DueT: 視覚・言語の Dual-adapter Tuning による基盤モデル

西田京介 [1*]　長谷川拓 [1*]　前田航希 [2†]　齋藤邦子 [1]
[1] 日本電信電話株式会社　NTT 人間情報研究所
[2] 東京工業大学
{kyosuke.nishida.rx, taku.hasegawa.ps}@hco.ntt.co.jp

概要

対照学習により構築する視覚・言語の基盤モデル CLIP の新たな転移学習方法として DueT を提案する．DueT は単一モーダルのコーパスで事前学習されたモデルにより画像・テキストエンコーダを初期化して固定し，両エンコーダに追加したゲート機構付のアダプタのみを学習する．英語・日本語ドメインの 0-shot 画像・テキスト検索において，単純な fine-tuning や画像エンコーダのみ転移・固定する従来手法に比べ，提案手法が精度やパラメータ効率性の観点で優れていたことを報告する．

1　はじめに

CLIP [1] が視覚と言語の融合理解における基盤モデルとして，テキストからの画像生成 [2, 3] や視覚情報を考慮した対話 [4] を始め様々なタスクで革新的な成果を挙げている．CLIP は 4 億件という膨大な画像・テキストのペアを用いた対照学習により，正しい（誤った）ペアに対して画像・テキストの各エンコーダが出力する特徴ベクトルの類似度が高く（低く）なるように scratch から学習された．ここで，CLIP よりも視覚と言語の意味的対応付けに優れた基盤モデルを少量の学習データから構築するためには，比較的学習しやすい単一モーダルの事前学習済モデルの転移学習が重要になると考える．

本研究では視覚・言語の基盤モデルの学習方法 **DueT**（Dual-adapter Tuning）を提案する．DueT は単一モーダルの事前学習済モデルを各エンコーダのパラメータの固定値とし，さらに両エンコーダにゲート機構を持つアダプタを追加して学習を行う（図 1 右）．英語・日本語ドメインの実験において，単純な転移学習や画像エンコーダのみ転移・固定する LiT [5] に比べて，DueT は優れた性能を達成できた．

* Equal contribution.
† NTT におけるインターンシップ期間中の貢献．

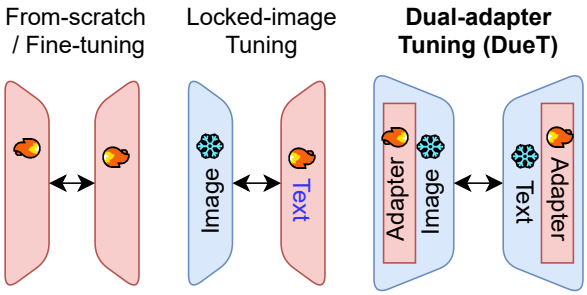

図 1　視覚・言語の基盤モデルの学習方法．左: 両エンコーダを学習（🔥）．**中央**: LiT (Locked-image Tuning) [5]. 事前学習済みの画像エンコーダを固定（❄）し，テキストエンコーダのみ学習する．**右**: DueT (Dual-adapter Tuning). 両エンコーダに追加したアダプタのみ学習する．

2　関連研究

2.1　視覚・言語の基盤モデル

視覚・言語の基盤モデルとして CLIP [1] および ALIGN [6] が提案されて以降，性能改善に向けて様々な観点から研究が行われている．

転移学習　全てのパラメータをランダムに初期化して学習する from-scratch [1, 6, 7, 8, 9, 10] と，画像・テキストエンコーダを各モーダルの事前学習済モデルで初期化する fine-tuning [11, 12, 13] のいずれかが多く採用されている（図 1 左）．最近では，画像エンコーダのみを事前学習済モデルで初期化・固定し，テキストエンコーダのみを学習する Locked-image Tuning（LiT） [5] が提案されている（図 1 中央）．

クロスアテンション　CLIP はクロスアテンション機構を持たないため，両モーダルを入力とするクロスエンコーダ [11, 12, 9]，クロスアテンションを行うテキストエンコーダ [13]・テキストデコーダ [13, 8] などの追加に関する検討が行われている．

目的関数　masked 言語モデリング [11, 12, 10], causal 言語モデリング [8, 13]，masked 画像モデリング [12]，画像・テキストマッチング [11, 12, 13] など

