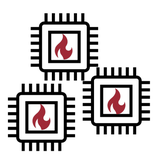
大規模Vision&Languageマルチタスクモデルの学習効率化とHuman-Object Interactionへの適用

軸屋敬介†，梁瀬和哉‡，表英輝†，土田裕登†，加藤邦人†
†：岐阜大学工学部，‡：日本車輌製造株式会社

研究背景

◆ 大規模Vision&Languageモデル
画像と**言語**の統合 ➡ 高い性能
◆ 大規模マルチタスクモデル
知識の組み合わせ ➡ 未学習タスクを解く（推論能力）

【問題点】


モデルサイズ
(=学習コスト)
の増加


学習方法
非公開


データセット
非公開

【対策】


一部重みの
凍結


学習方法
公開


公開データ
使用

これらを含むモデルの作成と下流タスクへの適用

【効率的な学習】
◆ 720データ（短い:長い=360:360）での計算

通常のマルチタスク学習

無駄なメモリを使用

短いタスク	abbey	<pad>	<pad>	<pad>	<pad>	<pad>	<pad>
長いタスク	cruise	ship	and	boats	by	the	dock

120バッチ ×6
バッチ内

提案手法

タスクごとのバッチサイズを設定
勾配を蓄積して勾配更新

短いタスク	abbey						
	bar						

360バッチ ×1

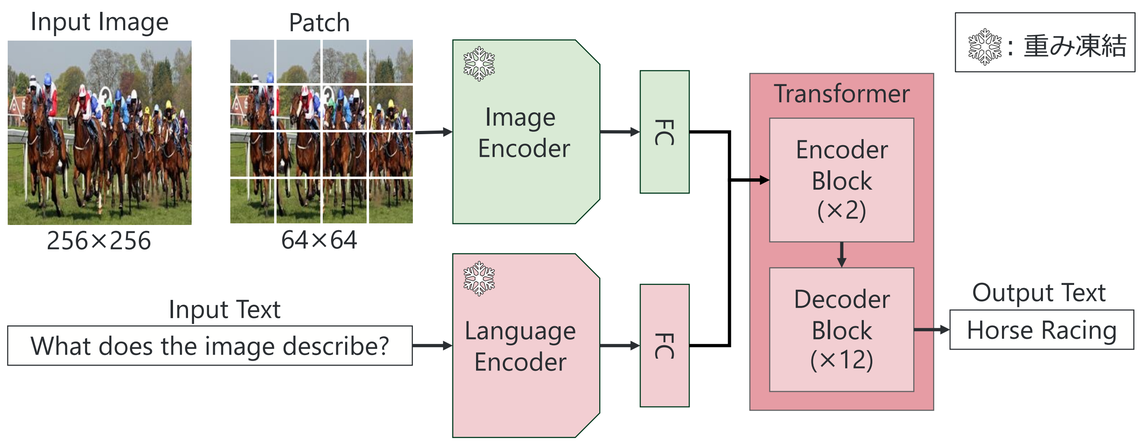
長いタスク	A	food	item	on	the	table	<pad>
	cruise	ship	and	boats	by	the	dock

