達摩院
ALIBABA DAMO ACADEMY

CVPR SEATTLE WASHINGTON
JUNE 16-18 2020

Visual-Linguistic Pre-training for Visual Question Answering

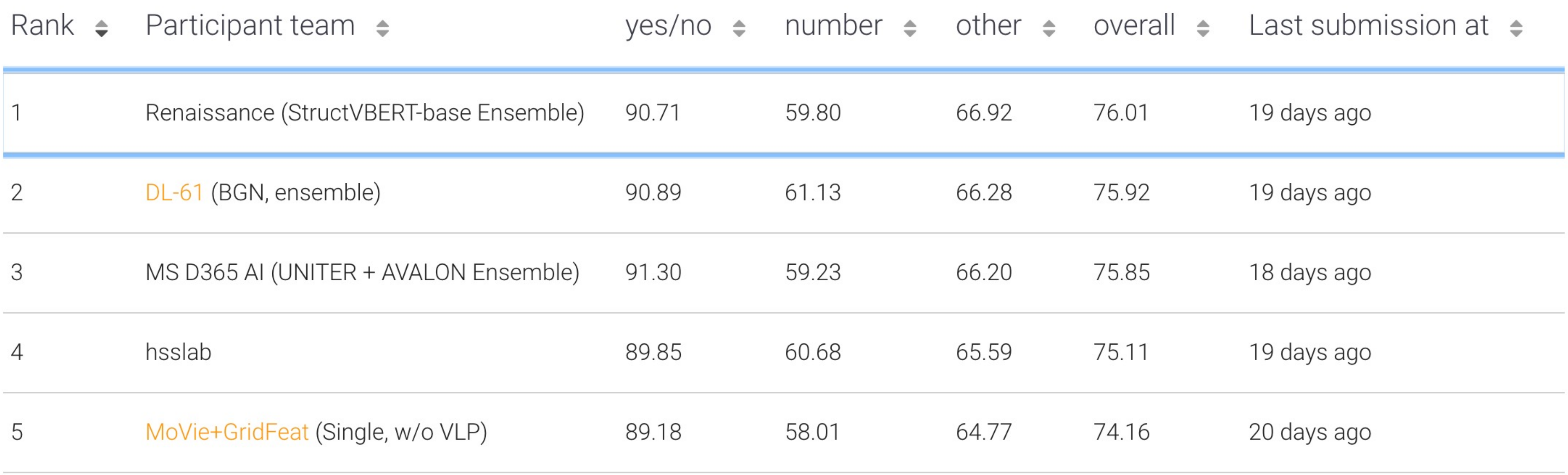
2020 VQA Challenge Runner-up

Team: Renaissance@DamoNLP

Ming Yan, Chenliang Li, Wei Wang, Bin Bi, Zhongzhou Zhao, Songfang Huang

Alibaba DAMO Academy

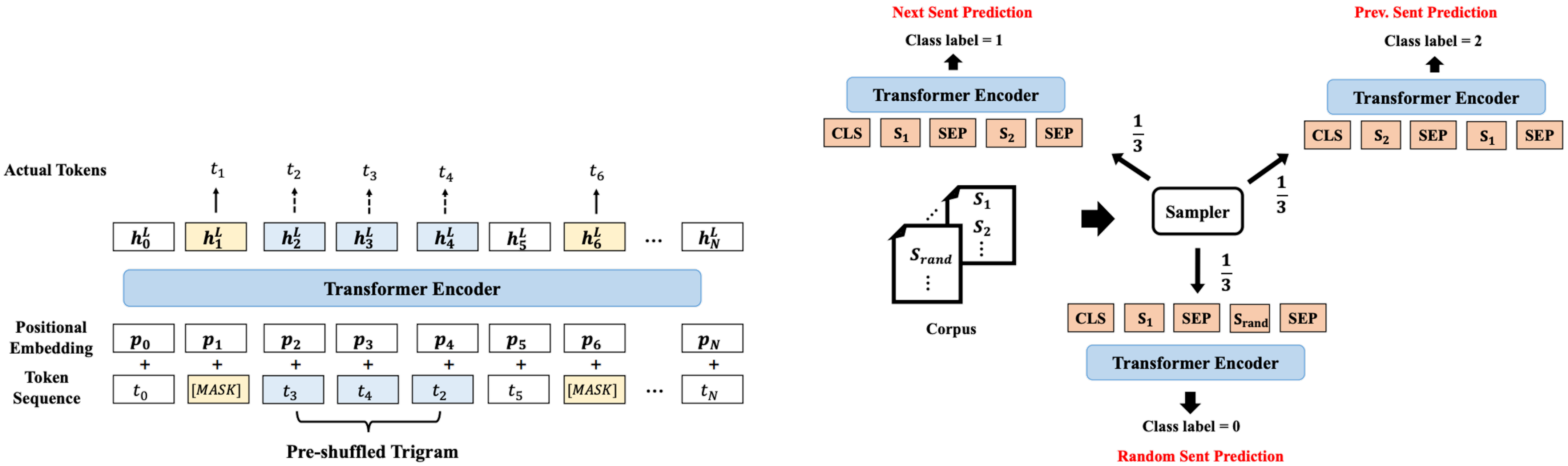
2020 VQA Leaderboard (Test-Standard)