様々な目的関数の追加が検討されている．また，対照学習自体もトークンレベルでの類似度 [7]，難解な負例の追加 [10, 14] による改善が検討されている．

2.2 パラメータ効率的な転移学習

事前学習済の言語モデルを効率的に下流タスクに適応させるための技術が盛んに研究されている．代表的な手法には adapter tuning [15, 16, 17, 18, 19]，prefix tuning [20, 21]，additive methods [22, 23, 24]，sparse-finetuning [25, 26] などがある．また，これらを統一的に扱うアプローチも提案されている [27, 28]．

特に adapter tuning は Transformer の層間に，アダプタと呼ばれる小さな追加モジュール（一般的には 2 層のフィードフォワードネットワーク; FFN）を残差接続付きで挿入し，アダプタのみを学習する．

2.3 本研究の位置付け

adapter tuning [15] のアイデアを下流タスクではなく CLIP の事前学習に導入した初めての研究である．単一モーダルの事前学習の忘却を防ぎ高精度かつパラメータ効率的な転移学習の実現を狙う．転移学習を除く前記した CLIP の拡張研究は扱わないが，これらは提案手法と組み合わせた利用が可能である．

3 提案手法

基盤モデルの学習方法として **DueT**（Dual-adapter Tuning）を提案する．モデルは画像・テキストの Transformer [29] エンコーダから構成される．事前学習済モデル（ViT [30] や BERT [31] など）で各エンコーダの初期化を行い，Transformer ブロックに追加した Gated Adapter Unit（GAU）のみを学習する．

入出力　画像（テキスト）エンコーダは，パッチ（トークン）の系列を入力として受け取り，d 次元ベクトルの系列 $\boldsymbol{H} = [\boldsymbol{h}_{\mathrm{CLS}}, \boldsymbol{h}_1, ..., \boldsymbol{h}_{\mathrm{SEP}}]$ を出力する．各エンコーダの出力 $\boldsymbol{h}_{\mathrm{CLS}} \in \mathbb{R}^d$ を学習可能パラメータによる線形変換および L_2 正規化にて d_m 次元の特徴ベクトル $\boldsymbol{x}, \boldsymbol{y}$ にそれぞれ射影し，最終的に内積 $\boldsymbol{x}^\top \boldsymbol{y}$ によって画像・テキストの類似度を求める．

GAU　一般的なアダプタを [15] を学習可能なゲート係数 α によりアダプタの入出力を混合する GAU に拡張する．各エンコーダの全ての Transformer ブロック（$l = 1...L$）に式 (1) の GAU を挿入する．

$$\mathrm{GAU}^l(\boldsymbol{H}^l) = \alpha^l \mathrm{FFN}^l(\mathrm{LN}(\boldsymbol{H}^l)) + (1-\alpha^l)\boldsymbol{H}^l, \quad (1)$$

$$\mathrm{FFN}^l(\boldsymbol{h}) = \phi(\boldsymbol{h}\boldsymbol{W}^l_{\mathrm{down}} + \boldsymbol{b}^l_{\mathrm{down}})\boldsymbol{W}^l_{\mathrm{up}} + \boldsymbol{b}^l_{\mathrm{up}}, \quad (2)$$

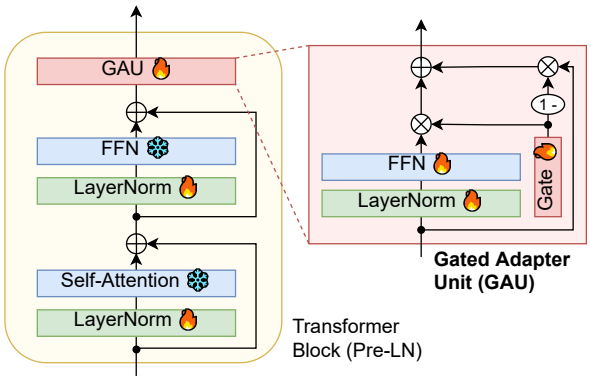

図 2　Gated Adapter Unit（GAU）．左: 各 Transformer ブロックに挿入した GAU および LN を学習し（🔥），FFN や Self-Attention のパラメータは更新しない（❄）．**右**: GAU は 2 層の FFN および学習可能なゲート係数 α を持つ．

