Unifying Vision-and-Language Tasks via Text Generation

Jaemin Cho, Jie Lei, Hao Tan, Mohit Bansal
UNC Chapel Hill

Presenters:
Siba Smarak Panigrahi

Reading Session X
Kharagpur Data Analytics Group
IIT Kharagpur

August 07, 2021

Important Notes

- We expect you all to have certain queries regarding the presentation, certain intricate doubts maybe... **Please put them in the chat of this meeting (if you feel shy) else unmute and speak**.
- Finally, a **Google Form** [Link] will be released for feedback but most importantly, we ask you to put up name of any AI-related paper to be presented in the upcoming sessions!

Link to GitHub Repo : Click Here

Link to join Slack Workspace : Click Here

Link to KDAG YouTube Channel : Click Here

Link to doc on good AI Blogs/Resources/Topics : Click Here

Major Contribution

A unified framework (VL-T5 & VL-BART):

- learns different tasks in a single architecture with **multimodal conditional text generation**
- learn to generate labels "as texts" on the visual and textual inputs

Major Contribution

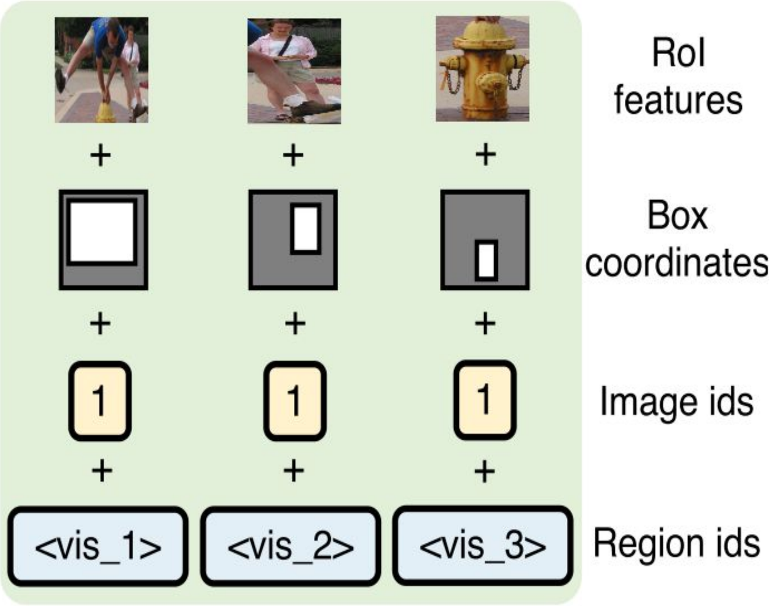
A unified framework (VL-T5 & VL-BART):

- learns different tasks in a single architecture with **multimodal conditional text generation**
- learn to generate labels “as texts” on the visual and textual inputs
- Earlier modeled as discriminative tasks, this paper provides a **generative approach**
- Generation of **open-ended natural language answers**, whereas with discriminative tasks we obtained one answer out of the fixed set of options

Model

- Visual Embeddings
  - Region of Interest (RoI) object features
  - RoI bounding box coordinates
  - Image ids (# of different images in input)
  - Region ids

$$e^v = \{e^v_1, \ldots, e^v_n\}$$



Final Visual embeddings:

