

トランسفォーマーによる画像や音声の処理

本谷 秀堅

名古屋工業大学

Vision Transformer

ViT: Vision Transformer

画像を AI にとって都合の良いベクトル（トークン）で表現したい

- 文章：各単語をトークンで表現
- 画像は？

復習：画像データ

画像データ：画素値の集積



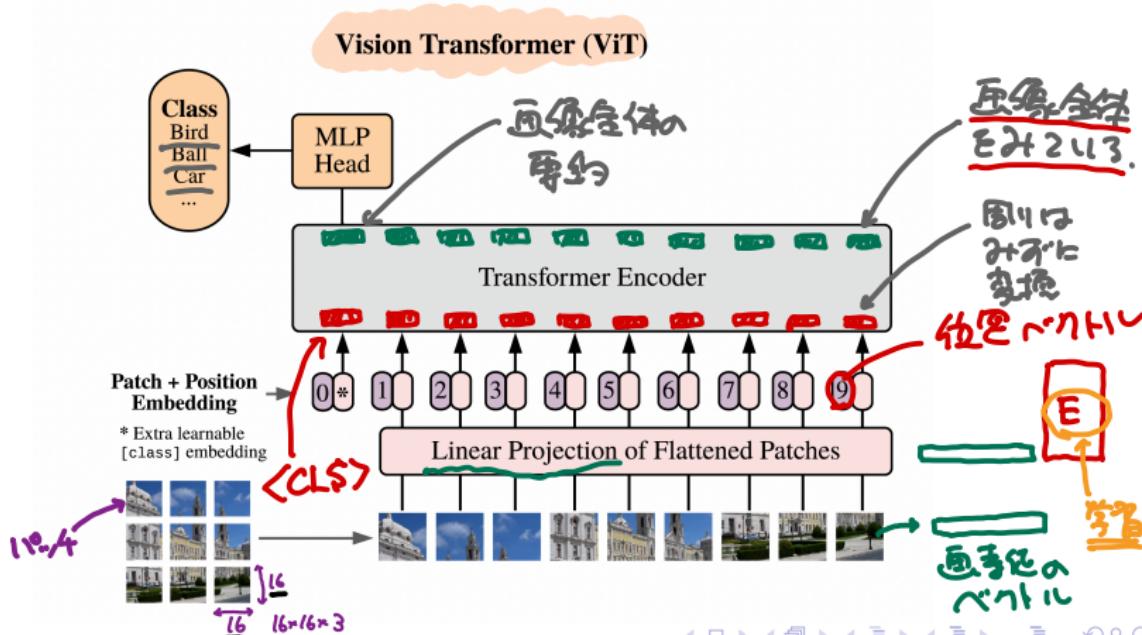
157	153	174	168	150	152	129	151	172	163	155	156
156	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	54	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	197	251	237	259	219	228	227	87	71	201
172	106	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	85	150	79	38	218	241
190	224	147	108	227	210	127	102	36	103	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	236	73	1	41	47	0	4	217	264	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	12	96	218

157	153	174	168	150	152	129	172	161	155	156
156	182	163	74	75	62	33	17	110	210	180
180	180	50	14	54	6	10	33	48	106	159
206	109	5	124	131	111	120	204	166	15	56
194	68	197	251	237	259	219	228	227	87	71
172	106	207	233	233	214	220	239	228	98	74
188	88	179	209	185	215	211	158	139	75	20
189	97	165	84	10	168	134	11	31	62	22
199	168	191	193	158	227	178	143	182	106	36
205	174	155	252	236	231	149	178	228	43	95
190	216	116	149	236	187	85	150	79	38	218
190	224	147	108	227	210	127	102	36	101	255
190	214	173	66	103	143	96	50	2	109	249
187	196	235	75	1	41	47	0	6	217	255
183	202	237	145	0	0	12	108	200	138	243
195	206	123	207	177	121	123	200	175	13	96

- カラー画像であれば各ピクセルは 3次元ベクトル (RGB: Red/Green/Blue) で表現されている
 - 画素ごとにトークンを求めるのは非現実的（画素数多過ぎ）

画像をトランスフォーマーで

- 画像を「パッチ」に分割
- それぞれのパッチを「単語」のように扱う
 - $P \times P$ の正方形 (多くの場合 $P = 16$)
 - 元画像が $H \times W$ だったときパッチは $N = HW/P^2$ 枚



