Unified Image-Text Modeling

Zhengyuan Yang

Microsoft
Agenda

• Image-text tasks overview; Motivation of unification
• Unified image-text models
• Summary and discussion
Close-set Classification

What color is the plate?

Text feature

Image feature

Multi-Modal Fusion

red
blue
yellow
white
......
yes
no

Image-text matching, NLVR2

True
False

VCR

a) b) c) d)

Visual Entailment

Entailment Neutral Contradiction

Image credit: from the original papers: NLVR2, VCR, Visual Entailment

The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

Why is [person4] pointing at [person3]?

- He is telling [person4] that [person3] ordered the pancakes.
- He just told a joke.
- He is feeling accusatory towards [person3].
- He is giving [person3] directions.

Two women are holding packages. The sisters are hugging goodbye while holding to go packages after just eating lunch. The men are fighting outside a deli.

Premise

Hypothesis

Answer
Open-ended Text Sequence

Optional text input/ Empty → Image feature → Multi-Modal Fusion → Text feature → Auto-regressive decoder

“Open-vocab”: language model tokenizer vocabulary, e.g., 30522, 50265

This image is of a family celebrating Christmas. They are all gathered around a dinner table, with a turkey and other food on it. The family is smiling and seems to be enjoying themselves. There is a Christmas tree in the background and some Christmas lights on the walls.

What color is the plate? The plate is white.

A donut on a white plate next to a cup of latte.

A donut on a plate on a coffee.

Paragraph Captioning

Open-ended VQA

Image captioning

Image credit: Yujia Xie
Box/mask Localization

The donut on the white plate

Image feature

Text feature

Multi-Modal Fusion

$[x_1, y_1, x_2, y_2]$: [90.1, 83.2, 184.9, 180.4]

Visual grounding (REC, phrase grounding)

Language-based segmentation (RES)

Image credit: from the original papers: Flickr30K Entities, PhraseCut, DMS, Mask R-CNN
Pixel Prediction

A donut on a white plate next to a cup of latte.  

Text-to-image synthesis  
Text-based image editing

Optional image input/ Empty

Image feature

Multi-Modal Fusion

Text feature

Pixel values

"White leaves"  "Blue leaves"  "Yellow leaves"

"Remove bottom-right large red cube"

"This bird has wings that are black, and has a red belly and a red head"

Image credit: Jing Shi
Why Unified Image-Text Modeling

• Better performance
• New capabilities
• Task-agnostic unified systems
Why Unified Image-Text Modeling

- Better performance
  - Similar abilities; Multi-task training
  - Extra data/annotations from other tasks

![Diagram of multi-modal fusion with text and image features connected to a single node](image)
Why Unified Image-Text Modeling

• New capabilities

• E.g., text + box; grounded captioning

• More comprehensive and interpretable image description
Why Unified Image-Text Modeling

• Task-agnostic unified systems
  – Ease framework design; Avoid model copies
  – Capability generalization

A man in a black jacket and black pants is playing with a dog in a park.

Generalizing grounded captioning to COCO
Agenda

• Image-text tasks overview; Motivation
• **Unified image-text models**
  – Classification as text generation
  – Model design and training
  – Unify text and box
  – Textualize visual outputs
• Summary and discussion
VL Research

• Models curated for tasks

• Fast-forward to vision-language pre-training (VLP)

Image credit: from the original papers
Vision-language Pre-training (VLP)

- Large-scale transformer-based self-supervised pre-training
- Reuse the same pre-training weight as initialization point
- Separate output head and finetune model copies for different downstream tasks

Image credit: CVPR 2020 VL Tutorial; Attention Is All You Need, NeurIPS 2017
Vision-language Pre-training (VLP)

- Large-scale transformer-based self-supervised pre-training
- Reuse the same pre-training weight as initialization point
- Separate output head and finetune model copies for different downstream tasks

Image credit: CVPR 2020 VL Tutorial; UNITER: UNiversal Image-TExt Representation Learning, ECCV 2020
12-in-1

- Single model for 12 tasks (12*270M=3B -> 270M)
- Relationships among tasks; better averaged performance
- Task-specific heads and objectives

Image credit: 12-in-1: Multi-task vision and language representation learning, CVPR 2020
VL-T5

- Image-text tasks as multimodal conditional text generation
- Avoid task-specific arch design and model copies

Image credit: Unifying vision-and-language tasks via text generation, ICML 2021
VL-T5

- Image-text tasks as multimodal conditional text generation
- Avoid task-specific arch design and model copies

Image credit: Unifying vision-and-language tasks via text generation, ICML 2021
Agenda

- Image-text tasks overview; Motivation
- Unified image-text models
  - Classification as text generation
  - Model design and training
  - Unify text and box
  - Textualize visual outputs
- Summary and discussion
Model design and training

• Output format unification is the first step, how to have different tasks and capabilities work well together
  – Partially-shared parameters
  – Modular network design
  – Data and training techniques
Partially-shared Parameters

- Mixture of modality experts
- Task-specific parameters

Image credit: VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts
VL-Adapter: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks
Modular Network Design

- Unimodal encoders for single-modality tasks
- Reuse adjusted submodules for different tasks

Image credit: FLAVA: A Foundational Language And Vision Alignment Model
BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation
Data and Training Techniques

- Training corpus, batch construction
- Optimizing and loss design

\[
\sum_{m=1}^{M} \lambda_m \cdot D_m \left[ - \sum_{\ell=1}^{L} \log p(y_\ell | y_{<\ell}, x_{<\ell}) \right]
\]

\[
G_{\text{txt}} = \frac{\partial L_{\text{txt}}}{\partial \theta}, \quad G_{\text{img}} = \frac{\partial L_{\text{img}}}{\partial \theta}
\]

\[
G_{\text{global}} = M \odot G_{\text{txt}} + (1 - M) \odot G_{\text{img}}
\]

Image credit: ZeroVL: A Strong Baseline for Aligning Vision-Language Representations with Limited Resources
Flamingo: a Visual Language Model for Few-Shot Learning
Towards a unified foundation model: Jointly pre-training transformers on unpaired images and text
Agenda

• Image-text tasks overview; Motivation

• **Unified image-text models**
  – Classification as text generation
  – Model design and training
  – **Unify text and box**
  – Textualize visual outputs

• Summary and discussion
Good for Text Outputs, Others?