ここで，入力 $\boldsymbol{H}^l$ は Transformer の FFN モジュールの残差接続後の出力である．LN は layer normalization [32] を表す[1)]．$\boldsymbol{W}^l_{\mathrm{down}} \in \mathbb{R}^{d\times m}$，$\boldsymbol{b}^l_{\mathrm{down}} \in \mathbb{R}^m$，$\boldsymbol{W}^l_{\mathrm{up}} \in \mathbb{R}^{m\times d}$，$\boldsymbol{b}^l_{\mathrm{up}} \in \mathbb{R}^d$，$\alpha^l \in \mathbb{R}$ は GAU 毎に独立した学習可能パラメータである．入出力・中間層の次元を d，m とする．活性化関数 ϕ は GeLU [33] である．

図 2 に GAU の例を示す．Transformer 本体は LN のパラメータを除いて固定し，GAU のみを学習する．なお，全てのゲート係数が 0 のとき，エンコーダは事前学習済モデルに等しくなる．

対照学習　ミニバッチ内に同じ画像やテキストが含まれた際のノイズを軽減するため，UniCL [34] のアイデアに基づく損失関数を用いる．画像・テキストのハッシュ値（s, t）のいずれかが同じ値となるペアを正例 $\mathbf{P}$，他のランダムに形成されるペアを負例 $\mathbf{B}$ として，以下の損失関数の合計を最小化する．

$$L_{\mathrm{i2t}} = -\frac{1}{|\mathbf{B}|}\sum_{i\in\mathbf{B}}\frac{1}{|\mathbf{P}_i|}\sum_{k\in\mathbf{P}_i}\log\frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{y}_k/\tau)}{\sum_{j\in\mathbf{B}}\exp(\boldsymbol{x}_i^\top \boldsymbol{y}_j/\tau)}, \quad (3)$$

$$L_{\mathrm{t2i}} = -\frac{1}{|\mathbf{B}|}\sum_{i\in\mathbf{B}}\frac{1}{|\mathbf{P}_i|}\sum_{k\in\mathbf{P}_i}\log\frac{\exp(\boldsymbol{y}_i^\top \boldsymbol{x}_k/\tau)}{\sum_{j\in\mathbf{B}}\exp(\boldsymbol{y}_i^\top \boldsymbol{x}_j/\tau)}. \quad (4)$$

ここで，τ は温度パラメータである．

学習効率性　画像・テキストエンコーダの層数 L と次元数 d が同じとき，DueT の学習対象パラメータ数（#TP）はおよそ $4Ldm$ となる．エンコーダとして ViT-B/16 と $\mathrm{BERT}_{\mathrm{base}}$ ($L = 12$, $d = 768$，合計パラメータ数 194.7M) を用いた場合，$m = 1536$ とした時の #TP は全体の約 29.1%（56.8M）となる．

1) サブモジュールの残差接続の内に LN を用いる pre-LN 型の Transformer（例:ViT [30]; 図 2）に GAU を挿入する際は，式 (1) のように FFN の前に LN を挿入する．post-LN 型（例:BERT [31]）の場合は FFN の後に LN を挿入する．

4 評価実験

英語・日本語の CLIP を構築して評価を行った．

4.1 事前学習済モデル

画像エンコーダ（IE） ImageNet-21k [35] から学習された ViT-B/16 [36, 37] を用いた．

テキストエンコーダ（TE） 英語モデルには BERT-base-uncased [31, 38] を用いた．日本語モデルには Wikipedia および CC-100-ja [39] から [31] の設定で学習した BERT-base-uncased を用いた．

4.2 訓練データセット

YFCC-CLIP Radford ら [1] により整備された英語の自然言語により記述されたタイトルあるいは説明を持つ約 15M 件の Flickr 画像（YFCC100M [40] のサブセット）を用いる．なお，ImageNet-21k には Flickr 画像が多く含まれているため YFCC-CLIP は IE の学習ドメインに近いデータセットと言える．

JWeb5M 我々が独自に広範囲の Web サイトから収集した 5M 件の画像・日本語テキストペアである．100 件のサンプリング調査を実施したところ，32 件の画像に日本語文字が含まれているなど IE の学習ドメインからは遠いデータセットと言える．単語数の中央値は 9 語であり，自然文ではない数単語のみのテキストも含まれる．詳細は付録 A に示す．