Rank	Participant team	yes/no	number	other	overall	Last submission at
1	Renaissance (StructVBERT-base Ensemble)	90.71	59.80	66.92	76.01	19 days ago
2	DL-61 (BGN, ensemble)	90.89	61.13	66.28	75.92	19 days ago
3	MS D365 AI (UNITER + AVALON Ensemble)	91.30	59.23	66.20	75.85	18 days ago
4	hsslab	89.85	60.68	65.59	75.11	19 days ago
5	MoVie+GridFeat (Single, w/o VLP)	89.18	58.01	64.77	74.16	20 days ago

Language Model Pre-training

StructBERT

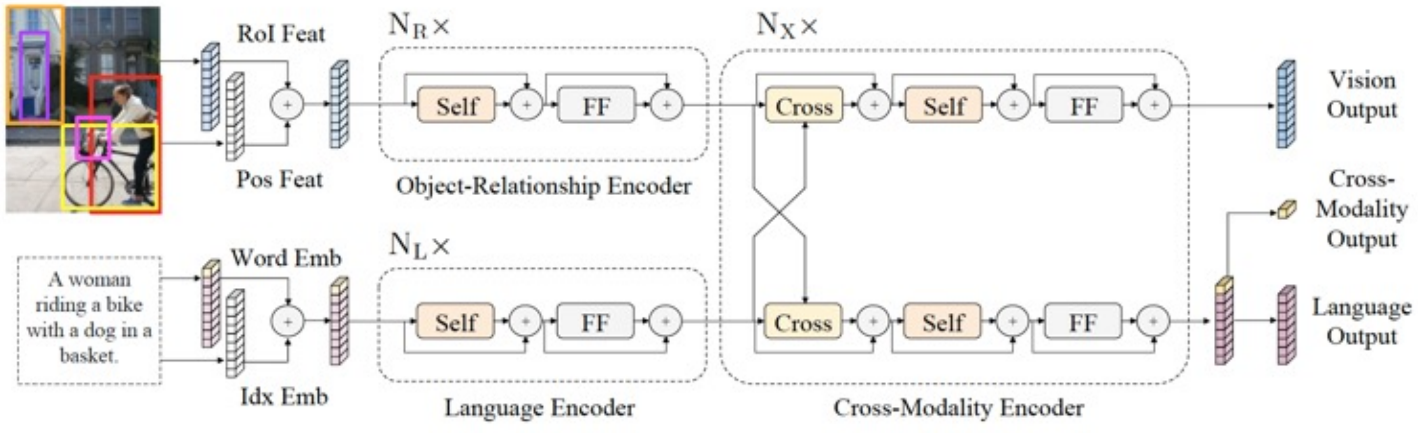


(a) Word Structural Objective

(b) Sentence Structural Objective

StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding, ICLR 2020

Visual-Linguistic Pre-training



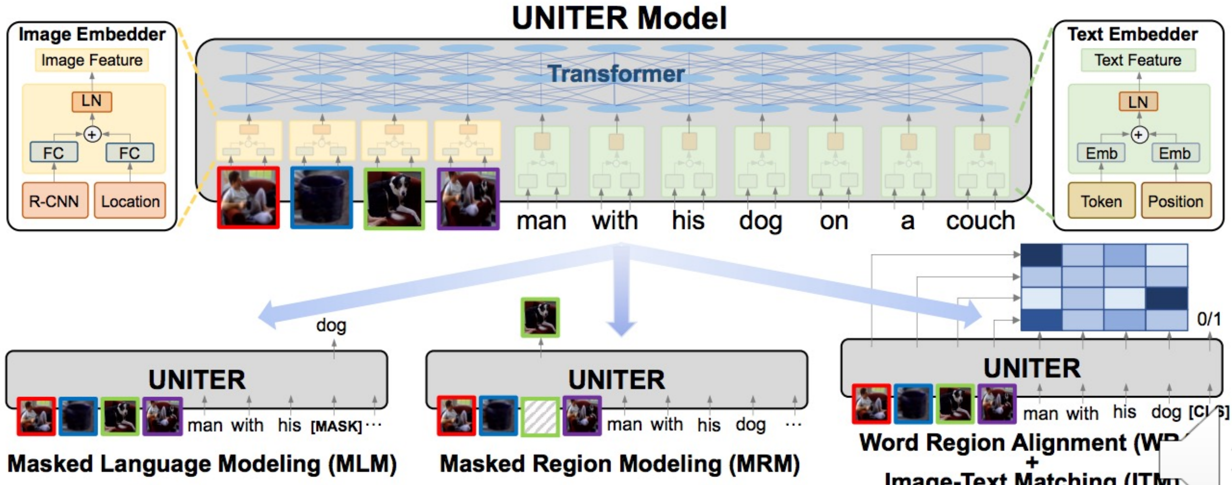
LXMERT

Two-stream Architecture

LXMERT: Learning Cross-Modality Encoder Representations from Transformers, EMNLP 2019

