Paper Sharing

袁凡

Multi-modal Dialogue

- 按功能分类：

  闲聊型、任务型、知识问答型和推荐型

- 按模式分类：

  基于检索模式和生成式模式

- 按领域分类：

  开放领域和封闭领域

- 单轮对话和多轮对话

Datasets

- Wikipedia
- Reddit
- msCOCO
- Flickr30k
- Image-chat
- OpenViDial
- MELD
- ……

Multi-Modal Open-Domain Dialogue

多模态开放领域对话

Kurt Shuster∗, Eric Michael Smith∗, Da Ju, Jason Westo

Introduction

- Much recent work has explored building and training dialogue agents that can blend different conversational skills throughout natural conversation, with the ultimate goal of providing an engaging and interesting experience for humans.

- In order to better approach human-like ability, however, it is necessary that agents can converse with both textual and visual context, similarly to how humans interact in the real world.

- Recent efforts have gone beyond classical, fact-based tasks such as image captioning or visual question answering to produce models that can respond and communicate about images in the flow of natural conversation.

- **In this paper, they explore the extension of large-scale conversational agents to image-based dialogue.**

Related Work

- Multi-Modal Models and Tasks
  - Rich Representations:
    - use standard Transformer-based models to jointly encode text and images
    - explore modifications to the standard self-attention scheme in Transformers by incorporating additional co-attention or crossattention layers
  - Visual Dialogue/Caption Generation:
    - (COCO Captions and Flickr30k) Sequence-to-sequence and retrieval-based models
    - (Visual Dialog) text generation
    - (Image-Chat and Image-grounded Conversations) Sequence-to-sequence and retrieval-based models

- Multi-Task Training / Using Pre-Trained Representations
  - adapting pre-trained representations to later downstream tasks has been shown to be successful in NLP
  - employ multi-task training, to both help generalize the applicability of the model and improve its performance on downstream tasks/evaluations

Model Architectures

- Image Encoders
  - use two different image encoders to determine the best fit for our tasks:
    1. ResNeXt WSL: output is 2048-dimensional vector
    2. ResNeXt WSL Spatial:2048×7×7-dimensional vector
    3. Faster R-CNN:2048×100-dimensional representations

- Multi-Modal Architecture

  - Transformer model has 2 encoder layers, 24 decoder layers, 2560-dimensional embeddings, and 32 attention heads, and the weights are initialized from a 2.7-billion parameter model pre-trained on 1.5B comments from a third-party Reddit dump hosted by pushshift.io

  - two possible fusion schemes:

    1. Late fusion: encoded image is projected to the same dimension as the text encoding, concatenated with this output and fed together as input to the decoder

    2. Early fusion: concatenate the projected image encoding with the token embeddings, assign each a different segment embedding, and jointly encode the text and image in the encoder

Training detail

- fix the weights of the pre-trained image encoders, except the linear projection to the Transformer output dimension, and fine-tune all of the weights of the Transformer encoder/decoder

- Datasets

  Domain-Adaptive Pre-Training： COCO & third-party Reddit dump

  Fine-tuning: ConvAI2, EmpatheticDialogues (ED), Wizard of Wikipedia (WoW), BlendedSkillTalk (BST), Image-Chat (IC)

Experiment

- Automatic Evaluations

  1. Results on Pre-Training Datasets: Training datasets, Image Features, Image Fusion

Image Features	Image Fusion	COCO (ppl)	pushshift.io Reddit (ppl)	Average
COCO & pushshift.io Reddit training data				
ResNeXt WSL	Late	11.11	13.80	12.45
	Early	6.69	13.50	10.10
ResNeXt WSL Spatial	Late	7.43	13.00	10.22
	Early	6.53	13.46	10.00
Faster R-CNN	Late	5.26	13.17	9.21
	Early	5.23	13.15	**9.13**
COCO training data only				
ResNeXt WSL	Late	5.82	19.52	12.67
	Early	6.21	21.30	13.76
ResNeXt WSL Spatial	Late	6.51	16.50	11.51
	Early	6.19	18.77	12.48
Faster R-CNN	Late	5.21	17.88	11.55
	Early	**4.83**	18.81	11.82

Table 1: Model performance, measured via perplexity on validation data, on domain-adaptive pre-training datasets, comparing various image features and image fusion techniques. The top three rows involve multi-task training on COCO Captions and pushshift.io Reddit, while the bottom three rows involve single task training on COCO Captions only. We note that early fusion with Faster R-CNN features yields the best performance on COCO Captions.