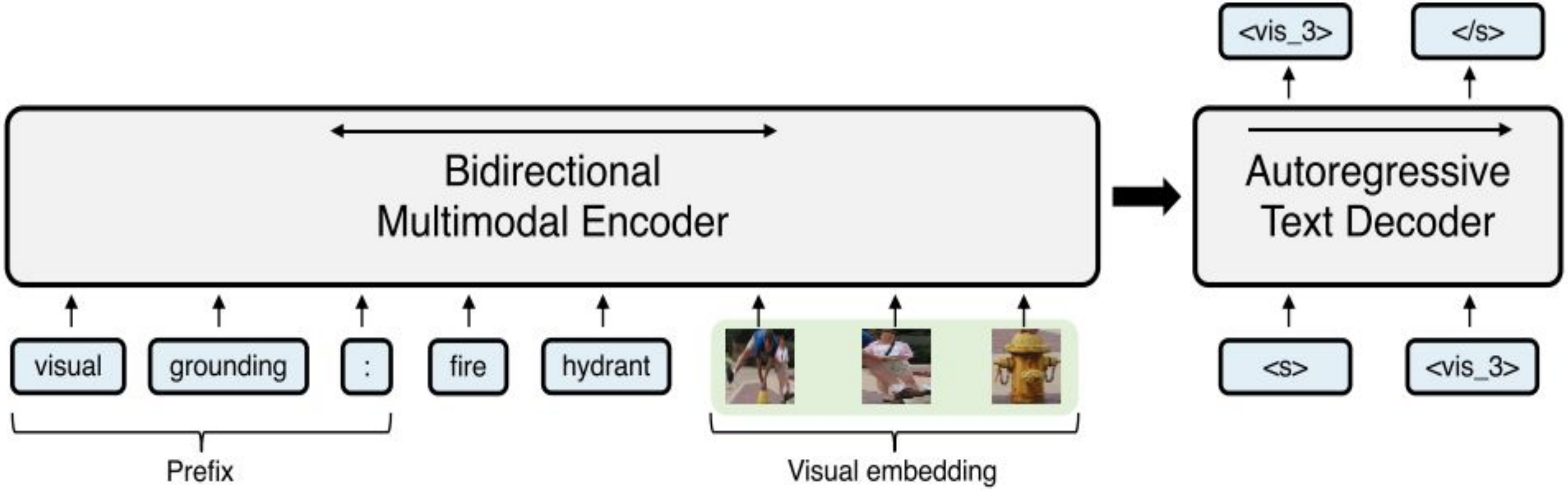
Model

- Text Embeddings
    - **Different prefixes** for different tasks
    - Addition of visual sentinel tokens: {<vis_1>, <vis_2>, ... <vis_n>}

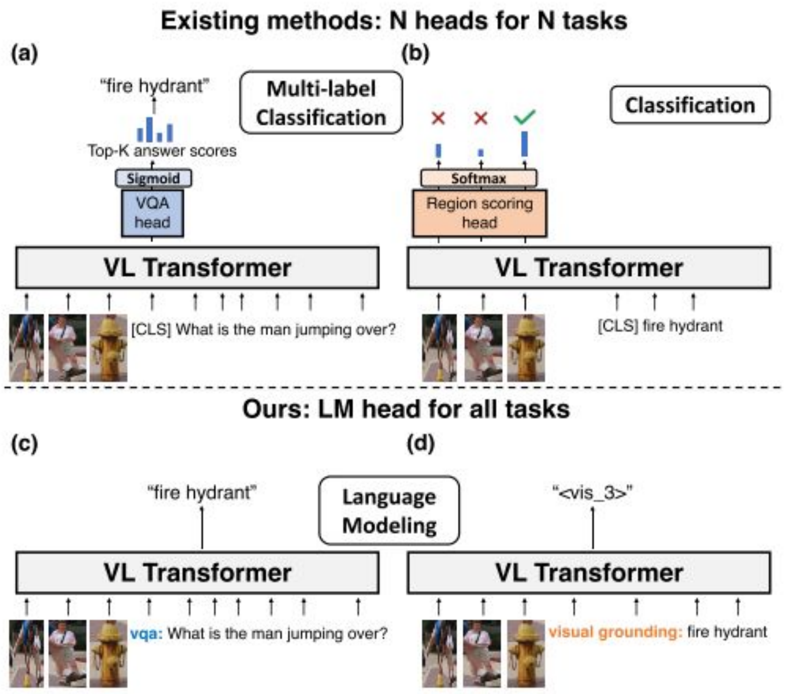
$$e^x = \{e_1^x, \ldots, e_{|x|}^x\}$$

    - Augmented text **x** encoded as learned embedding (after tokenization)

Framework for “visual-grounding task”



