OFA: Towards Unified Multimodal Multitask Pretraining

Junyang Lin

Twitter: @JustinLin610

Email: justinlin930319@hotmail.com

DAMO Academy, Alibaba Group

Overview

- Review of Multimodal Pretraining
- Introduction to OFA (One-For-All)
    - Methodology
    - Experiments
    - Extension: Prompt Tuning, Chinese Models, …
    - Opensource and Demos
- Future Work

Review of Multimodal Pretraining

Vision-Language Tasks

Image Captioning

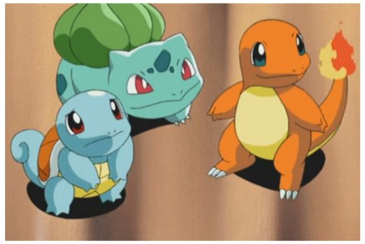
the album cover of the beatles abbey road

Visual Question Answering

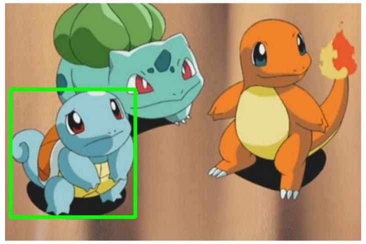
What is the style of the painting?

Impressionism

Visual Grounding



a blue turtle-like pokemon with round head



Text-to-Image Generation

A clock tower looms underneath a clear sky.

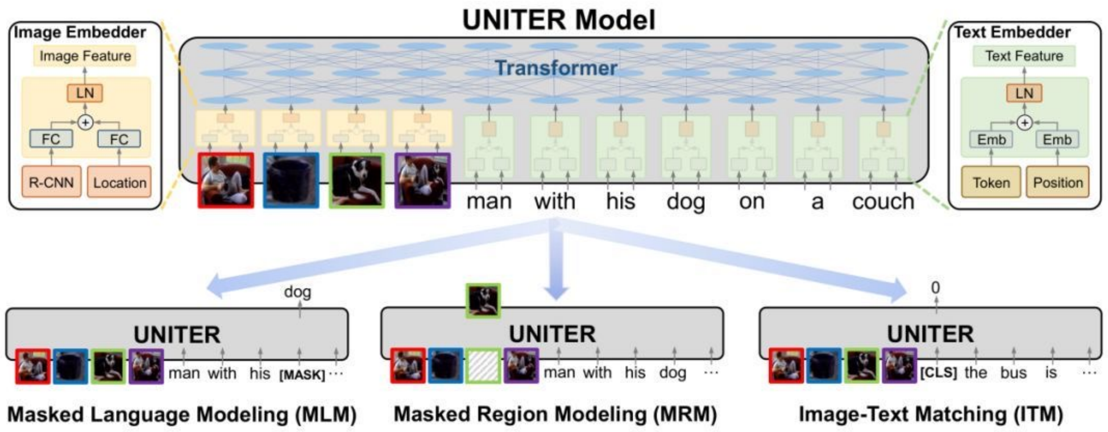
Pretraining on Large-scale Datasets

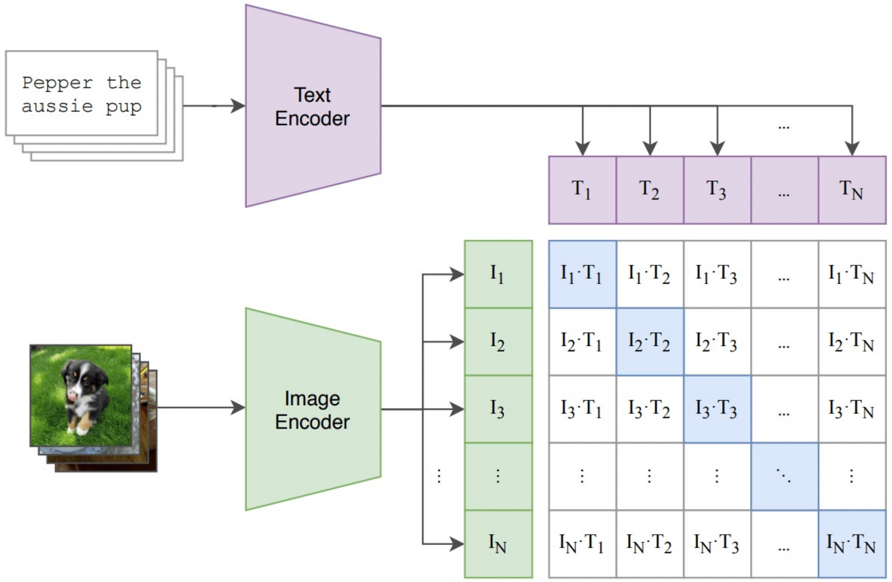
- Large datasets of image-text pairs
- Pretraining with ”language” modeling & image-text pairing, …
- Transfer to downstream tasks with finetuning