4.3 評価データセット

英語 / 日本語の 0-shot（下流タスクでの fine-tuning を行わない）の画像・テキスト検索タスクにて，訓練データにて学習した基盤モデルを評価する．

COCO [41] / STAIR Captions [42] 5000 件のテスト画像（[43] の分割）．各画像に 5 件の自然文キャプション．それぞれの単語数の中央値は 11 / 12 語．

Flickr30k [44] / F30kEnt-JP [45] 1000 件のテスト画像（[43] の分割）．各画像に 5 件の自然文キャプション．それぞれの単語数の中央値は 12 / 17 語．

上記セットの画像は Flickr 由来であるため，IE の事前学習と YFCC-CLIP に近いドメインと言える．詳細は付録 B に示す．また，0-shot の評価とは異なるが，YFCC-CLIP・JWeb5M の訓練セットから開発・テスト用に 10,000 ペアを分割して利用した．

評価指標 [11] と同様に，クエリに対する k 件の検索結果に正しい事例が含まれる割合である Recall@k（$k = 1, 5, 10$）の平均値を報告する．

表 1 英語モデルのテキスト・画像検索（I→T，T→I）性能．#TP は学習対象パラメータ数．†0-shot 検索タスク．

Method	#TP	YFCC-CLIP I→T	YFCC-CLIP T→I	COCO† I→T	COCO† T→I	Flickr30k† I→T	Flickr30k† T→I
FS	194.7	78.77	78.48	43.31	26.79	64.36	42.53
FT	194.7	**88.70**	**88.21**	60.64	40.91	83.07	62.43
LiT	108.9	67.30	66.29	52.64	35.07	75.61	55.21
LiT-FT	108.9	71.16	69.81	57.07	37.64	78.83	57.24
DueT	113.4	87.74	87.18	**60.93**	**41.69**	**83.61**	**63.90**

表 2 日本語モデルの検索性能．†0-shot 検索タスク．

Method	#TP	JWeb5M I→T	JWeb5M T→I	STAIR† I→T	STAIR† T→I	F30kEnt-JP† I→T	F30kEnt-JP† T→I
FS	194.7	50.38	50.67	33.16	23.63	40.82	30.15
FT	194.7	72.07	72.77	60.44	50.23	77.87	64.64
LiT	108.9	50.45	48.60	47.90	32.82	65.72	50.49
LiT-FT	108.9	54.58	53.02	52.21	35.87	71.53	54.30
DueT	113.4	**74.44**	**74.21**	**62.08**	**52.74**	**81.62**	**70.54**

4.4 実験設定

提案手法 $m = 3072$ とした GAU を両エンコーダの全層に挿入した．学習対象パラメータ数（#TP）は 113.4M であった．特徴ベクトル $\boldsymbol{x}$，$\boldsymbol{y}$ の次元を $d_m = 512$ とし，バッチサイズ 8192 で 16 エポック学習した．Ablation テストでは $m = 1536$ とした（#TP = 56.8M）．その他の詳細は付録 C に示す．

ベースライン エンコーダ（IE・TE）の初期化およびパラメータ固定について 4 つの手法を用意した．その他の設定は提案手法と揃えた．

- **from-scratch（FS）** IE・TE の両方をランダムに初期化して学習する（194.7M）．
- **fine-tuning（FT）** IE・TE の両方を事前学習済モデルで初期化して学習する（194.7M）．
- **LiT** [5] IE を事前学習済モデルで初期化および固定．TE はランダム初期化して学習（108.9M）．
- **LiT-FT** IE・TE の両方を事前学習済モデルで初期化し，TE のみを学習する（108.9M）．

4.5 実験結果

表 1・2 に英語・日本語モデルの評価結果を示す．提案手法 DueT は 0-shot 画像・テキスト検索においてベースラインの評価スコアを上回った．特に，F30kEnt-JP の画像検索においては，学習パラメータ数を約 58%に抑えつつ fine-tuning（from-scratch）に比べてスコアを 5.9（40.4）ポイント改善した．