120バッチ ×3
バッチ内

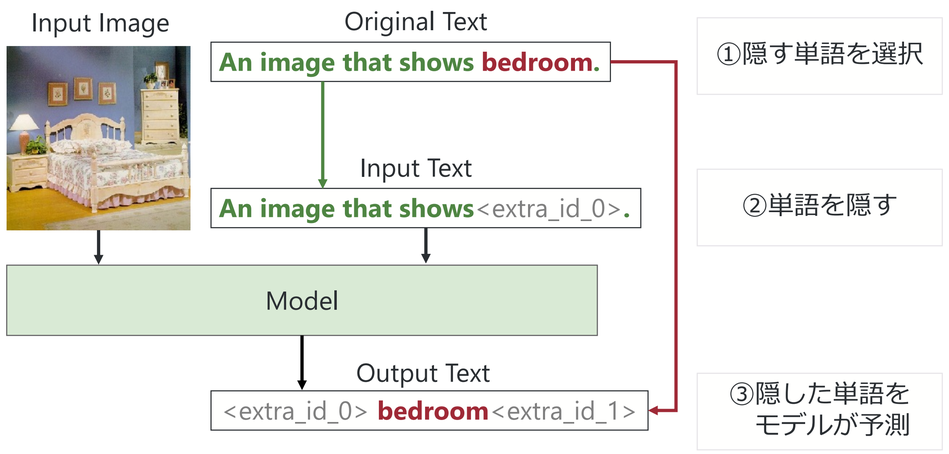
➡ 6 から 4 (=1+3) 回 へ **計算回数の削減**

提案手法

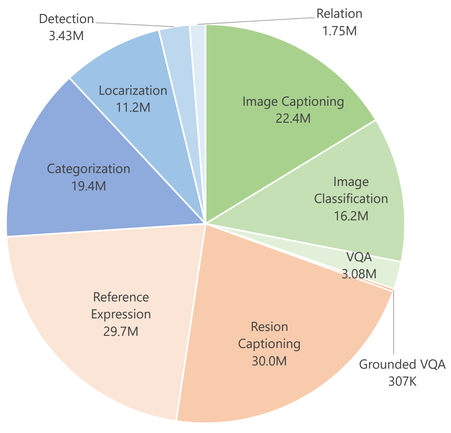
【モデル図】



【事前学習】
文章の一部を隠して学習



【タスク学習】
様々なタスクで
汎用的な知識を得る

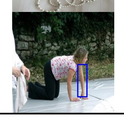


【ファインチューニング】
特定のタスクに適用

Human-Object Interaction
人と物体の**関係**を予測するタスク


人がボールを**持っている**

タスクの例

タスク		入力	出力
Region Captioning		What does the region <loc_81><loc_1192> describe?	heavily embellished wedding sandals
Categorization		What is the category of the region <loc_825><loc_1349>?	Human arm

実験結果

データセット
V-COCO

データセットの詳細

学習	テスト	関係
8,543	7,811	29

入力と出力

入力: What is the interaction between **person<loc_40><loc_1558>** and **donut<loc_376><loc_1024>**?

出力: hold, eat

評価指標
Marco F1

$$\text{Macro F1} = \frac{1}{m}\sum_{i=1}^{m}(F_1)_i$$

比較対象
OFA，Kosmos-2

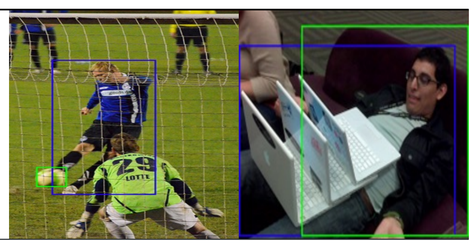
結果

vs $\text{OFA}_{\text{Large}}$
11倍⧗ 13.3%↙

vs OFA_{Huge}
24倍⧗ 11.5%↙

vs Kosmos-2
120倍⧗

結果の例

画像と領域			
正解	**look, kick**	**lay**
$\text{OFA}_{\text{Large}}$	look	sit
OFA_{Huge}	look	sit
Ours	**look, kick**	sit
Kosmos-2	**look, kick**	**lay**

パラメータと学習時間およびMacro F1

モデル	$\text{OFA}_{\text{Large}}$	OFA_{Huge}	Ours	Kosmos-2
総パラメータ	473M	929M	699M	1798M
学習パラメータ	412M	853M	**162M**	1798M
学習時間 [1エポック] (6000Ada×4)	11 min	24 min	**1 min**	**120 min**
RTX4090での バッチサイズ	16	**不可**	**160**	**不可**
Macro F1 (Oursとの比較)	68.0% **(−13.3%)**	69.8% **(−11.5%)**	81.3% (±0%)	84.3% **(+3.0%)**

今後の課題

パラメータ数が最多のKosmos-2が最も良い精度
■ モデルやデータの改良で
軽量さを維持しつつKosmos-2を超える性能を実現したい