Two Trends in Multimodal Pretraining

Generative Pretraining

Contrastive Pretraining

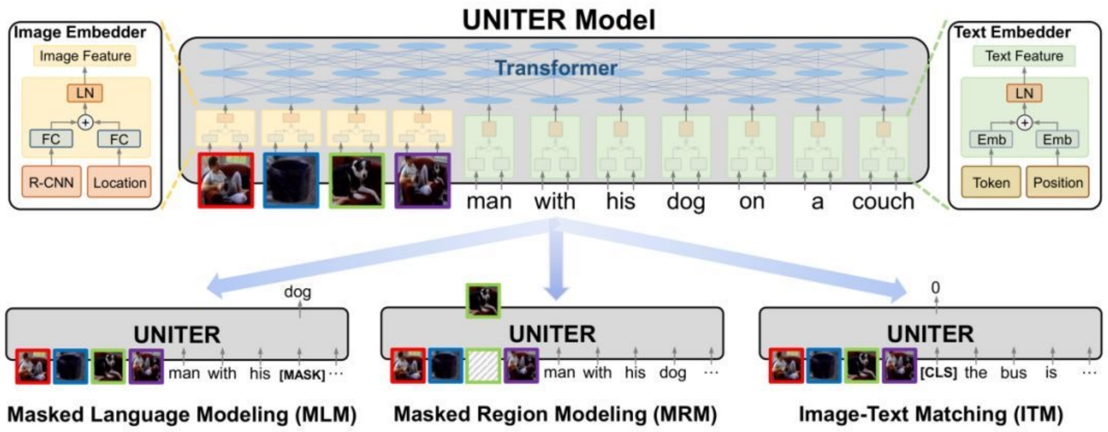


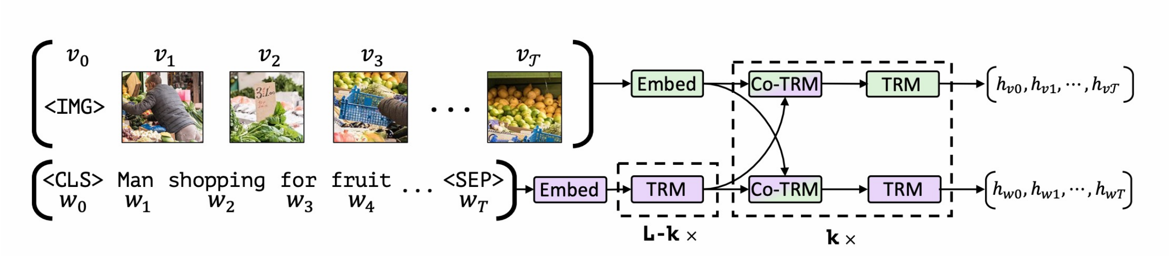


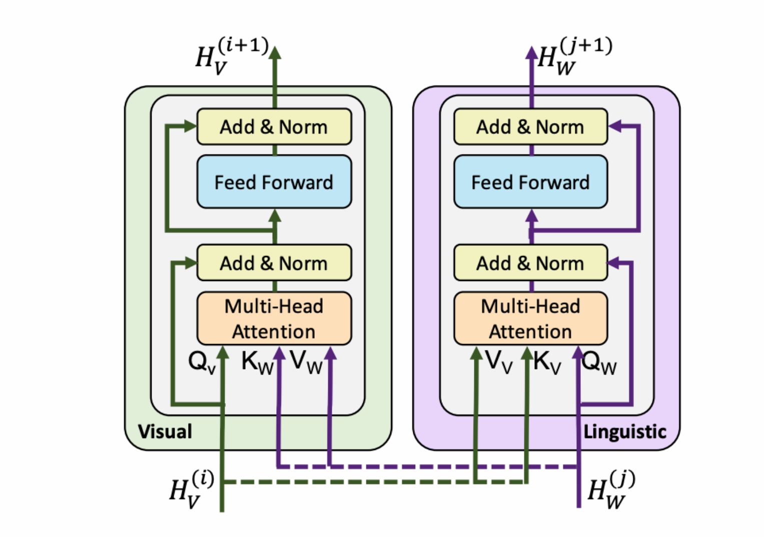
Transfer of BERT to VL

Single Stream

Dual Stream



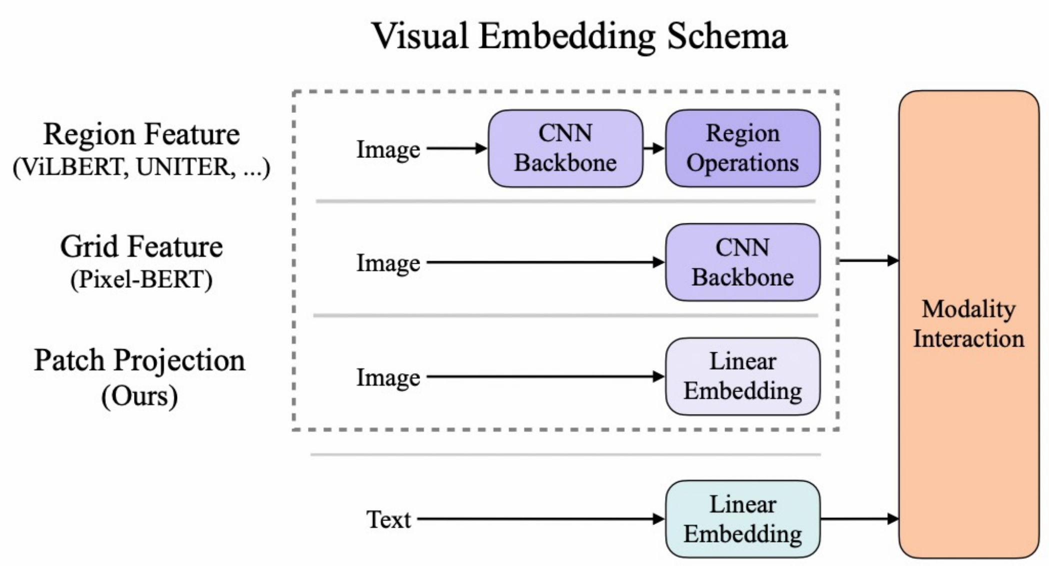


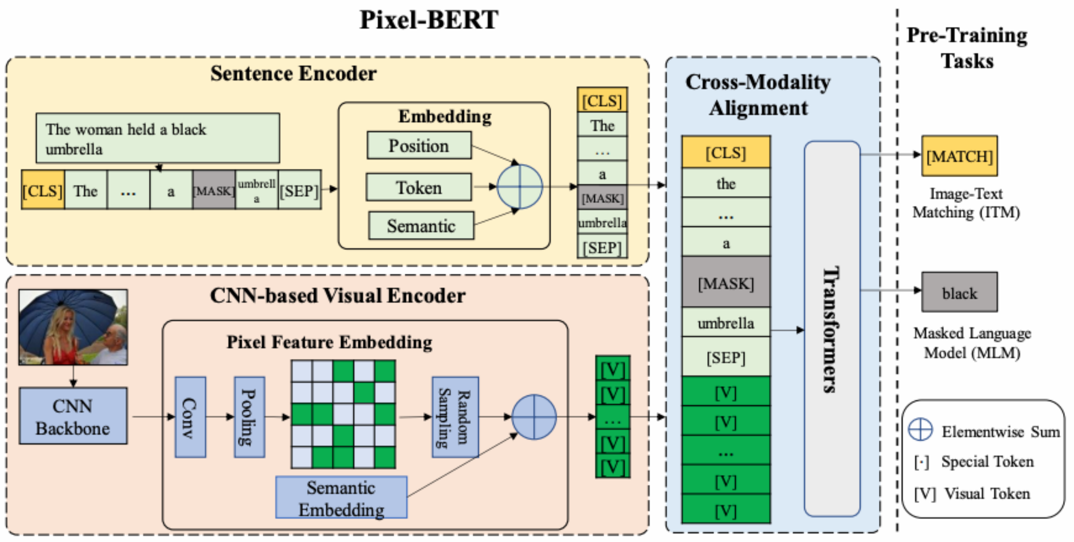


From Objects to Raw Image Features

Patch Projection

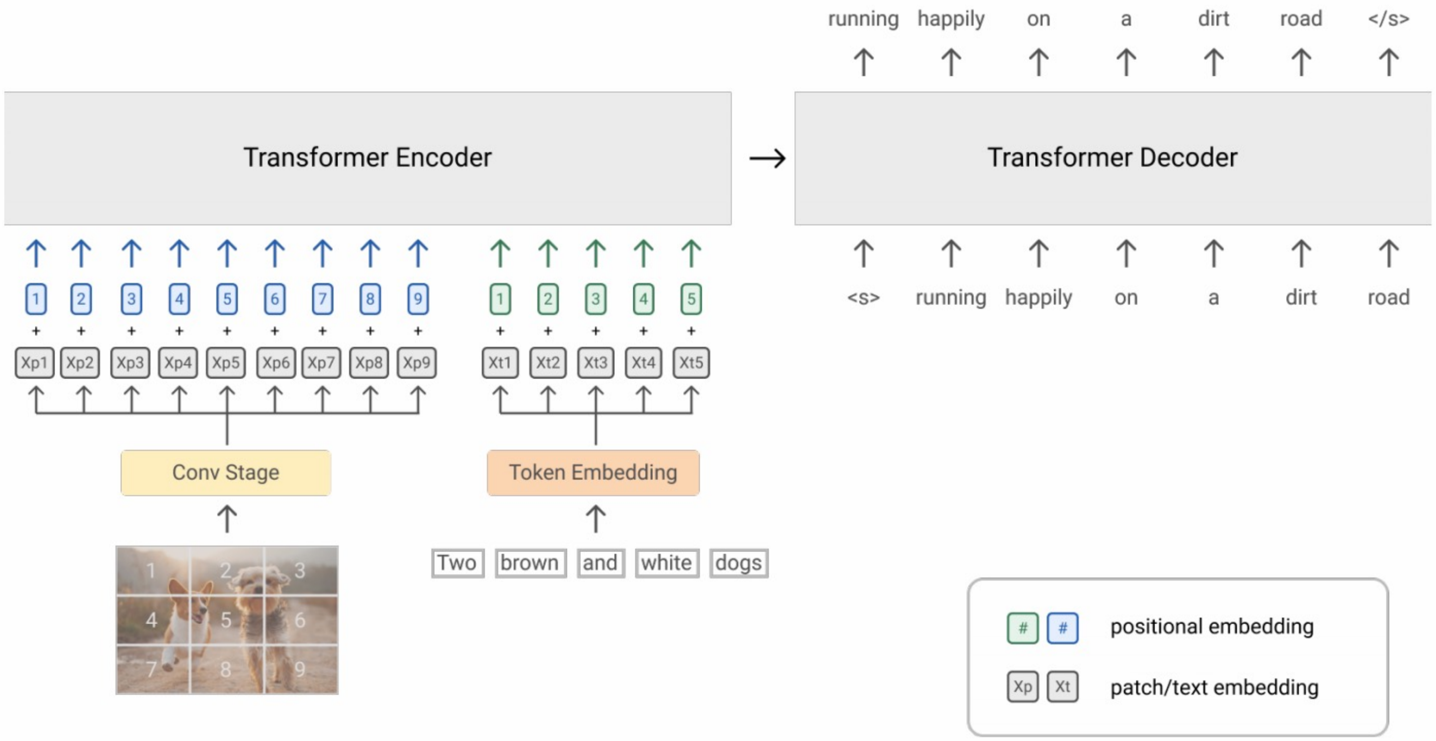
Vision Backbone





Adapting Understanding and Generation

Encoder-Decoder Framework



Summary

- Pretraining is important for vision-language representation learning
- A simple end-to-end model is expected
- Stepping forward to Unification (OFA, Gato, Unified-IO, GIT, etc.)

OFA: Multimodal Multitask Pretraining for a “One-For-All”Model

Features for a Unified Model

Task Agnostic

Unified task representation to support different types of tasks

Modality Agnostic

Unified input and output representation shared among all tasks to handle different modalities