ベースラインの中では fine-tuning が優れていた．

表 3 GAU の学習パラメータ数による検索性能の変化．水色セルは fine-tuning（FT）と同程度の性能を表す．

		JWeb5M		STAIR		F30kEnt-JP	
m	#TP	I→T	T→I	I→T	T→I	I→T	T→I
96	3.7	64.63	64.19	57.26	47.95	77.83	65.72
192	7.2	67.50	66.95	59.08	49.75	80.54	67.53
384	14.3	70.41	69.92	60.37	50.57	80.48	69.18
768	28.4	72.57	72.07	61.85	52.25	**82.26**	70.03
1536	56.8	73.25	73.30	61.91	52.38	81.07	69.38
3072	113.4	**74.44**	**74.21**	**62.08**	**52.74**	81.62	**70.54**
FT	194.7	72.07	72.77	60.44	50.23	77.87	64.64

表 4 GAU を挿入する層の範囲による検索性能の変化．

		JWeb5M		STAIR		F30kEnt-JP	
Image	Text	I→T	T→I	I→T	T→I	I→T	T→I
1-12	N/A	63.95	64.19	52.77	45.16	69.90	62.77
1-12	8-12	71.99	71.71	60.57	50.85	79.41	69.47
1-12	4-12	73.14	72.99	61.62	52.07	81.06	**69.38**
N/A	1-12	66.07	64.95	58.50	47.36	76.31	62.16
8-12	1-12	71.58	71.29	59.90	50.41	77.46	65.28
4-12	1-12	**73.32**	72.89	61.65	51.87	79.78	69.26
1-12	1-12	73.25	**73.30**	**61.91**	**52.38**	**81.07**	**69.38**

YFCC-CLIP のテストセットでは最高スコアを達成したが，訓練データに過学習し易く 0-shot 検索の評価セットでは英語・日本語共に全て提案手法のスコアを下回った．特に，画像エンコーダの事前学習と訓練セットのドメインが異なる日本語モデルの評価では提案手法のスコアを大きく下回った．

LiT は事前学習済の画像エンコーダがカバーしていないドメインへの適応能力が低く，日本語モデルの評価において大きくスコアを落とした．また，転移学習を行わない from-scratch の場合は，5-15M 程度の学習データ数では不足していることが分かった．

4.6 Ablation テスト

学習データに JWeb5M を用いて，本研究の主な貢献である GAU の導入に関する評価を行った．

パラメータ効率的な学習が可能か？ 表 3 に示す通り，DueT は STAIR（F30kEnt-JP）では 14.3M（3.7M）個のパラメータ更新により，194.7M 個を更新する fine-tuning と同等の性能が得られており，パラメータ効率的な学習が実現できた．

全ての層にアダプタは必要か？ 表 4 に示す様に GAU を挿入する Transformer ブロック数を減らした場合の性能は低下した．画像・テキストエンコーダの両方の適応が CLIP で用いられる対照学習において重要であることが示唆された．

表 5 ゲート機構の違いによる検索性能の変化．

	JWeb5M		STAIR		F30kEnt-JP	
Gate	I→T	T→I	I→T	T→I	I→T	T→I
N/A	72.68	72.46	59.88	49.81	75.90	67.16
$\text{FFN}_{\text{token}}$	**74.02**	**73.97**	61.65	52.10	79.53	**70.09**
FFN_{sent}	73.50	73.40	61.15	52.32	80.74	69.17
scalar	73.25	73.30	**61.91**	**52.38**	**81.07**	69.38

表 6 ゲート係数の初期値と学習による検索性能の変化．

		JWeb5M		STAIR		F30kEnt-JP	
α_{init}	fixed	I→T	T→I	I→T	T→I	I→T	T→I
1.0		72.48	72.22	59.90	50.57	77.37	68.89
0.02	✓	68.28	67.64	59.57	50.72	79.14	67.15
0.02		**73.25**	**73.30**	**61.91**	**52.38**	**81.07**	**69.38**

アダプタのゲート係数は有効か？ 表 5 にゲートを用いない場合（$\alpha = 1.0$ 固定）および 1 層の FFN によりトークンあるいは文レベルで入力に適応的なゲート機構（詳細は付録 D に示す）との比較結果を示す．まず，ゲート係数の有無の比較により，ゲート係数を導入することの有効性を確認できた．一方で，FFN による適応的なゲート機構とゲート係数の間に明確な性能差は確認されなかった．