トークン化の意味

パッチは、そのままでもベクトルで表現できる

- 元画像が C チャンネルだとする
 - 白黒画像 : $C = 1$
 - カラー画像 : $C = 3$

パッチのサイズを $P \times P$ とする。画像中の i 番目のパッチの j 番目の画素値を $x_{ij} \in \mathbb{R}^C$ で表すとき、次式でパッチ画像をベクトル表現できる：

$$\mathbf{x}_i = [\mathbf{x}_{i1}^\top, \mathbf{x}_{i2}^\top, \dots, \mathbf{x}_{iJ}^\top]^\top \in \mathbb{R}^{P^2 C}$$

ただし、 $J = P^2$ 。例えば $C = 3$ のとき、 $\mathbf{x}_i = [r_i, g_i, b_i]^\top \in \mathbb{R}^3$ など。

- 重要：このままでは AI に向き（次ページ）

トークン化の意味

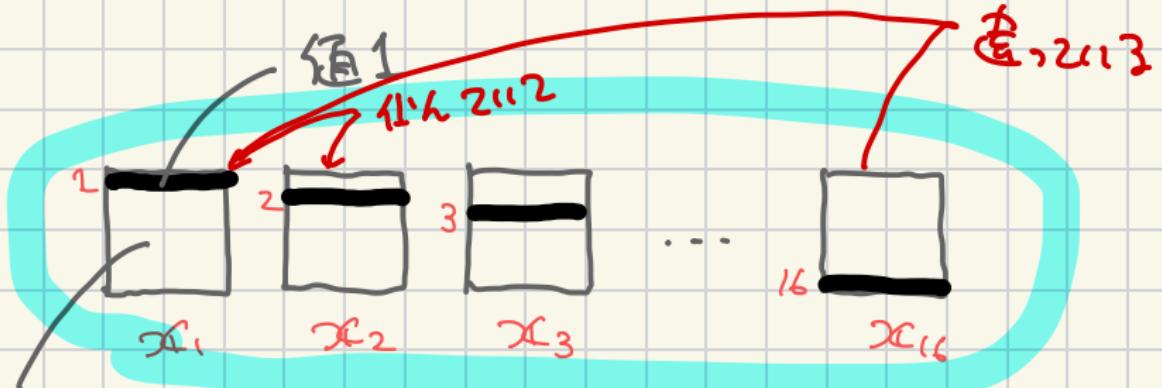
例題 $C = 1$, $P = 16$ とする。黒地に上から i 番目の行に幅 1 ピクセルの白の水平線が描かれたパッチ画像を x_i で表す。黒地部分の画素値は 0 で、白地部分の画素値は 1 とする。例えば x_2 は次式のような $P^2 = 256$ 次元の二値ベクトルである。

$$x_2 = \underbrace{[0, 0, \dots, 0]}_{16}, \underbrace{1, 1, \dots, 1}_{16}, \underbrace{0, 0, \dots, 0}_{16 \times 14}]^\top$$

$$\underbrace{x_i^\top x_j}_{\text{if } 0} \quad (i \neq j) \text{ を求めよ。}$$

この例では x_i ($i = 1, 2, \dots, P$) は互いに直交している。それぞれのベクトルの向きは、水平線の位置を一切反映していない。

- パッチ画像の画素値を並べたベクトル表現, x , は「パターンの類似度」評価に適していない



值0

$$\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \} 16$$

$$\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \} 16$$

$$\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \} 16$$

トランスフォーマーによる画像処理

トランスフォーマーへの入力準備

- ① 画像を $P \times P$ のパッチに切り分ける

- ② i 番目のパッチ: $x_i \in \mathbb{R}^{P^2 C}$ ($i = 1, 2, \dots, N$)

- $P = 16, C = 3$ のときは 768 次元

(例: 4x4と(画像全体は考慮しない))

- ③ 線形変換する: $(\tilde{z}_i^{(0)})^\top = x_i^\top E$, $E \in \mathbb{R}^{P^2 C \times D}$

- 例えば $D = 768$



- ④ 各パッチの位置ベクトルを足す: $z_i^{(0)} = \tilde{z}_i^{(0)} + r_i$

$$z_i^{(0)} = \tilde{z}_i^{(0)} + r_i$$

(例: 4x4の位置ベクトル)

