Hoon Lee and Sae-Young Chung. Unsupervised embedding adaptation via early-stage feature reconstruction for few-shot classification. In **ICML**, Vol. 139, pp. 6098–6108, 2021.

[22] TY Lin, M Maire, S Belongie, J Hays, P Perona, D Ramanan, P Dollar, and CL Zitnick. Microsoft coco: Common objects in context. In **ECCV**, pp. 740–755, 2014.

[23] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. Show, adapt and tell: Adversarial training of cross-domain image captioner. In **ICCV**, pp. 521–530, 2017.

[24] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In **ICLR**, 2017.

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.

[26] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In **Autodiff@NIPS**, 2017.

[27] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In **ACL: System Demonstrations**, pp. 38–45, 2020.

[28] Steven Bird, Ewan Klein, and Edward Loper. **Natural Language Processing with Python**. O'Reilly Media, Inc., 2009.

[29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **JMLR**, Vol. 21, pp. 1–67, 2020.

A　Few-shot 画像分類の詳細

N-way K-shot 画像分類は，データから N クラス・K サポート，M クエリをサンプリングすることで 1 タスクを作成する．N-way K-shot 画像分類の訓練時は，複数の N クラス分類タスクをバッチサイズ個独立にサンプリングして同時に学習する episodic training [24] を行う．テスト時では複数の N クラス分類タスクにおける平均性能を評価する．

B　実験設定の詳細

B.1　実装

実験には NVIDIA Quadro RTX 8000 (48GB) GPU 1 枚を用いた．ハイパーパラメータを表 6 に示す．最適化手法には Adam [25] を用いた．実装には PyTorch [26] と Transformers [27] を用いた．ストップワードの定義には NLTK [28] を用い，‘bird’ を追加した．画像の前処理及び画像エンコーダの構成は先行研究 [5] と同一である．トークナイザとテキストエンコーダは BERT-base-uncased を用いた．テキストデコーダは事前学習済みエンコーダデコーダである T5-base [29] のデコーダと同じ構成としたが，事前学習済みパラメータは用いなかった．

表 6　ハイパーパラメータ．

	Pre-Training	Fine-Tuning
batch size	128	100
epochs	100	1500
learning rate for main model	1e-3	1e-3
learning rate for text encoder	1e-3	1e-4
learning rate for text decoder	1e-5	1e-5
λ_{text}	—	10
λ_{cntr}	—	0.1
τ	—	0.05

B.2　局所内在次元の推定

表 4 に示した局所内在次元は，データ点 x の近傍でのデータ多様体の次元を示す．局所内在次元は最尤法を用いて

$$\hat{\mathrm{LID}}(x) = -\left\{\frac{1}{n_{nn}}\sum_{i=1}^{n_{nn}}\log\left(\frac{r_i(x)}{r_{n_{nn}}(x)}\right)\right\}^{-1}$$

と推定できる．ここで，n_{nn} は近傍数であり，既存研究同様に 20 と設定した．r_i は点 x とその第 i 近傍点のユークリッド距離である．評価実験においては，テストデータ点における局所内在次元の平均を推定値とした．

B.3　一貫性の評価

表 5 では生成したキャプションの正確さと画像分類の予測の正確さの相関を調査するため，Spearman の順位相関係数を用いた．テストデータ 2953 枚に対してキャプション生成を行い，それぞれの指標でキャプションを評価した．キャプション評価値の昇順でデータを 30 のビンに分割し，ビンごとに画像分類の平均精度を求めることで，相関を調べた．