Task Comprehensive

Enough task variety to accumulate generalization ability robustly

3 Unifications

I/O

Shard I/O across different modalities and tasks

Architecture

A shared encoder-decoder framework without task-specific layers

Task

Varieties of tasks are unified to the sequence-to-sequence format

I/O

Byte-Pair Encoding for texts

The boy was playing on the ground

1169, 7081, 9776, 17916, 261, 1169, 2833

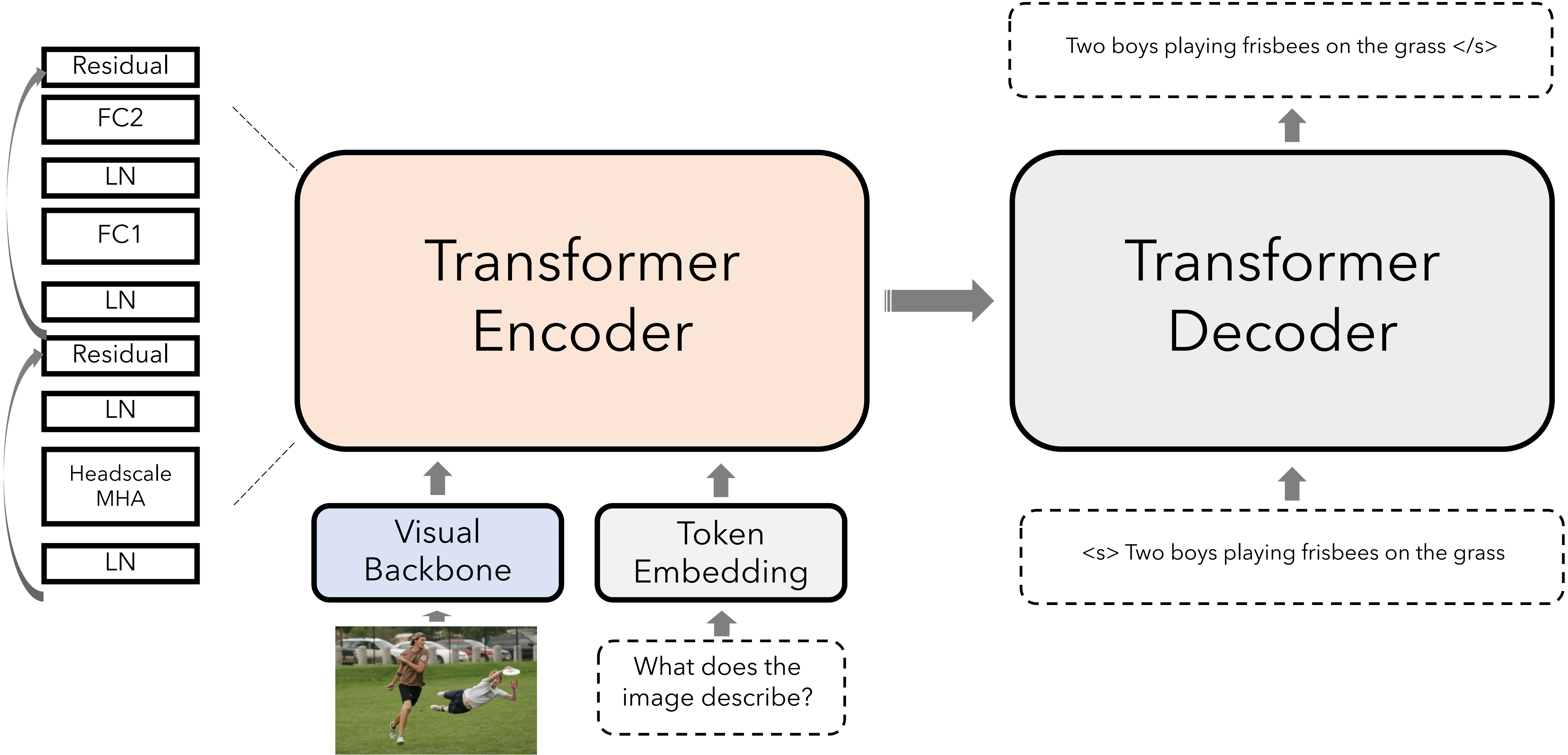
Vector Quantization for images

<code_314>, <code_78>, <code_23> ……

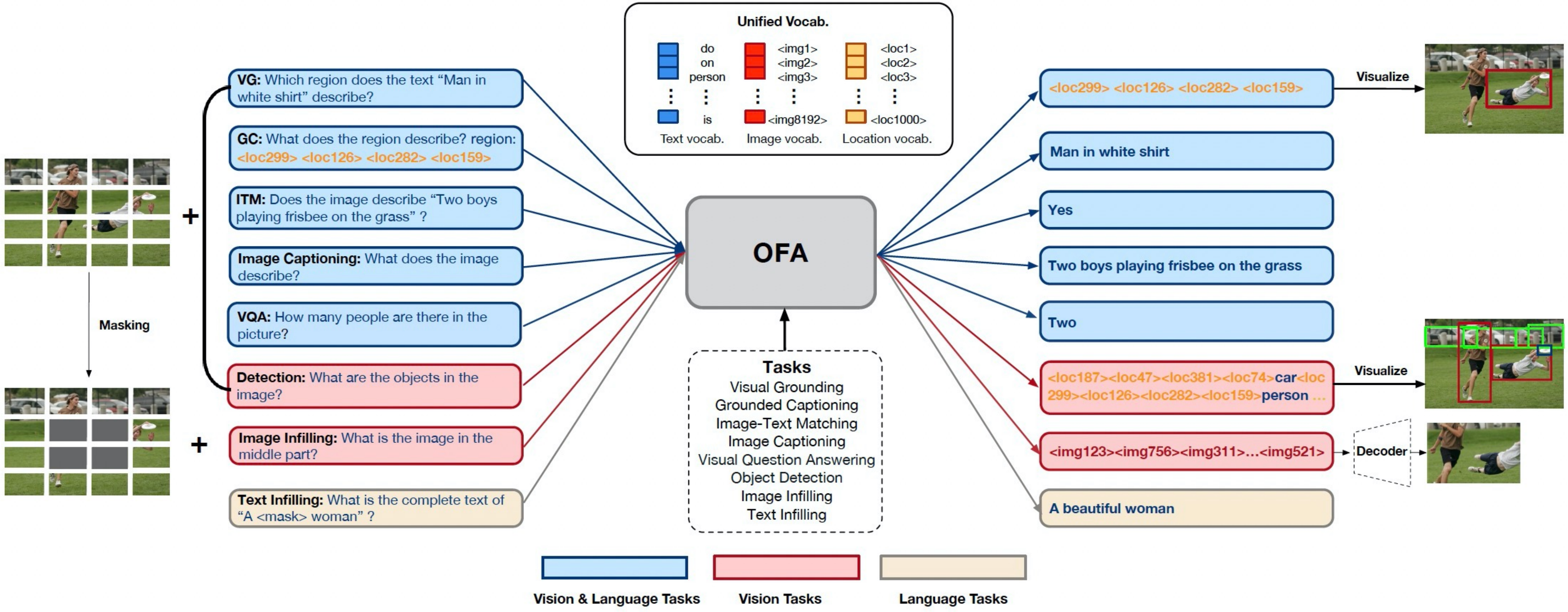
Discretization for bounding boxes

<loc_324>, <loc_512>, <loc_33>, <loc_629>

Architecture



Task


Unified Vocab.
do on person … is
Text vocab.
<img1> <img2> <img3> … <img8192>
Image vocab.
<loc1> <loc2> <loc3> … <loc1000>
Location vocab.
VG: Which region does the text “Man in white shirt” describe?
GC: What does the region describe? region: <loc299> <loc126> <loc282> <loc159>
ITM: Does the image describe “Two boys playing frisbee on the grass” ?
Image Captioning: What does the image describe?
VQA: How many people are there in the picture?
Detection: What are the objects in the image?
Image Infilling: What is the image in the middle part?
Text Infilling: What is the complete text of “A <mask> woman” ?
Masking
OFA
Tasks
Visual Grounding
Grounded Captioning
Image-Text Matching
Image Captioning
Visual Question Answering
Object Detection
Image Infilling
Text Infilling
<loc299> <loc126> <loc282> <loc159>
Visualize
Man in white shirt
Yes
Two boys playing frisbee on the grass
Two
<loc187><loc47><loc381><loc74>car<loc299><loc126><loc282><loc159>person …
Visualize
<img123><img756><img311>…<img521>
Decoder
A beautiful woman
Vision & Language Tasks
Vision Tasks
Language Tasks

Pretraining Datasets

Type	Pretraining Task	Source	#Image	#Sample
Vision & Language	Image Captioning Image-Text Matching	CC12M, CC3M, SBU, COCO, VG-Cap	14.78M	15.25M
	Visual Question Answering	VQAv2, VG-QA, GQA	178K	2.92M
	Visual Grounding Grounded Captioning	RefCOCO, RefCOCO+, RefCOCOg, VG-Cap	131K	3.20M
Vision	Detection	OpenImages, Object365, VG, COCO	2.98M	3.00M
	Image Infilling	OpenImages, YFCC100M, ImageNet-21K	36.27M	-
Language	Masked Language Modeling	Pile (Filtered)	-	140GB*

Model Card

Model	#Param.	Backbone	Hidden size	Intermediate Size	#Head	#Enc. Layers	#Dec. Layers
OFA_{Tiny}	33M	ResNet50	256	1024	4	4	4
$\text{OFA}_{\text{Medium}}$	93M	ResNet101	512	2048	8	4	4
OFA_{Base}	182M	ResNet101	768	3072	12	6	6
$\text{OFA}_{\text{Large}}$	472M	ResNet152	1024	4096	16	12	12
OFA_{Huge}	930M	ResNet152	1280	5120	16	24	12

Experiments

- Multimodal:
  - Cross-modal understanding: VQA, SNLI-VE.
  - Image-to-text generation: MSCOCO Caption
  - Visual Grounding: RefCOCO, RefCOCO+, RefCOCOg
  - Text-to-Image Generation: MSCOCO
- Unimodal:
  - NLU: GLUE
  - NLG: Gigaword
  - Image Classification: ImageNet