- ⑤ 追加でひとつ Class トークンを追加する: $z_0^{(0)}$

- ⑥ それぞれをレイヤ正規化する

トランسفォーマーによる画像処理

トランسفォーマー層での処理

- ① クエリ、キー、バリューそれぞれのベクトルへと変換

$$q_i^\top = z_i^{(0)} \mathbf{W}^{(q)}, \quad k_i^\top = z_i^{(0)} \mathbf{W}^{(k)}, \quad v_i^\top = z_i^{(0)} \mathbf{W}^{(v)}$$

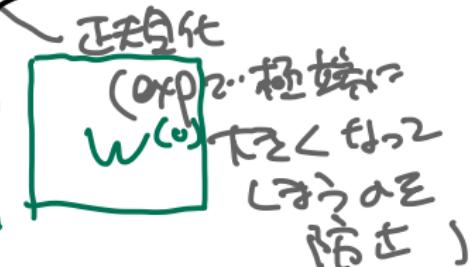
- ② 各ヘッドで注意（アテンション）の強さの計算

$$a_{ij} = \text{SM} \left(\frac{q_i^\top k_j}{\sqrt{D_k}} \right)$$

- ③ 各ヘッドでトークンの更新

attention
pooling

$$h_i = \sum_j a_{ij} v_j$$



- ④ マルチヘッドの統合

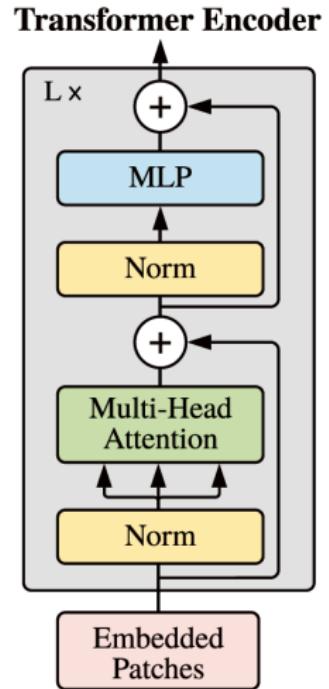


$$\tilde{z}_i^\top = [h_i^\top(1), h_i^\top(2), \dots, h_i^\top(H)] \mathbf{W}^{(o)}$$

- ⑤ 残差接続 : $z'_i = \tilde{z}_i + z_i^{(0)}$

トランスフォーマーによる画像処理

- ① レイヤ正規化: $\text{LN}(z')$
- ② MLP を適用 : $z^{(1)} = \text{MLP}(\text{LN}(z')) + z'$
- ③ 出力 : $y = \text{LN}(z^{(1)})$