Comparison between generative and discriminative architectures

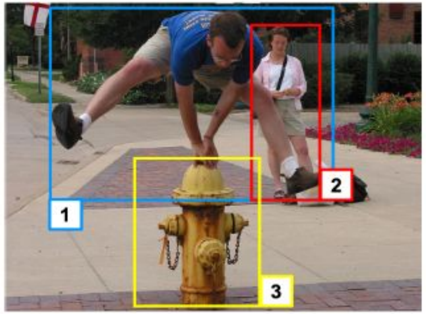


Model

- Encoder-Decoder architecture
  - **Encoder:**
    - Bi-directional, multimodal
    - Stacked m-transformer blocks; self-attention
    - fully connected layer + residual connections
  - **Decoder:**
    - similar to encoder with additional cross-attention layer in each block

$$\mathcal{L}_{\theta}^{\mathrm{GEN}} = -\sum_{j=1}^{|y|} \log P_{\theta}(y_j | y_{<j}, x, v)$$

Input-Output format for different tasks in pre-training

Tasks	Input image	Input text	Target text
Pretraning tasks (Sec. 4)			
Multimodal LM (VL-T5)		span prediction: A <text_1> is <text_2> over a fire hydrant.	<text_1> man <text_2> jumping
Multimodal LM (VL-BART)		denoise: A <mask> is <mask> over a fire hydrant.	A man is jumping over a fire hydrant
[a]Visual question answering		vqa: what is the color of the man's shirt?	blue
Image-text matching		image text match: A man with blue shirt is jumping over fire hydrant.	true
Visual grounding		visual grounding: yellow fire hydrant	<vis_3>
Grounded captioning		caption region: <vis_3>	yellow fire hydrant
Downstream tasks (Sec. 5)			
VQA		vqa: [Q]	[A]
GQA		gqa: [Q]	[A]
[b]NLVR^2		nlvr: [text]	true/false
VCR Q→A		vcr qa: question [Q] answer: [A]	true/false
VCR QA→R		vcr qar: question [Q] answer: [A] rationale: [R]	true/false
RefCOCOg		visual grounding: [referring expression]	[region id]
COCO captioning		caption:	[caption]
COCO captioning (w/ object tags)		caption with tags: [Tag1 Tag2 ..]	[caption]
Multi30K En-De translation		translate English to German: [English text]	[German text]

Results on downstream tasks

Visual Question Answering

Method	In-domain	Out-of-domain	Overall
Discriminative			
$\text{UNITER}_{\text{Base}}$	**74.4**	10.0	**70.5**
VL-T5	70.2	7.1	66.4
VL-BART	69.4	7.0	65.7
Generative			
VL-T5	71.4	13.1	67.9
VL-BART	72.1	**13.2**	68.6

- In-domain performance is comparable, while the out-of-domain performance is significantly higher.

Results on downstream tasks

Visual Question Answering

Method	In-domain	Out-of-domain	Overall
Discriminative			
$\text{UNITER}_{\text{Base}}$	**74.4**	10.0	**70.5**
VL-T5	70.2	7.1	66.4
VL-BART	69.4	7.0	65.7
Generative			
VL-T5	71.4	13.1	67.9
VL-BART	72.1	**13.2**	68.6

- In-domain performance is comparable, while the out-of-domain performance is significantly higher.
- Also show that a single model can successfully handle multiple VQA tasks without dataset-specific prefixes

Results on downstream tasks

Natural Language Visual Reasoning (NLVR)

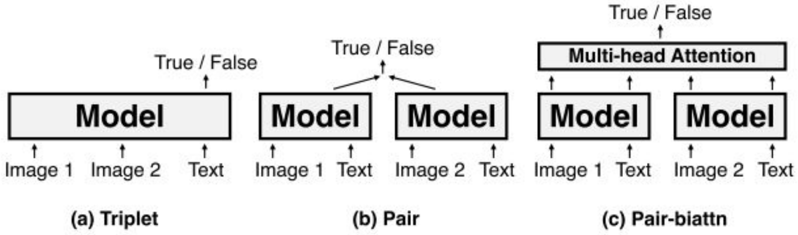
Method	Setting	dev	test-P
UNITER_{Base}	Triplet	73.0	73.9
UNITER_{Base}	Pair	75.9	75.8
UNITER_{Base}	Pair-biattn	**77.2**	**77.9**
LXMERT	Pair	74.9	74.5
Oscar_{Base}	Pair	78.1	78.4
VL-T5	Triplet	74.6	73.6
VL-BART	Triplet	71.7	70.3