Vision-Language Understanding

Model	VQA		SNLI-VE	
	test-dev	test-std	dev	test
UNITER	73.8	74.0	79.4	79.4
OSCAR	73.6	73.8	-	-
VILLA	74.7	74.9	80.2	80.0
VL-T5	-	70.3	-	-
VinVL	76.5	76.6	-	-
UNIMO	75.0	75.3	81.1	80.6
ALBEF	75.8	76.0	80.8	80.9
METER	77.7	77.6	80.9	81.2
VLMo	79.9	80.0	-	-
SimVLM	80.0	80.3	86.2	86.3
Florence	80.2	80.4	-	-
OFA_{Tiny}	70.3	70.4	85.3	85.2
$\text{OFA}_{\text{Medium}}$	75.4	75.5	86.6	87.0
OFA_{Base}	78.0	78.1	89.3	89.2
$\text{OFA}_{\text{Large}}$	80.3	80.5	90.3	90.2
OFA	**82.0**	**82.0**	**91.0**	**91.2**

Image Captioning

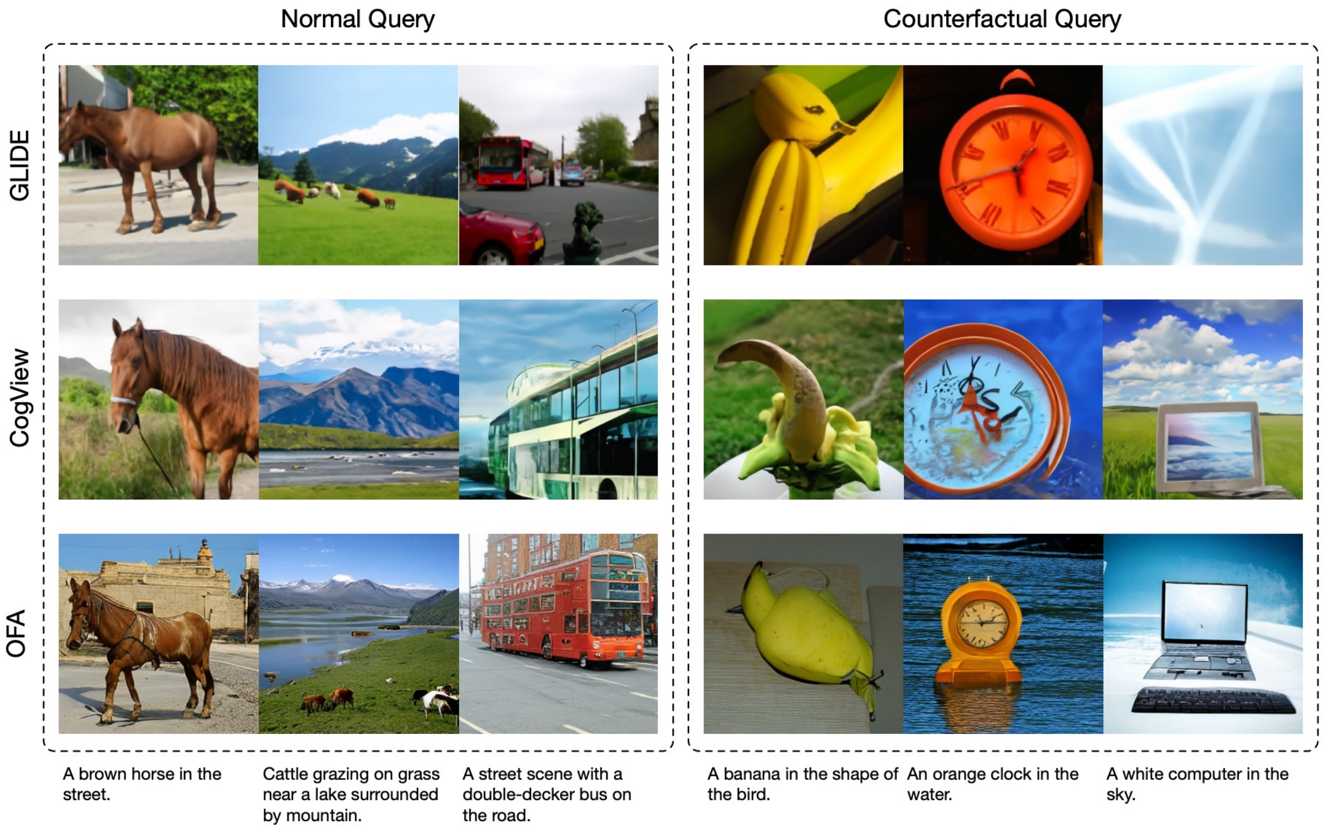
Model	Cross-Entropy Optimization BLEU@4	METEOR	CIDEr	SPICE	CIDEr Optimization BLEU@4	METEOR	CIDEr	SPICE
VL-T5 [57]	34.5	28.7	116.5	21.9	-	-	-	-
OSCAR [15]	37.4	30.7	127.8	23.5	41.7	30.6	140.0	24.5
UNICORN [58]	35.8	28.4	119.1	21.5	-	-	-	-
VinVL [17]	38.5	30.4	130.8	23.4	41.0	31.1	140.9	25.2
UNIMO [47]	39.6	-	127.7	-	-	-	-	-
LEMON [24]	41.5	30.8	139.1	24.1	42.6	31.4	145.5	25.5
SimVLM [22]	40.6	**33.7**	143.3	**25.4**	-	-	-	-
OFA_{Tiny}	35.9	28.1	119.0	21.6	38.1	29.2	128.7	23.1
$\text{OFA}_{\text{Medium}}$	39.1	30.0	130.4	23.2	41.4	30.8	140.7	24.8
OFA_{Base}	41.0	30.9	138.2	24.2	42.8	31.7	146.7	25.8
$\text{OFA}_{\text{Large}}$	42.4	31.5	142.2	24.5	43.6	32.2	150.7	26.2
OFA	**43.9**	31.8	**145.3**	24.8	**44.9**	**32.5**	**154.9**	**26.6**

Visual Grounding

Model	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val-u	test-u
VL-T5	-	-	-	-	-	-	-	71.3
UNITER	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77
VILLA	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
MDETR	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
UNICORN	88.29	90.42	83.06	80.30	85.05	71.88	83.44	83.93
OFA_{Tiny}	80.20	84.07	75.00	68.22	75.13	57.66	72.02	69.74
OFA_{Medium}	85.34	87.68	77.92	76.09	83.04	66.25	78.76	78.58
OFA_{Base}	88.48	90.67	83.30	81.39	87.15	74.29	82.29	82.31
OFA_{Large}	90.05	92.93	85.26	85.80	89.87	79.22	85.89	86.55
OFA	**92.04**	**94.03**	**88.44**	**87.86**	**91.70**	**80.71**	**88.07**	**88.78**

Text-to-Image Generation

Model	FID↓	CLIPSIM↑	IS↑
DALLE	27.5	-	17.9
CogView	27.1	33.3	18.2
GLIDE	12.2	-	-
Unifying	29.9	30.9	-
NÜWA	12.9	34.3	27.2
OFA	**10.5**	**34.4**	**31.1**



Text-to-Image Generation

An art painting of a soldier, in the style of cyperpunk.

The golden palace of the land of clouds.

Rustic interior of an alchemy shop.

An art painting of a city, in the style of cyberpunk.

A painting of the sunset cliffs in the style of fantasy art.