トランスフォーマーによる画像識別器の構築

Class トークンを使って入力画像が何の画像かを識別する

- 画像中の対象を強く注意することを期待する



クラストークンを
させると、
attention
強さ。

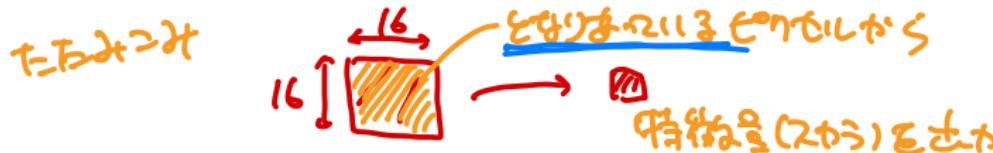


飞机

- CNN の受容野を思い出す
- ヘッドごとに注意する領域が異なる / 層の深さにも依存
- 必ずしも常に対象のみに注意するわけではなさそう

[1]: Dosovitskiy, AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE, 2021

CNN vs トランスフォーマー

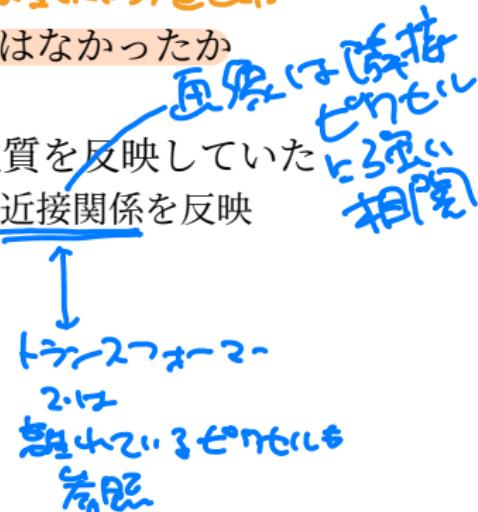


画像に適しているアーキテクチャは CNN ではなかったか

- ① CNN: アーキテクチャが画像の有する性質を反映していた

- 置込みフィルタによるピクセル間の近接関係を反映
- 並進同変性

- トランスフォーマー
 - 画像処理に適した構造を持たない



CNN vs トランスフォーマー

CNN と Transformer: ネットワークの規模を概ね揃えて、同じデータベースを使って学習して性能比較

文献 [1] によると：

- 学習データ数が少ないとき：トランスフォーマーは過学習して汎化性能は劣る
- 学習データ数が大規模なとき：トランスフォーマーが CNN を凌駕

「大規模」とは…

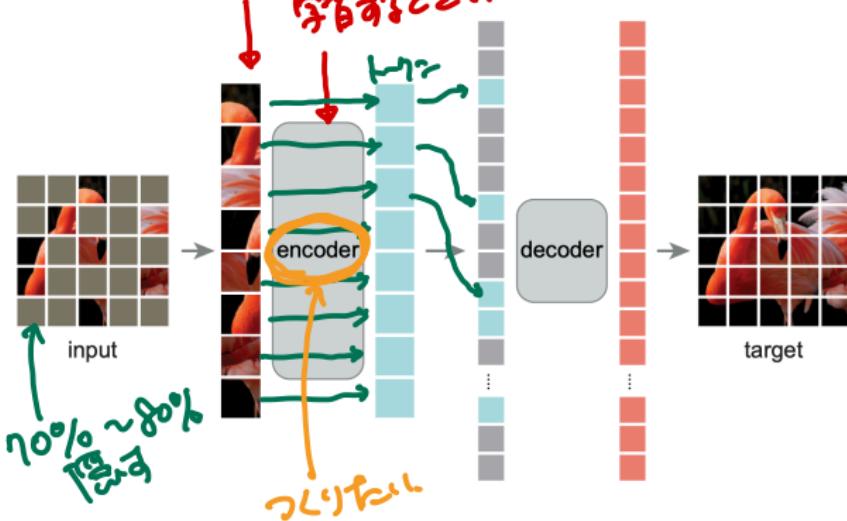
- JFT-300M: 3億3千枚の画像・1万8千クラスのラベル
- ImageNet-21k: 1400万枚の画像・2万1千クラスのラベル
- ImageNet: 130万枚の画像・1千クラスのラベル