Experiment

- Automatic Evaluations

  2. Results on Fine-Tuned Datasets:

Image Features	Training Data	Image Fusion	ConvAI2	ED	WoW	BST	IC 1st Turn	IC	Text Avg.	All Avg.
None	None	None	12.31	10.21	13.00	12.41	32.36	21.48	11.98	13.88
	BST^+		8.74	8.32	8.78	10.08	38.94	23.13	8.98	14.76
	BST^+ + IC		8.72	8.24	8.81	10.03	16.03	13.21	**8.95**	9.83
ResNeXt WSL	BST^+ + IC	Late	8.71	8.25	8.87	10.09	16.20	13.27	8.98	9.84
	BST^+ + IC	Early	8.80	8.32	8.79	10.17	15.16	12.99	9.02	9.81
	BST^+ + IC + COCO + Reddit	Late	9.27	8.87	9.45	10.74	17.56	14.44	9.58	10.56
	BST^+ + IC + COCO + Reddit	Early	9.34	8.90	9.48	10.78	15.87	13.88	9.62	10.48
	BST^+ + IC + COCO	Late	8.79	8.36	9.00	10.21	16.00	13.31	9.09	9.93
	BST^+ + IC + COCO	Early	8.91	8.38	8.99	10.29	14.64	12.85	9.14	9.88
ResNeXt WSL Spatial	BST^+ + IC	Late	8.71	8.24	8.88	10.10	15.39	13.02	8.98	9.78
	BST^+ + IC	Early	8.79	8.29	8.92	10.15	15.34	13.02	9.04	9.83
	BST^+ + IC + COCO + Reddit	Late	8.76	8.31	8.88	10.14	15.20	13.04	9.02	9.83
	BST^+ + IC + COCO + Reddit	Early	9.30	8.82	9.46	10.76	15.67	13.79	9.56	10.43
	BST^+ + IC + COCO	Late	8.73	8.31	8.87	10.13	15.04	12.98	9.01	9.84
	BST^+ + IC + COCO	Early	8.81	8.34	8.99	10.22	14.76	12.87	9.09	9.80
Faster R-CNN	BST^+ + IC	Late	8.70	8,24	8.92	10.07	13.97	12.48	8.98	**9.68**
	BST^+ + IC	Early	8.81	8.33	8.81	10.15	13.66	12.43	9.03	9.71
	BST^+ + IC + COCO + Reddit	Late	8.75	8.31	8.93	10.14	13.83	12.49	9.03	9.73
	BST^+ + IC + COCO + Reddit	Early	8.78	8.31	8.85	10.15	**13.51**	**12.36**	9.02	9.69
	BST^+ + IC + COCO	Late	8.74	8.33	8.87	10.13	13.85	12.51	9.02	9.72
	BST^+ + IC + COCO	Early	8.81	8.34	8.93	10.19	13.57	12.39	9.07	9.73

Image Features	Training Data	Image Fusion	IC First Turn	IC
None	None	None	32.36	21.48
	Image Chat		28.71	13.17
ResNeXt WSL	IC	Late	14.80	12.83
	IC	Early	16.00	13.21
	IC + COCO + Reddit	Late	16.73	13.92
	IC + COCO + Reddit	Early	15.71	13.53
	IC + COCO	Late	14.70	12.95
	IC + COCO	Early	14.62	12.92
ResNeXt WSL Spatial	IC	Late	15.34	13.01
	IC	Early	15.27	13.00
	IC + COCO + Reddit	Late	15.09	12.95
	IC + COCO + Reddit	Early	15.55	13.50
	IC + COCO	Late	15.02	12.95
	IC + COCO	Early	14.62	12.87
Faster R-CNN	IC	Late	13.99	12.51
	IC	Early	13.76	12.42
	IC + COCO + Reddit	Late	13.75	12.43
	IC + COCO + Reddit	Early	**13.44**	**12.29**
	IC + COCO	Late	13.82	12.48
	IC + COCO	Early	13.56	12.37