- Similar performance in triplet setting (which is half the computational cost of other settings)

Results on downstream tasks

Natural Language Visual Reasoning (NLVR)

Method	Setting	dev	test-P
UNITER_{Base}	Triplet	73.0	73.9
UNITER_{Base}	Pair	75.9	75.8
UNITER_{Base}	Pair-biattn	**77.2**	**77.9**
LXMERT	Pair	74.9	74.5
Oscar_{Base}	Pair	78.1	78.4
VL-T5	Triplet	74.6	73.6
VL-BART	Triplet	71.7	70.3



- Similar performance in triplet setting (which is half the computational cost of other settings)

Results on downstream tasks

Referring Expression Comprehension: RefCOCOg

Method	V&L PT	val^d	test^d
MattNet		66.9	67.3
UNITER$_{Base}$	✓	**74.3**	**74.5**
VL-T5		63.4	62.9
VL-T5	✓	71.2	71.3
VL-BART		21.8	23.0
VL-BART	✓	23.6	22.4

- With pre-training VL-T5 reaches similar performance similar with UNITER-base.
- **Poor performance of TL-BART**: BART adds learned absolute positional embedding to text token embedding, whereas T5 uses relative position biases in self-attention layers instead

Results on downstream tasks

Multimodal Machine Translation: Multi30K

Method	V&L PT	test2016	test2017	test2018
MSA		38.7	-	-
MeMAD		38.9	32.0	-
MSA†		39.5	-	-
MeMAD†		45.1	40.8	-
MeMAD†*		45.5	41.8	38.5
T5 (text only)		44.6	41.6	39.0
VL-T5		45.3	**42.4**	**39.5**
VL-T5	✓	**45.5**	40.9	38.6
BART (text only)		41.2	35.4	33.3
VL-BART		41.3	35.9	33.2
VL-BART	✓	37.7	29.7	28.1

- Best performance across all the test-sets.
- **Vision & Language Pre-training** didn't help: the source text contains sufficient information for translation

Multi-task fine-tuning

Single-task vs. Multi-task Fine-tuning

Method	Finetuning tasks	# Params	Discriminative tasks					Generative tasks	
			VQA Karpathy test Acc	GQA test-dev Acc	NLVR^2 test-P Acc	RefCOCOg test^d Acc	VCR val Acc	COCO Caption Karpathy test CIDEr	Multi30K En-De test2018 BLEU
VL-T5	single task	7P	67.9	60.0	73.6	71.3	57.5	116.1	38.6
VL-T5	all tasks	P	67.2	58.9	71.6	69.4	55.3	110.8	37.6

- Fine-tune a single VL-T5 for 20 more-epochs; tackle 7 tasks simultaneously.
- **Multi-task model** achieves **comparable performance** to the **separately optimized single-task models** on all 7 tasks with a single set of parameters

Multi-task fine-tuning

Single shared head vs. Task-specific heads

Method	# Params	VQA Karpathy test Acc	GQA test-dev Acc	COCO Caption Karpathy test CIDEr
Single shared head	P	68.3	59.3	110.6
Task-specific heads	P+7H=1.8P	68.5	59.3	110.9

- 7 additional task-specific heads are added for each downstream tasks
- With **fewer parameters**, the single shared head has **similar performance** to task specific heads

Conclusion

- VL-T5 and VL-BART to tackle vision-and-language tasks with a **unified text generation objective**, a single architecture to have fewer parameters and without losing much performance.
- VL-T5 and VL-BART can **achieve comparable performance** with state-of-the-art vision-and-language transformers on diverse vision-and-language tasks **without hand-crafted architectures and objectives**
- Generative approach is better suited for open-ended visual question answering.

Thank You