ゲート係数の初期値および学習の影響はあるか？ 表 6 に示す通り，ゲート係数 α を学習により更新することで性能が改善された．また，初期値を大きく設定すると，学習初期から未学習の GAU の影響が強くなるため転移学習が進みにくい問題があった．

5 おわりに

対照学習により構築する視覚・言語の基盤モデル CLIP [1] の事前学習に adapter tuning [15] による転移学習を導入した DueT を提案した．単一モーダルの事前学習済モデルを各エンコーダのパラメータの固定値とし，両エンコーダに [15] から拡張したゲート機構付のアダプタ GAU を追加して学習を行う．

本研究の貢献 膨大な学習データを要する CLIP において，単一モーダルの事前学習済モデルの転移学習は重要なテーマである．fine-tuning や LiT [5] よりも精度およびパラメータ効率性の観点で優れた手法を提案できたことは，基盤モデルの研究に関する大きな貢献と言える．そして，評価実験を通じて日本語 CLIP の構築における転移学習に要する追加パラメータ数などの知見が得られた．本研究の成果は，視覚と言語の融合的な理解を必要とする対話やナビゲーション，コンテンツ生成・検索など産業上重要なサービスの発展に貢献できる．

参考文献

[1] Alec Radford, et al. Learning transferable visual models from natural language supervision. In **ICML**, pp. 8748–8763, 2021.

[2] Aditya Ramesh, et al. Hierarchical text-conditional image generation with CLIP latents. **arXiv:2204.06125**, 2022.

[3] Robin Rombach, et al. High-resolution image synthesis with latent diffusion models. In **CVPR**, pp. 10674–10685, 2022.

[4] Jean-Baptiste Alayrac, et al. Flamingo: a visual language model for few-shot learning. **arXiv:2204.14198**, 2022.

[5] Xiaohua Zhai, et al. LiT: Zero-shot transfer with locked-image text tuning. In **CVPR**, pp. 18123–18133, 2022.

[6] Chao Jia, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In **ICML**, pp. 4904–4916, 2021.

[7] Lewei Yao, et al. FILIP: fine-grained interactive language-image pre-training. In **ICLR**, 2022.

[8] Jiahui Yu, et al. CoCa: Contrastive captioners are image-text foundation models. **arxiv:2205.01917**, 2022.

[9] Lu Yuan, et al. Florence: A new foundation model for computer vision. **arxiv:2111.11432**, 2021.

[10] Yangguang Li, et al. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In **ICLR**, 2022.

[11] Junnan Li, et al. Align before fuse: Vision and language representation learning with momentum distillation. In **NeurIPS**, pp. 9694–9705, 2021.

[12] Amanpreet Singh, et al. FLAVA: A foundational language and vision alignment model. In **CVPR**, pp. 15638–15650, 2022.

[13] Junnan Li, et al. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In **ICML**, pp. 12888–12900, 2022.

[14] Manling Li, et al. Clip-event: Connecting text and images with event structures. In **CVPR**, pp. 16420–16429, 2022.

[15] Neil Houlsby, et al. Parameter-efficient transfer learning for NLP. In **ICML**, pp. 2790–2799, 2019.

[16] Jonas Pfeiffer, et al. Adapterhub: A framework for adapting transformers. In **EMNLP**, pp. 46–54, 2020.

[17] Jonas Pfeiffer, et al. Adapterfusion: Non-destructive task composition for transfer learning. In **EACL**, pp. 487–503, 2021.

[18] Ruidan He, et al. On the effectiveness of adapter-based tuning for pretrained language model adaptation. In **ACL/IJCNLP**, pp. 2208–2222, 2021.

[19] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. In **NeurIPS**, pp. 1022–1035, 2021.

[20] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In **ACL-IJCNLP**, pp. 4582–4597, 2021.

[21] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In **EMNLP**, pp. 3045–3059, 2021.

[22] Demi Guo, Alexander M. Rush, and Yoon Kim. Parameter-efficient transfer learning with diff pruning. In **ACL-IJCNLP**, pp. 4884–4896, 2021.

[23] Edward J. Hu, et al. LoRA: Low-rank adaptation of large language models. In **ICLR**, 2022.

[24] Jeffrey O. Zhang, et al. Side-tuning: A baseline for network adaptation via additive side networks. In **ECCV**, pp. 698–714, 2020.