A painting of the superman.

An art painting of a dog, in the style of steampunk, white background.

A strawberry splashing in the coffee in a mug under the starry sky.

Elf elk in the forest illustration, HD, fantasy art.

An art painting of a city, in the style of steampunk.

A painting of the sunset cliffs in the style of dark fantasy art.

A painting of the superman, in the dark style.

Text Classification

Model	SST-2	RTE	MRPC	QQP	MNLI	QNLI
Multimodal Pretrained Baseline Models						
VisualBERT [38]	89.4	56.6	71.9	89.4	81.6	87.0
UNITER [14]	89.7	55.6	69.3	89.2	80.9	86.0
VL-BERT [8]	89.8	55.7	70.6	89.0	81.2	86.3
VilBERT [13]	90.4	53.7	69.0	88.6	79.9	83.8
LXMERT [40]	90.2	57.2	69.8	75.3	80.4	84.2
Uni-Perceiver [61]	90.2	64.3	86.6	87.1	81.7	89.9
SimVLM [22]	90.9	63.9	75.2	90.4	83.4	88.6
FLAVA [60]	90.9	57.8	81.4	90.4	80.3	87.3
UNIMO [46]	96.8	-	-	-	89.8	-
Natural-Language-Pretrained SOTA Models						
BERT [2]	93.2	70.4	88.0	91.3	86.6	92.3
RoBERTa [28]	96.4	86.6	90.9	92.2	90.2	93.9
XLNet [25]	**97.0**	85.9	90.8	92.3	90.8	94.9
ELECTRA [82]	96.9	88.0	90.8	92.4	90.9	95.0
DeBERTa [83]	96.8	88.3	**91.9**	92.3	**91.1**	**95.3**
Ours						
OFA	96.6	**91.0**	91.7	**92.5**	90.2	94.8

Text Generation

Model		Gigaword	
	ROUGE-1	ROUGE-2	ROUGE-L
BERTSHARE [85]	38.13	19.81	35.62
MASS [86]	38.73	19.71	35.96
UniLM [29]	38.45	19.45	35.75
PEGASUS [87]	39.12	19.86	36.24
ProphetNet [88]	39.55	20.27	36.57
UNIMO [46]	39.71	20.37	36.88
OFA	**39.81**	**20.66**	**37.11**

Image Classification

Model	Top-1 Acc.
EfficientNet-B7 [89]	84.3
ViT-L/16 [6]	82.5
DINO [90]	82.8
SimCLR v2 [32]	82.9
MoCo v3 [35]	84.1
BEiT$_{384}$-L/16 [36]	**86.3**
MAE-L/16 [37]	85.9
OFA	85.6

Ablation Study on Tasks

Model	Caption CIDEr	VQA Test-dev	ImageNet Top-1 Acc.	Image Generation FID / CLIPSIM / IS
OFA_{Base}	135.6	76.0	82.2	20.8 / 31.6 / 21.5
w/o text infill.	134.8	75.6	83.2	20.3 / 31.7 / 21.8
w/o image infill.	136.3	76.3	81.8	23.2 / 31.0 / 20.0
w/o det.	133.3	75.4	81.4	20.9 / 31.5 / 21.6
w/o ground.	134.2	75.5	82.0	21.2 / 31.5 / 21.5

Zero-shot Performance

Model	SST-2 Acc.	RTE Acc.	MRPC F1	QQP F1	QNLI Acc.	MNLI Acc.	SNLI-VE Acc. (dev/test)
Uni-Perceiver	70.6	55.6	76.1	53.6	51.0	**49.6**	-
OFA_{Base}	**71.6**	**56.7**	**79.5**	**54.0**	**51.4**	37.3	**49.71 / 49.18**

Old:
What color is the car?

New:
What color is the car in the region? region: <loc301>...

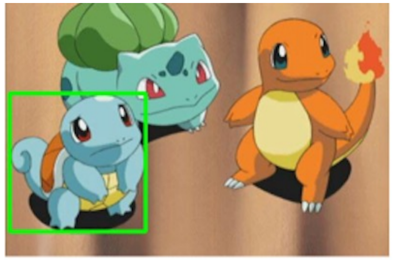
Q: what color is the car in the region? region: <loc301> <loc495> <loc501> <loc596>

A: tan

Q: what color is the car in the region? region: <loc512> <loc483> <loc675> <loc576>

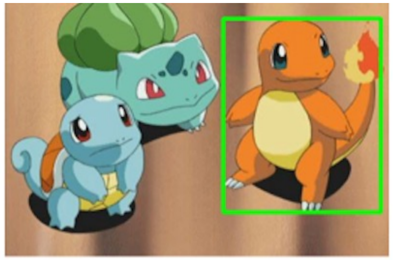
A: gray

Out-of-Domain



A blue turtle-like pokemon with round head.

A green toad-like pokemon with seeds on its back.



A red dinosaur-like pokemon with a flaming tail.

a man with green hair in green clothes with three swords at his waist

a man in a straw hat and a red dress

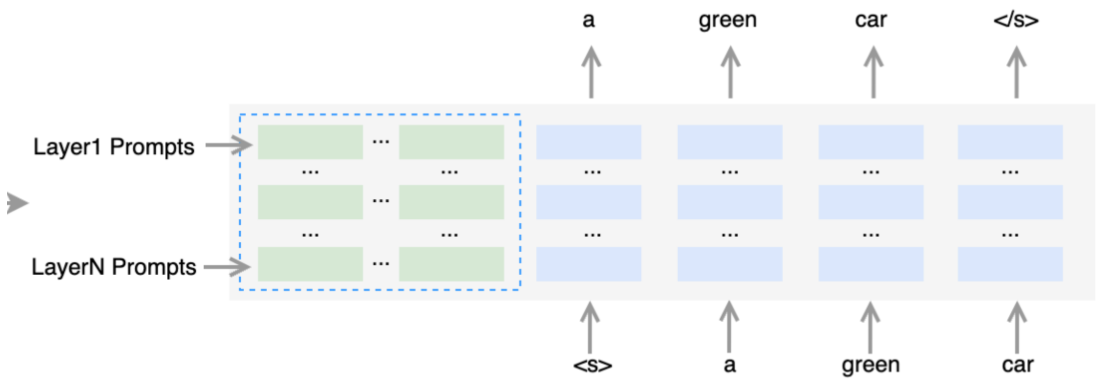
a blond-haired man in a black suit and brown tie