UNITER

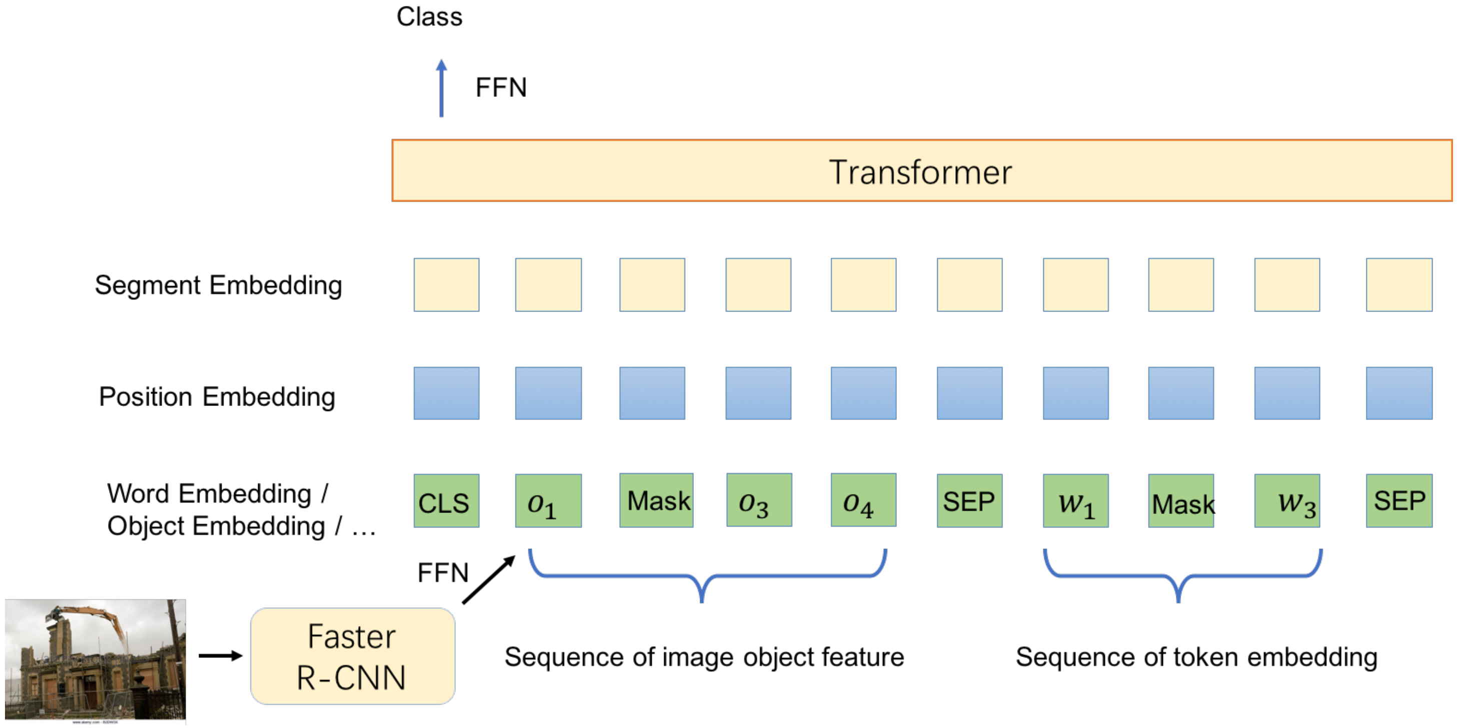
One-stream Architecture



UNITER: UNiversal Image-TExt Representation Learning

Our Model

- One-stream 12-layer Transformer (BERT base architecture)
  - Project the image feature and textual feature into the same semantic space
  - Pre-training with COCO caption, VG caption, VG QA, GQA, VQA dataset
  - Bottom up top down feature with faster rcnn