Experiment

- Human Evaluations

  1. Human/Model Chats Without Images

	MMB Style	BlenderBot
Contradiction	2.15%	3.37%
Improper English	0.27%	0.26%
Repetitive	1.34%	1.55%
Unrelated	2.42%	2.33%
Non-Sensical	4.03%	2.07%
None (All Good)	91.13%	91.45%
Mean engagingness	4.70±0.60	4.70±0.60

Table 6: Per-turn annotations and mean engagingness ratings of human/model conversations with MMB Style and BlenderBot. Both models perform roughly equivalently on these metrics. Ranges given are plus/minus one standard deviation.

		Baseline	vs	MMB
Human Engaging	DialoGPT std. beam	17 *		83 *
	DialoGPT min beam 20	29 *		71 *
	Meena	37 *		63 *
Human	DialoGPT std. beam	33 *		67 *
	DialoGPT min beam 20	40 *		60 *
	Meena	36 *		64 *

Table 8: ACUTE-Evals (engagingness and humanness) show that MMB Style outperforms DialoGPT with standard generation parameters (GPT-2 medium, beam search with beam width 10), DialoGPT with the same parameters but a min beam length of 20 (to match BlenderBot's setting), and Meena. Asterisk indicates significance (two-tailed binomial test, $p < 0.05$).

Experiment

- Human Evaluations

2. Initial Responses to Images

Win %	Loss % MMB	Loss % Dodeca	Loss % 2AMMC
MMB Style		65 *	49
Dodeca	35 *		39 *
2AMMC	51	61 *	

Table 10: ACUTE-Evals on the image-response metric show that MMB Style and 2AMMC significantly outperform Dodeca fine-tuned on Image-Chat. ACUTE-Evals are measured on the models' first response to an image only.

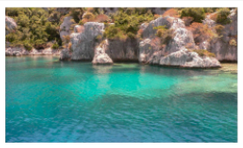
Image		Output
	*(Style)*	*Sweet*
	MMB	I would love to take my wife here for our anniversary. It would be so romantic.
	Dodeca	What a beautiful view!
	2AMMC	Oh what a great honeymoon spot with the lovely view of the mountains.
	*(Style)*	*Maternal (Mother-like)*
	MMB	I would love to take my children here to show them the beauty of the earth and its inhabitants.
	Dodeca	I would love to take my kids here.
	2AMMC	I would like to hide my kids safe in that cavern from a storm.
	*(Style)*	*Fearful*
	MMB	I hope that dog doesn't bite me. I don't want to get bitten by a dog.
	Dodeca	I hope that dog doesn't bite me.
	2AMMC	This dog looks ready to attack. It's focused on something, maybe another animal? I hope it doesn't go after it or anyone around!
	*(Style)*	*Playful*
	MMB	I would love to climb to the top of that sculpture and play in the fountain.
	Dodeca	I would love to visit this place!
	2AMMC	Who wants to put bubble bath in the fountain with me?
	*(Style)*	*Pompous (Self-important, Arrogant)*
	MMB	I have a statue just like this one in my backyard, but it's much nicer.
	Dodeca	I could take a better picture than that.
	2AMMC	I could make a fence that looks way better
	*(Style)*	*Empathetic*
	MMB	These flowers remind me of my grandmother's garden. I miss her so much.
	Dodeca	I feel bad for the flowers.
	2AMMC	I can't imagine how difficult it is for the farmer to round them all up. He works so hard.
	*(Style)*	*Narcissistic (Self-centered, Egotistical)*
	MMB	I would never go to a beach like this. My beach is so much nicer.
	Dodeca	I could take a better picture than that.
	2AMMC	I could swim out to those rocks. The current is strong but I am a stronger swimmer.
	*(Style)*	*Boisterous (Rowdy, Loud)*
	MMB	I would love to jump in that puddle and play in the mud! It would be so fun!
	Dodeca	This looks like a great place to party!
	2AMMC	I would ride through here on a loud bike and throw a party!
	*(Style)*	*Happy*
	MMB	I can't wait to go to the beach and fly a kite with my best friend!
	Dodeca	What a beautiful day!
	2AMMC	Face full of wind, smell of salt water, cool of the ocean, and fantastic memories being made!