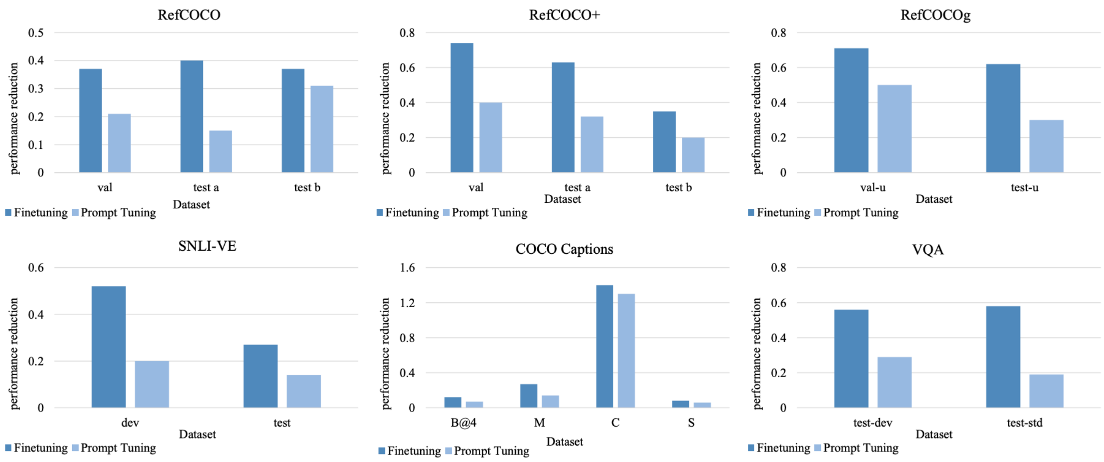
a sexy lady wearing sunglasses and a crop top with black hair

a man with a long nose in a hat and yellow pants

a strange skeleton

Extension: Prompt Tuning





Model	RefCOCO			RefCOCO+			RefCOCOg		SNLI-VE		COCO Captions				VQA	
	val	testA	testB	val	testA	testB	val-u	test-u	dev	test	B@4	M	C	S	test-dev	test-std
Base-size Models																
Finetuning	88.48	90.67	83.30	81.39	87.15	74.29	82.29	82.31	89.30	89.20	41.00	30.90	138.2	24.20	78.00	78.10
Prompt Tuning	84.53	85.21	77.36	76.34	81.44	67.68	75.61	76.57	88.18	88.59	39.70	30.10	134.2	23.50	74.31	74.47
Large-size Models																
Finetuning	90.05	92.93	85.26	85.80	89.87	79.22	85.89	86.55	90.30	90.20	42.40	31.50	142.2	24.50	80.40	80.70
Prompt Tuning	90.05	92.31	85.59	84.54	89.40	77.77	85.27	85.89	90.04	90.12	41.81	31.51	141.4	24.42	78.30	78.53

Extension: Chinese Models

- Large-scale Chinese datasets for pretraining
- Downstream transfer to Chinese image captioning and visual grounding

MUGE Caption

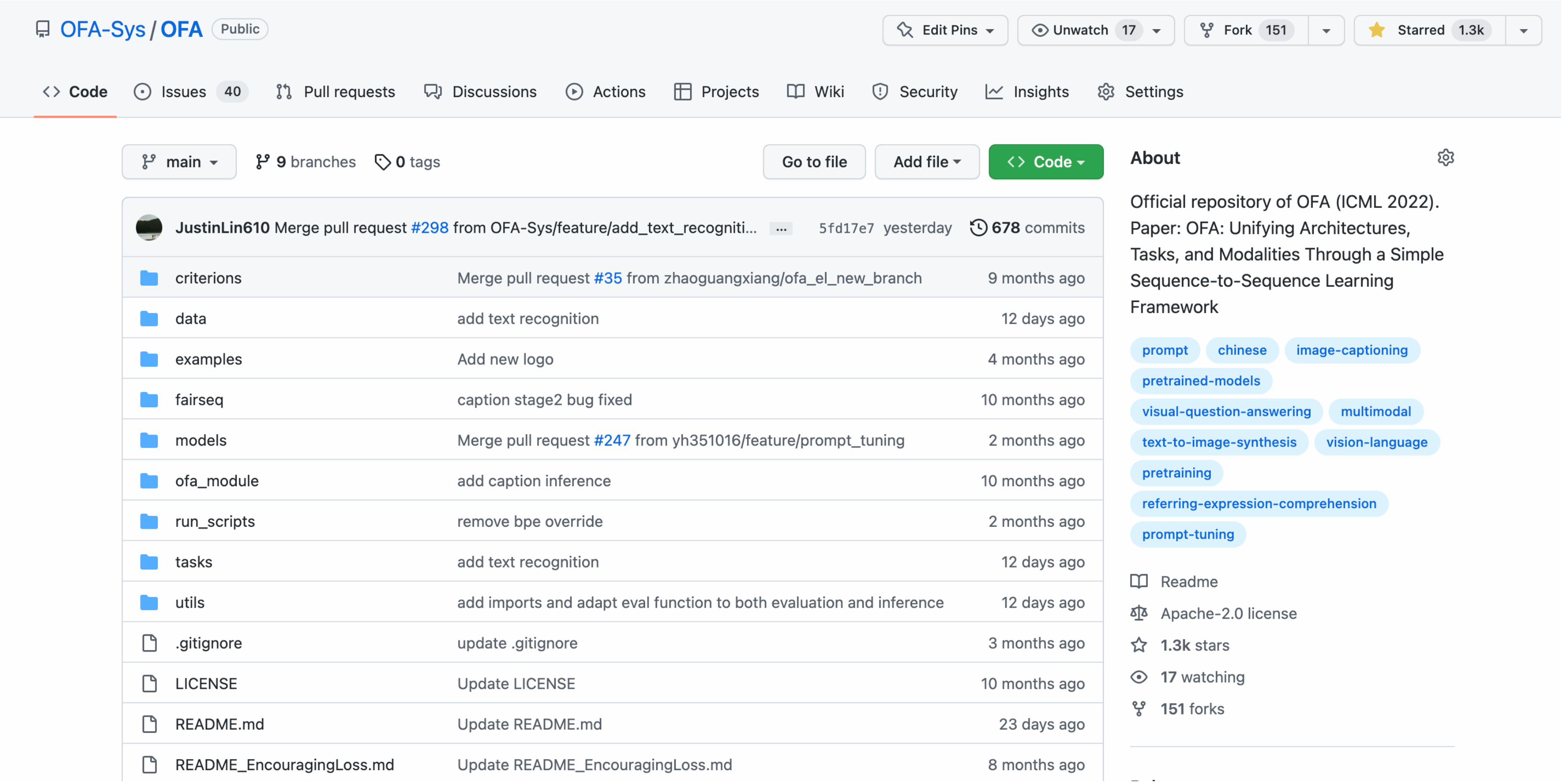
Model	BLEU@4	ROUGE-L	CIDEr-D
Trm	7.33	51.51	11.00
M6	16.19	55.06	30.75
OFA_{Base}	26.23	58.95	50.70
OFA_{Large}	**27.32**	**59.20**	**53.51**

RefCOCO-CN Series

Model	RefCOCO(val/testA/testB)
OFA_{Base}(random-init)	30.13/35.07/25.03
OFA_{Base}	82.18/86.07/**76.68**
OFA_{Large}	**82.84/86.54**/76.50

Opensource

https://github.com/OFA-Sys/OFA



Opensource

Pretraining

Below we provide methods for pretraining OFA.