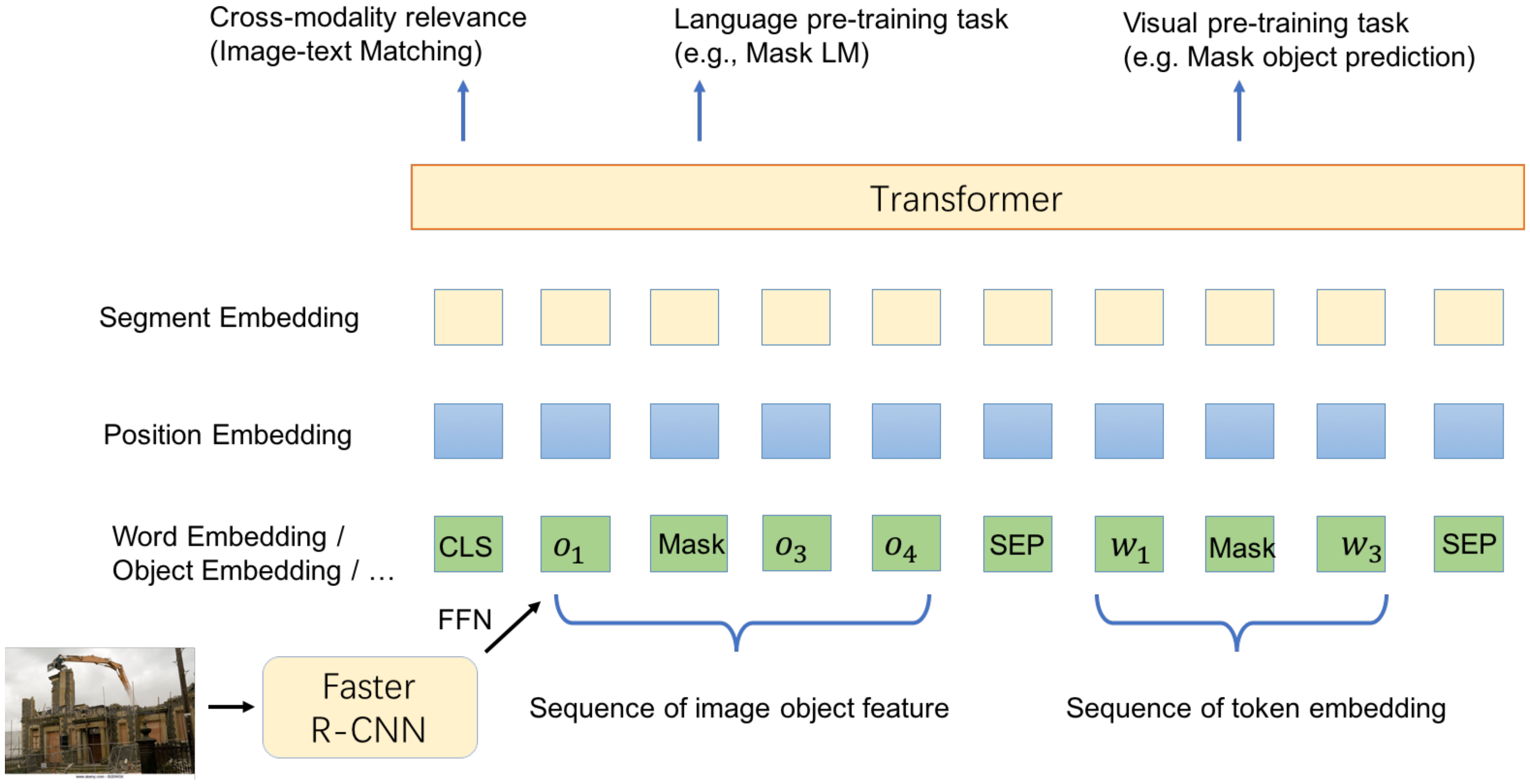


Main Techniques To Improve Performance

- ✓ One-stream v.s. Two-stream
- ✓ Pre-training Tasks (Multi-task Pre-training)
- ✓ Multi-stage Progressive Pre-training
- ✓ Model Ensembling

One-stream v.s. Two-stream

- One-stream can be better with less parameters
  - Masked Cross-Modality LM + ROI Feature Regression + Detected Object Classification + Detected Attribute Classification + Image-Text Matching



	# param	Performance
Two stream (LXMERT)	200M+	72.5
One stream Transformer (12layer base)	130M	**72.8**

Main Techniques To Improve Performance