[25] Yi-Lin Sung, Varun Nair, and Colin Raffel. Training neural networks with fixed sparse masks. In **NeurIPS**, pp. 24193–24205, 2021.

[26] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In **ACL**, pp. 1–9, 2022.

[27] Yuning Mao, et al. Unipelt: A unified framework for parameter-efficient language model tuning. In **ACL**, pp. 6253–6264, 2022.

[28] Junxian He, et al. Towards a unified view of parameter-efficient transfer learning. In **ICLR**, 2022.

[29] Ashish Vaswani, et al. Attention is all you need. In **NIPS**, pp. 5998–6008, 2017.

[30] Alexey Dosovitskiy, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In **ICLR**, 2021.

[31] Jacob Devlin, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In **NAACL-HLT**, pp. 4171–4186, 2019.

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

[33] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). **arxiv:1606.08415**, 2016.

[34] Jianwei Yang, et al. Unified contrastive learning in image-text-label space. In **CVPR**, pp. 19163–19173, 2022.

[35] Jia Deng, et al. ImageNet: A large-scale hierarchical image database. In **CVPR**, pp. 248–255, 2009.

[36] Andreas Steiner, et al. How to train your ViT? data, augmentation, and regularization in vision transformers. **arxiv:2106.10270**, 2021.

[37] Andreas Steiner. Vision transformer and MLP-mixer architectures, 2021. https://github.com/google-research/vision_transformer.

[38] Google. BERT, 2018. https://github.com/google-research/bert.

[39] Alexis Conneau, et al. Unsupervised cross-lingual representation learning at scale. In **ACL**, pp. 8440–8451, 2020.

[40] Bart Thomee, et al. YFCC100M: the new data in multimedia research. **Commun. ACM**, Vol. 59, No. 2, pp. 64–73, 2016.

[41] Tsung-Yi Lin, et al. Microsoft COCO: common objects in context. In **ECCV**, pp. 740–755, 2014.

[42] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR captions: Constructing a large-scale japanese image caption dataset. In **ACL**, pp. 417–421, 2017.

[43] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. **IEEE Trans. Pattern Anal. Mach. Intell.**, Vol. 39, No. 4, pp. 664–676, 2017.

[44] Bryan A. Plummer, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. **Int. J. Comput. Vis.**, Vol. 123, No. 1, pp. 74–93, 2017.

[45] Hideki Nakayama, Akihiro Tamura, and Takashi Ninomiya. A visually-grounded parallel corpus with phrase-to-region linking. In **LREC**, pp. 4204–4210, 2020.

[46] Armand Joulin, et al. Bag of tricks for efficient text classification. In **EACL**, pp. 427–431, 2017.

[47] Paulius Micikevicius, et al. Mixed precision training. In **ICLR**, 2018.

[48] Tianqi Chen, et al. Training deep nets with sublinear memory cost. **arXiv:1604.06174**, 2016.

[49] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **ICLR**, 2019.

[50] Christian Szegedy, et al. Going deeper with convolutions. In **CVPR**, pp. 1–9, 2015.

[51] Samuel G. Müller and Frank Hutter. TrivialAugment: Tuning-free yet state-of-the-art data augmentation. In **ICCV**, pp. 754–762, 2021.

A JWeb5M

JWeb5M は本研究のために我々が広範囲な Web サイトから収集した画像・日本語キャプションを基に構築したデータセットである．ユニーク画像数 4,942,737（md5 ハッシュ値による確認），キャプション数は 4,369,144 となる．画像は短辺のサイズが 256 以上になるようにダウンロード時にリサイズした．言語判定には学習済の fastText [46] モデル [2] を利用した．キャプションを mecab-unidic により単語分割した際の単語数の分布を図 3 に示す．文字数および単語数の平均（中央）値は 26.3（20.0）文字，11.9（9.0）単語であった．自然文ではなく 1〜2 単語のキャプションも含まれる．また，100 画像をサンプリングして調査した際，日本語の文字（漢字，ひらがな，カタカナ）を含む画像は 32 枚であった．