▶ **1. Prepare the Dataset**

▶ **2. Pretraining**

Image Captioning

We provide procedures to reproduce our results of image captioning on our paper below.

▶ **1. Prepare the Dataset & Checkpoints**

▶ **2. Finetuning**

▶ **3. Inference**

Text-to-Image Generation

This part provides procedures for the finetuning and inference of text-to-image generation. See below.

▶ **1. Prepare the Dataset & Checkpoints**

▶ **2. Shuffle the Training Data**

▶ **3. Finetuning**

▶ **4. Inference**

Opensource

Image Captioning

We provide procedures to reproduce our results of image captioning on our paper below.

▼ **1. Prepare the Dataset & Checkpoints**

Download data (see datasets.md) and models (see checkpoints.md) and put them in the correct directory. The dataset zipfile `caption_data.zip` contains caption_stage1_train.tsv, caption_stage2_train.tsv, caption_val.tsv and caption_test.tsv. Each image corresponds to only 1 caption in `caption_stage1_train.tsv` and corresponds to multiple captions in other TSV files (about 5 captions per image). Each line of the dataset represents a caption sample with the following format. The information of uniq-id, image-id, caption, predicted object labels (taken from VinVL, not used), image base64 string are separated by tabs.

```
162365  12455   the sun sets over the trees beyond some docks.  sky&&water&&dock&&pole  /9j/4AAQSkZ
```

▼ **2. Finetuning**

Following previous standard practice, we divide the finetuning process of image captioning into two stages. In stage 1, we finetune OFA with cross-entropy loss on 4 NVIDIA-V100 GPUs with 32GB memory (expected to obtain ~139.5 CIDEr on the validation set at this stage). In stage 2, we select the best checkpoint of stage 1 and train with CIDEr optimization on 8 NVIDIA-V100 GPUs. **Note that CIDEr optimization is very unstable and requires careful hyperparameter tuning. If you encounter training errors in the stage2 finetuning, you can increase the batch size or reduce the learning rate. If neither of these works, you can directly set** `--freeze-resnet` **to freeze the inner states of batch normalization.**

```
cd run_scripts/caption
nohup sh train_caption_stage1.sh > train_stage1.out &  # stage 1, train with cross-entropy loss
nohup sh train_caption_stage2.sh > train_stage2.out &  # stage 2, load the best ckpt of stage1 and
```

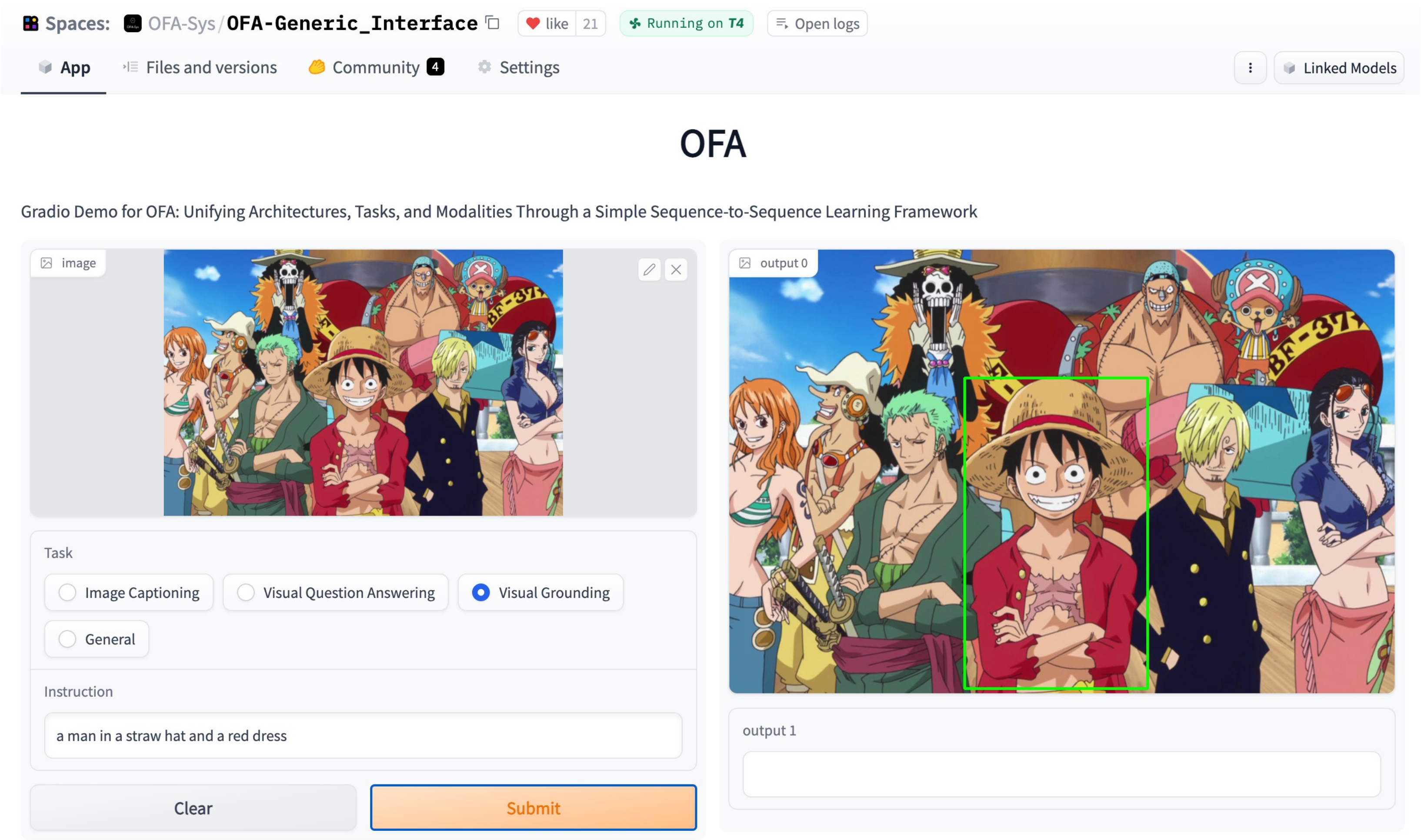
▼ **3. Inference**

Run the following commands to get your results and evaluate your model.

```
cd run_scripts/caption ; sh evaluate_caption.sh  # inference & evaluate
```

Demo

https://huggingface.co/spaces/OFA-Sys/OFA-Generic_Interface



Future Work

Future Work

- A Step forward:
  - A multimodal multitask system for extensive modality and task combinations
  - A single line of code to specify tasks and modalities
- Unified Models for Application
  - Small models matter!
  - More applications…

Thanks!

Github: https://github.com/OFA-Sys

Huggingface: https://huggingface.co/OFA-Sys

Papers

- OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. https://arxiv.org/abs/2202.03052

- Prompt Tuning for Generative Multimodal Pretrained Models. https://arxiv.org/abs/2202.03052