✓ One-stream v.s. Two-stream

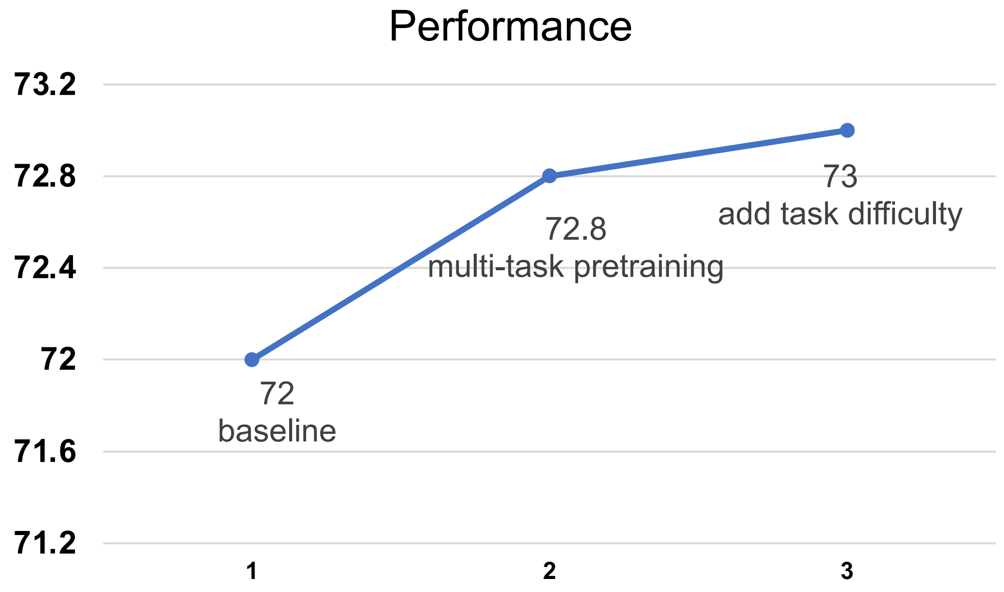
✓ Pre-training Tasks (Multi-task Pre-training)

✓ Multi-stage Progressive Pre-training

✓ Model Ensembling

Improve Pre-training Tasks

- Increase the task difficulty
  - **Language Modality:** use whole word masking to mask continuous word span
  - **Visual Modality:** also mask overlapped Image objects with significant overlap (> 0.5 IoU)
- Multi-task Pre-training
  - Pre-training with all the self-supervised tasks together
  - Add in-domain question-answering data in pre-training can further improve the performance