B 評価データセット

COCO [41] は 123K 件の画像を含み，各画像に 5 件のキャプションが付与されている．画像は Flickr から収集され，クラウドワーカにより自然文のキャプションが作成された．nltk-punct による単語数の平均（中央）値は 11.3（11）単語であった．日本語版である STAIR Captions [42] は，COCO と同じ画像に対してクラウドワーカによって新たに各画像 5 件の日本語キャプションが作成された．mecab-unidic による単語数の平均（中央）値は 12.5（12）単語であった．

Flickr30K は 31K 件の画像を含み，各画像に 5 件のキャプションが付与されている．単語数の平均（中央）値は 31.4（12）単語であった．日本語版である F30kEnt-JP [45] は，Flickr30k の各キャプションを専門家が翻訳してキャプションが作成された．単語数の平均（中央）値は 18.4（17）単語であった．COCO および Flickr30k については，従来研究と同様に，[43] におけるデータセットの分割に従ってテストデータの画像を選択した．

C 実験設定

学習には 8 枚の NVIDIA A100 80GB GPU を用いた．バッチサイズを 8192 とし，mixed precision training [47] および gradient checkpointing [48] を用いて 16 エポック学習した．オプティマイザには

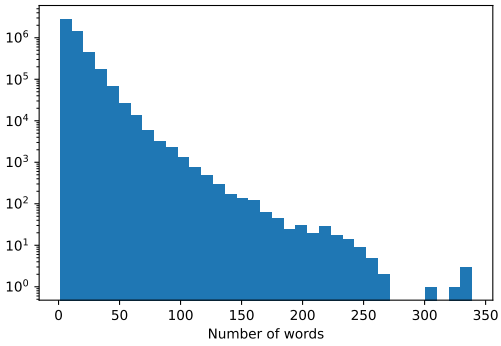

図 3 JWeb5M の各キャプションの単語数の分布．横軸：単語数（bin 幅 10），縦軸: 頻度（対数）

AdamW [49] を用い，学習率は 5e-4 とした．温度パラメータ τ は 0.015625（1/64）で固定とした．1 エポック毎にチェックポイントを保存し，開発セットの評価値（画像検索・テキスト検索それぞれの R@1，R@5，R@10 の平均値）の上位 3 つのチェックポイントのモデルに対するテストセットの評価値の平均を実験結果として報告した．

YFCC-CLIP の学習時は Inception-style random cropping [50] を用い，解像度は 224×224 とした．JWeb5M の学習時は上記の設定からクロップ範囲のスケールの下限のみ 0.9 に変更した．また，共通の augmentation 手法として TrivialAugment Wide [51] を採用し，画像の正規化は CLIP [1] と同様に行った．テキストエンコーダへの入力は最大 77 トークンとした．テスト時は，解像度 224×224 へのリサイズ・中央からのクロップと正規化のみ行った．テキストエンコーダにおけるプロンプトテキストは学習・テスト時のいずれも利用していない．

D 適応的ゲート機構

表 5 の実験では，式 (1) のゲート係数 α^l を 1 層の FFN に変更して，文・トークンレベルの適応的なゲート機構の評価を行った．$\mathrm{GAU}^l(\boldsymbol{H}^l)$ を

$$\mathrm{FFN}^l_\alpha(\boldsymbol{H}^l)\mathrm{FFN}^l(\mathrm{LN}(\boldsymbol{H}^l)) + (1 - \mathrm{FFN}^l_\alpha(\boldsymbol{H}^l))\boldsymbol{H}^l,$$

としたとき，

$$\mathrm{FFN}^l_{\alpha,\mathrm{sent}}(\boldsymbol{H}^l) = \sigma(\boldsymbol{h}^l_{\mathrm{CLS}}{}^\top \boldsymbol{w}^l_{\mathrm{gate}} + b^l_{\mathrm{gate}}) \in \mathbb{R} \quad (5)$$

$$\mathrm{FFN}^l_{\alpha,\mathrm{token}}(\boldsymbol{H}^l) = \sigma(\boldsymbol{H}^l \boldsymbol{w}^l_{\mathrm{gate}} + b^l_{\mathrm{gate}}) \in \mathbb{R}^n \quad (6)$$

と定義した．σ はシグモイド関数である．$\boldsymbol{w}^l_{\mathrm{gate}} \in \mathbb{R}^d$, $b^l_{\mathrm{gate}} \in \mathbb{R}$ は学習可能パラメータ，n は系列の長さである．

2） https://fasttext.cc/docs/en/language-identification.html