• Back to unifying I/O formats
• Output formats:
  ✓ Classification, text sequence
  ? Box/mask
  ? Pixel value
• Text+box as a case study
A donut on a white plate next to a cup of latte.

OD/grounding: white plate, donut, coffee mug

1. Support both outputs
2. Word-box alignments
• **Text and box outputs:** detector for image -> regions

• **Word-box alignments:** region index prediction
  
  Related to the modeling in region-based VL models, but with detector E2E finetuned

Image credit: Towards General Purpose Vision Systems, CVPR 2022
MDETR, UniT

- Avoid explicit detection module?
- **Text and box outputs:** (box): coordinate regression; (text): heads for classification output
- **Word-box alignments:** input word index, or OD vocab

Image credit: MdeTR-modulated detection for end-to-end multi-modal understanding, ICCV 2021

UniT: Multimodal Multitask Learning with a Unified Transformer, ICCV 2021
FIBER

- Challenge: resolution, computing cost trade-offs
- Coarse-to-fine two-stage vision-language pre-training
- Text and box outputs; Fusion in-the-backbone

Image credit: Coarse-to-Fine Vision-Language Pre-training with Fusion in the Backbone
Agenda

• Image-text tasks overview; Motivation
• **Unified image-text models**
  – Classification as text generation
  – Model design and training
  – Unify text and box
  – **Textualize visual outputs**
• Summary and discussion
Textualize Visual Outputs

- Visual (box, mask, pixel) outputs and text outputs (cls, text) are often modeled differently and require different modules:
  - Object detector (OD)
  - Coordinate regression head
- A single model that unifies text and box(visual) outputs?

<table>
<thead>
<tr>
<th>Representative Models</th>
<th>Visual Modeling</th>
<th>Text Output</th>
<th>Box Output</th>
<th>Word-box Align</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViLBERT [12], OSCAR [38], UNITER [12], VinVL [73], etc. [37,59,35,58,78,43]</td>
<td>Offline OD</td>
<td>Task-specific Heads</td>
<td>Region Index</td>
<td>×</td>
</tr>
<tr>
<td>PixelBERT [29], SOHO [28], ViLT [33], SimVLM [64], etc. [56,36,68,19]</td>
<td>Image Patches</td>
<td>Task-specific Heads</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>GPV [23]</td>
<td>Online OD</td>
<td>Single Output Seq.</td>
<td>Region Index</td>
<td>Extra Prediction</td>
</tr>
<tr>
<td>MDERT [31]</td>
<td>Image Patches</td>
<td>Task-specific Heads</td>
<td>Box Coordinate</td>
<td>×</td>
</tr>
<tr>
<td>UniT [26]</td>
<td>Image Patches</td>
<td>Task-specific Heads</td>
<td>Box Coordinate</td>
<td>×</td>
</tr>
<tr>
<td>UniTAB (Ours)</td>
<td>Image Patches</td>
<td>Single Output Seq.</td>
<td>Box Coordinate</td>
<td>Inline Indicated</td>
</tr>
</tbody>
</table>
UNICORN

• Textualize bounding box for object detection [1]
• Text and box outputs: Unified text+box decoding vocabulary
• Word-box alignments: in-line in output sequence

UNICORN. Crossing the Format Boundary of Text and Boxes: Towards Unified Vision-Language Modeling

[1] Pix2seq: A language modeling framework for object detection, ICLR 2022
Model and Training

- Encoder-decoder architecture
- Single LM objective

\[ \mathcal{L}_{LM}(\theta) = - \sum_{t=1}^{T} \log P_\theta(s_t | s_{<t}, v, l) \]

Image credit: UNICORN. Crossing the Format Boundary of Text and Boxes: Towards Unified Vision-Language Modeling
Unifying Different VL Tasks

- Textualized outputs: text, box, alignment
- Multi-task finetuning, capability generalization
Capability Generalization

- **MSCOCO; Grounded description**
- For eval: text, box, alignment
- Metrics: captioning

- **ImageNet; Object localization**
- For eval: text, box, alignment
- Metrics: accuracy
Textualize Visual Outputs

- OFA
- Image tokens for pixel outputs
Textualize Visual Outputs

- **Pix2Seq-V2**
- **Masks as polygon, Keypoints**

Image credit: A Unified Sequence Interface for Vision Tasks
Textualize Visual Outputs

- Unified-IO

all vision tasks (seg., depth, surface, etc.) that require dense prediction as (conditional) image generation
Agenda

• Image-text tasks overview; Motivation
• Unified image-text models
• Summary and discussion
Take-away Messages

• Unified image-text modeling from the view of I/O format
• Textualized visual outputs for unified image-text modeling
• Format unification is the first step, improving unified models
• Grand vision of general-purpose visual understanding
Challenges and Future Directions

• How to better show the advantage of unified models
  – Relationship among tasks; Gain from MTL
  – New capability showcase; Generalization setups

• How to better train the unified models
  – Balance the degree of unification
  – Better ways of format unification

• Foundation models at what granularity
Thank you!
Any Questions?