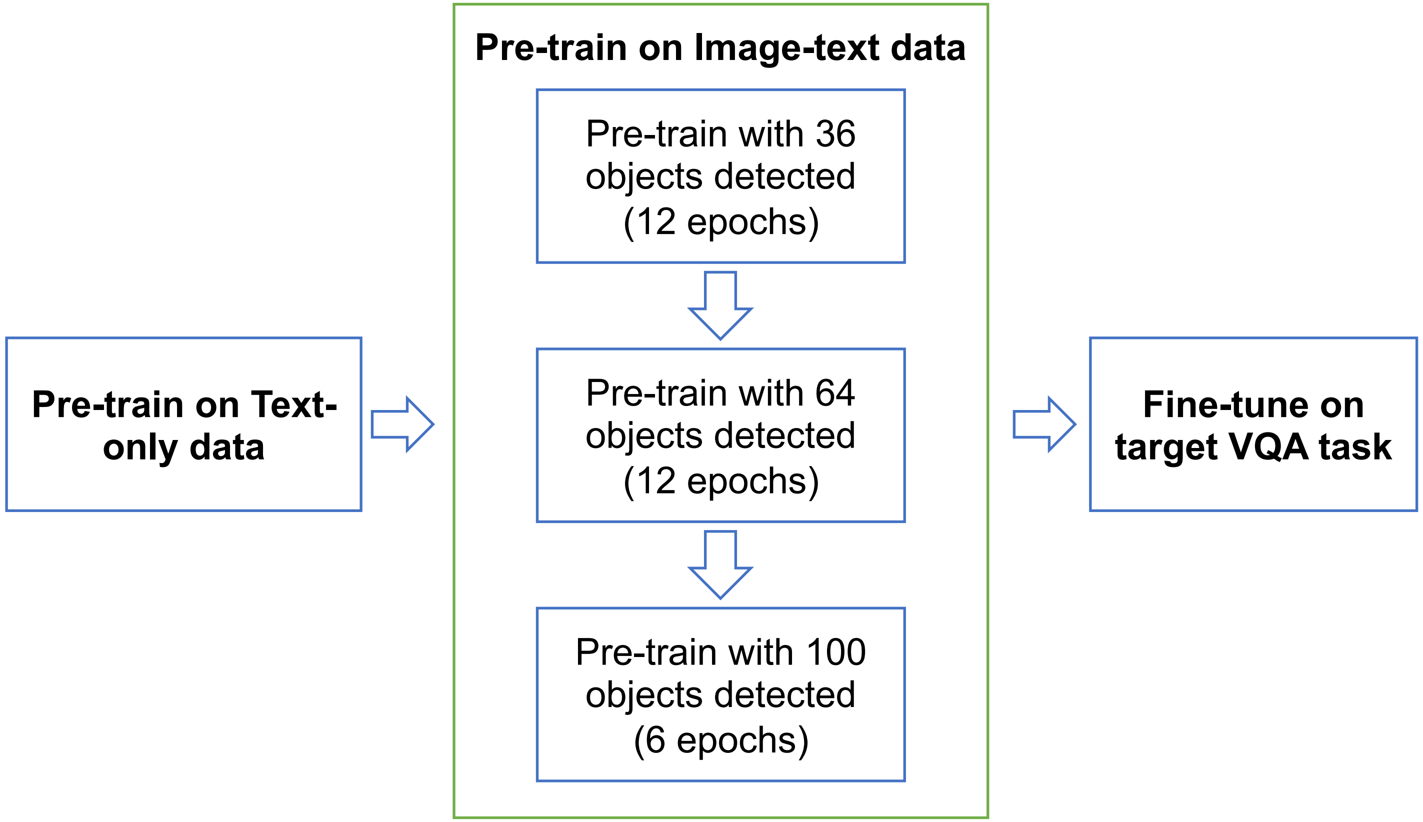


Main Techniques To Improve Performance

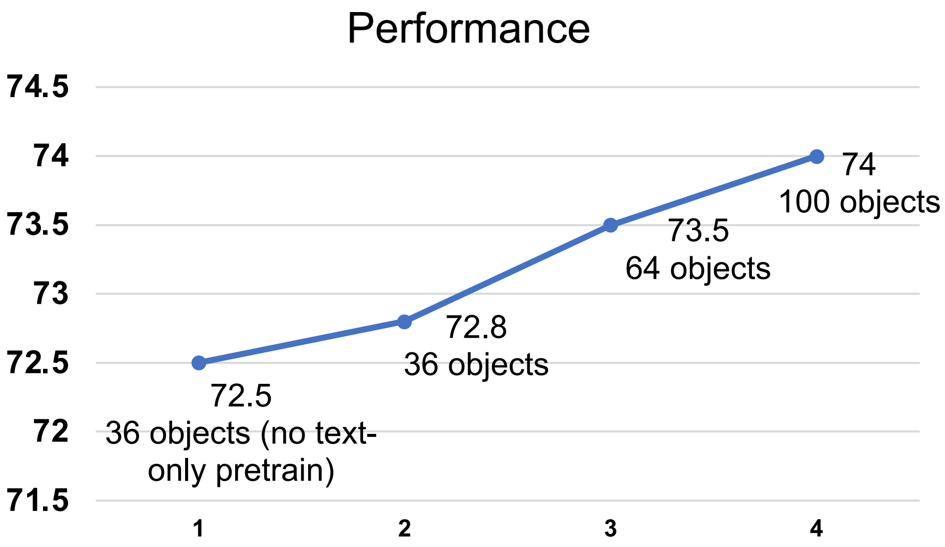
- ✓ One-stream v.s. Two-stream
- ✓ Pre-training Tasks (Multi-task Pre-training)
- ✓ Multi-stage Progressive Pre-training
- ✓ Model Ensembling