[1]: Dosovitskiy, AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE, 2021

- マスクされた部分の画素値を予測
 - マスクされた部分はエンコード対象から除外（軽量）
 - エンコーダより小さいデコーダ

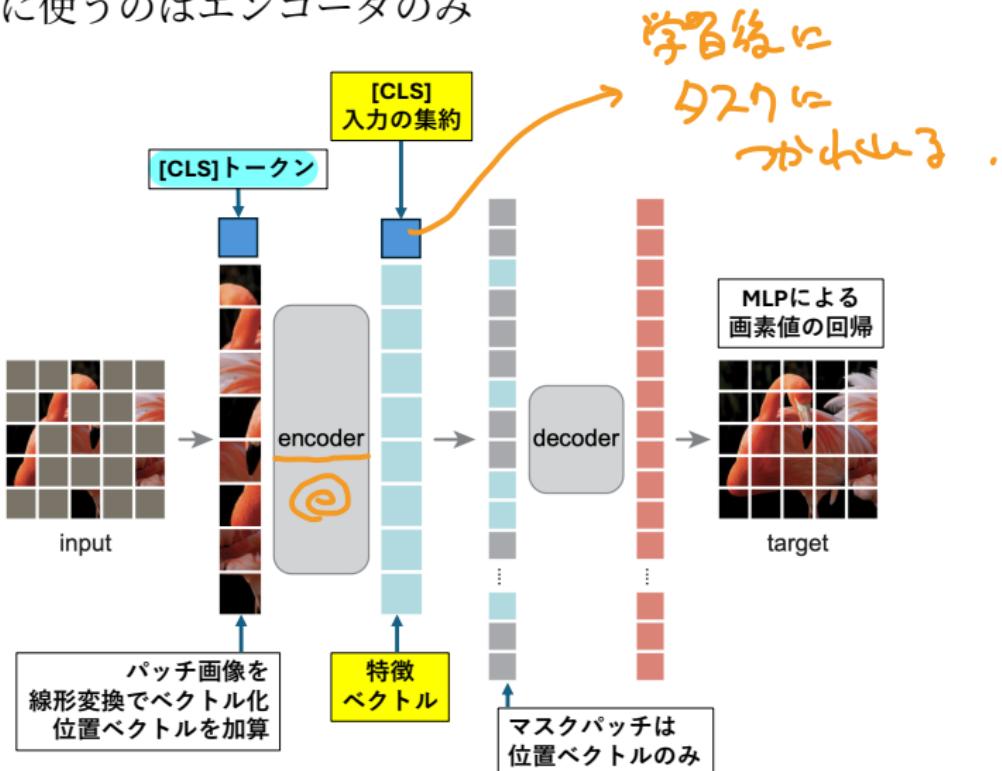
見立38=30-8t

字首可加小篆上聲

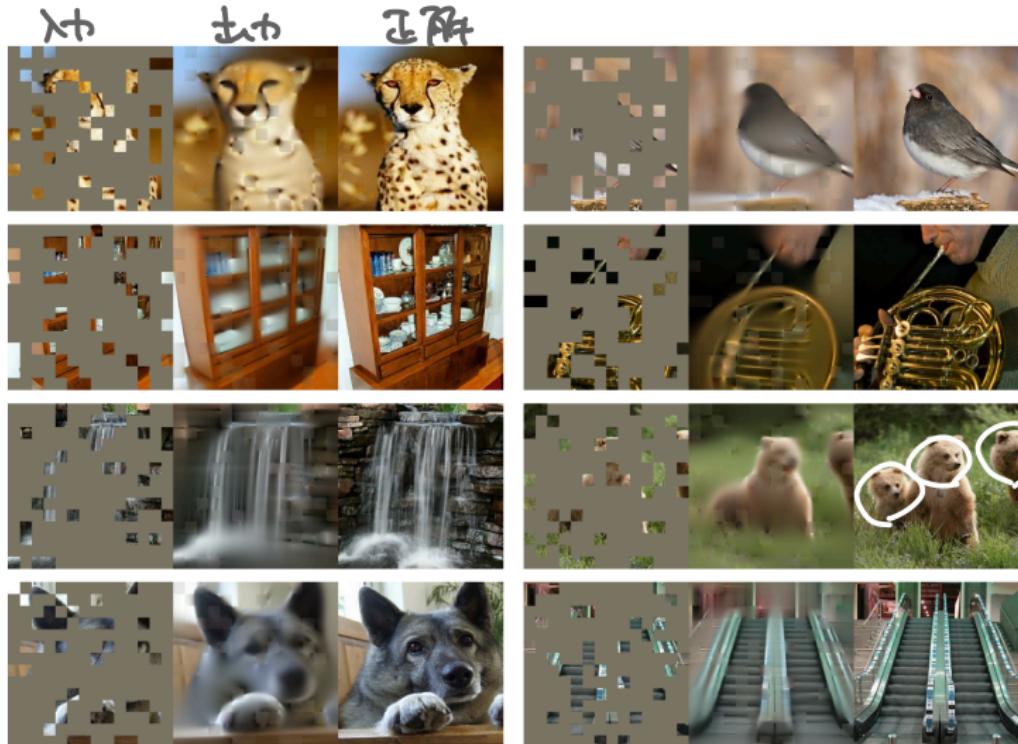


MAE: Masked Auto Encoder

- 学習後に使うのはエンコーダのみ

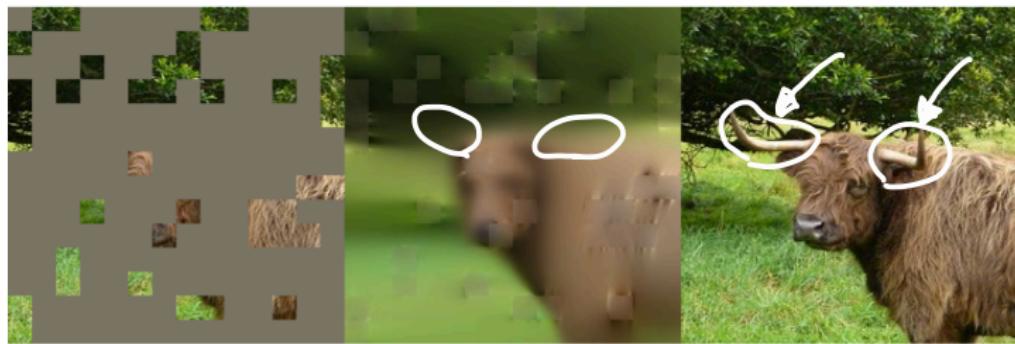
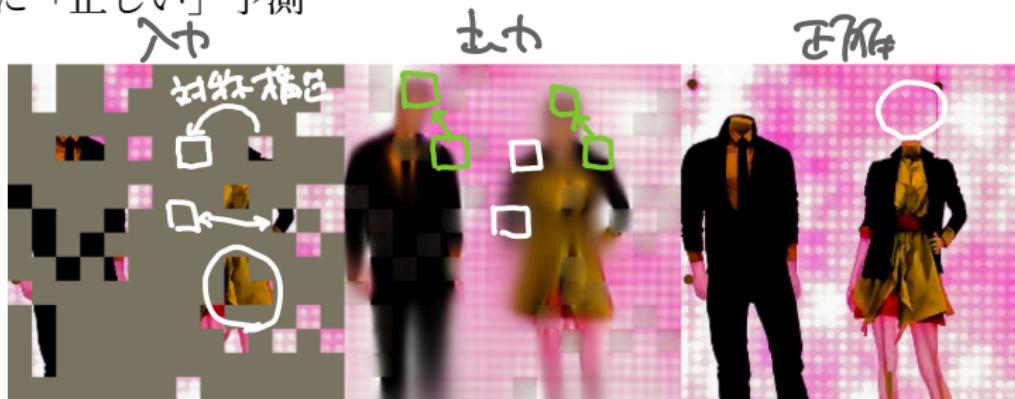


MAE: Masked Auto Encoder



MAE: Masked Auto Encoder

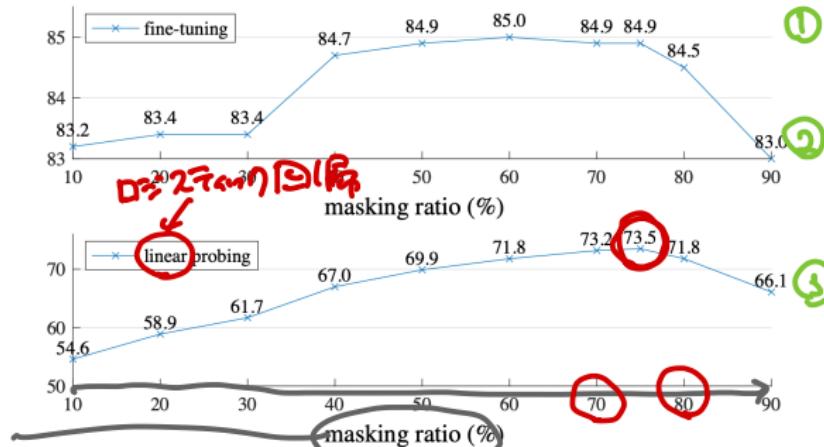
文脈的に「正しい」予測



MAE: Masked Auto Encoder

[CLS] トークンを使って線形識別 (ImageNet-1K)

1000 例
タグラス識別



① MAE で
トークンを学習

② 初め2つ3つは
から <CLS>
132 トークン

③ 線形識別器
で学習

おもてごりの
領域では
特に
効果

Figure 5. **Masking ratio.** A high masking ratio (75%) works well for both fine-tuning (top) and linear probing (bottom). The y-axes are ImageNet-1K validation accuracy (%) in all plots in this paper.

全体の 75% をマスクして学習したときが最良

画像の冗長性を排除して学習するにはマスクを大きくしたほうが良い

33113103

(5212)

କ୍ଷୁର ଟ-ହେ ଅନ୍ତର୍ଭାବରେ

"କାଟିଗ'ି" ରେ ଉପସ୍ଥିତ କରିବାକୁ

ପାଇଁ.