Multi-stage Progressive Pre-training

- Progressive Pre-training within Unified Visual-Semantic Space
  - **Horizontal:** first pre-train on text-only data (helpful in one-stream architecture), and then pre-train image-text pairs
  - **Vertical:** first pre-train on 36 objects, then pre-train on 64 objects and finally pre-train on 100 objects (use faster rcnn to detect more fine-grained objects can promote the performance)



Pre-train on Text-only data
Pre-train on Image-text data
Pre-train with 36 objects detected (12 epochs)
Pre-train with 64 objects detected (12 epochs)
Pre-train with 100 objects detected (6 epochs)
Fine-tune on target VQA task

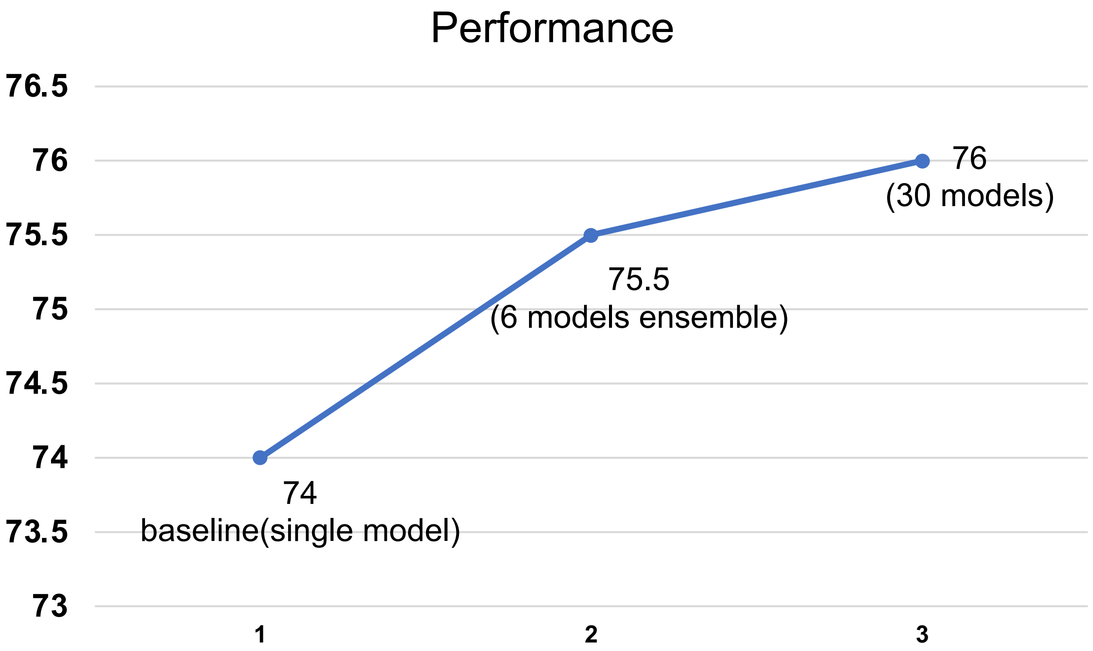


Main Techniques To Improve Performance

- ✓ One-stream v.s. Two-stream
- ✓ Pre-training Tasks (Multi-task Pre-training)
- ✓ Multi-stage Progressive Pre-training
- ✓ Model Ensembling

Model Ensembling

- Diverse Model Ensembling
  - **6 Diverse Models:** train with different object numbers (36 objects, 64 objects, 100 objects), one-stream architecture and two-stream architecture, without qa data in pre-training, MCAN model (w/o VLP)
  - **24 More Ensemble:** learning rate, seed, checkpoint, etc



Thank you!