Table 9: Example outputs from MMB Style, Dodeca fine-tuned on Image-Chat, and 2AMMC, on images from Shuster et al. (2020). Image-Chat styles fed to the models are shown above the models' responses.

Experiment

- Human Evaluations

  3. Human/Model Chats About Images

Win %			Loss % MMB	Loss % Dodeca	Loss % 2AMMC
Human Engaging		MMB Style		**70** *	**66** *
		Dodeca	30 *		38 *
		2AMMC	34 *	**62** *	
	Image	MMB Style		**70** *	**58** *
		Dodeca	30 *		51
		2AMMC	42 *	49	
	Image	MMB Style		**61** *	52
		Dodeca	39 *		44
		2AMMC	48	56	

Table 12: ACUTE-Evals show that MMB Style significantly outperforms Dodeca and often 2AMMC on various metrics on human/model conversation about an image.

Multimodal Dialogue Response Generation

多模态对话响应生成

Qingfeng Sun **Yujing Wang** **Can Xu** **Kai Zheng**
Yaming Yang **Huang Hu** **Fei Xu** **Jessica Zhang**
Xiubo Geng **Daxin Jiang**
Microsoft STC Aisa Microsoft Research Asia

Introduction

- a good intelligent conversational agent should not only be able to converse freely with plain text, but also have the ability to perceive and share the real physical world.

  But recent large-scale pre-trained text-only dialogue like DialoGPT still cannot rely exclusively on plain text to completely simulate the rich experience of visual perception.

- various vision-language tasks have been introduced and attracted widespread attention, such as visual question answering, photo sharing.

  But performance of retrieval-based method is limited in specific domains. On the other hand, to make an image reply appropriately for the context, a better way is to generate a new one accordingly.

Related Work

- Textual Dialogue Response Generation:
    - End-to-end response generation for textual open domain dialogues is inspired by the neural sequence-to sequence models on machine translation. But the previous works on open domain dialogue response generation that converse freely with plain text.

- Text-to-Image Generation:
    - Draw generative model could generate images from natural language description
    - Reed et al. (2016) proposed a generative adversarial network to improve the image fidelity
    - Cho et al. (2020) use uniform masking with a large range of masking ratios and align the right pre-training datasets to the right objectives

Idea

- formulate a new problem: Multimodal Dialogue Response Generation(MDRG): given the dialogue context, the model should decide to generate an informative text or a high-resolution image as response.

- Due to human effort is expensive, extending the assumption of MDRG to a low-resource setting where only a few multi-modal dialogues containing both texts and images are available

- incorporate text-to-image generation into text-only open domain dialogue generation. Key idea is to make parameters that rely on multimodal dialogues small and independent by disentangling textual dialogue response generation and text-to-image generation, and thus can learn the major part of the generation model from text-only dialogues and <text description, image> pairs that are much easier to be obtained.

- **Divter**, a novel conversational agent powered by large-scale visual world experiences

Model

- Textual Dialogue Response Generator:
  - it consists of a 24-layers Transformer with a hidden size of 1024 and 16 heads.
  - if the target is a text response: loss is

$$\mathcal{L}_{\mathcal{G}} = \mathbb{E}_{r^e \sim \mathcal{D}}[-\log p(r^e)] \quad (1)$$

$$p(r^e) = \prod_t p(w_t | U, w_{1:t-1}) \quad (2)$$

  - if the target is a text image description: loss is

$$\mathcal{L}_{\mathcal{G}} = \mathbb{E}_{c \sim \mathcal{D}}[-\log p(c)] \quad (3)$$

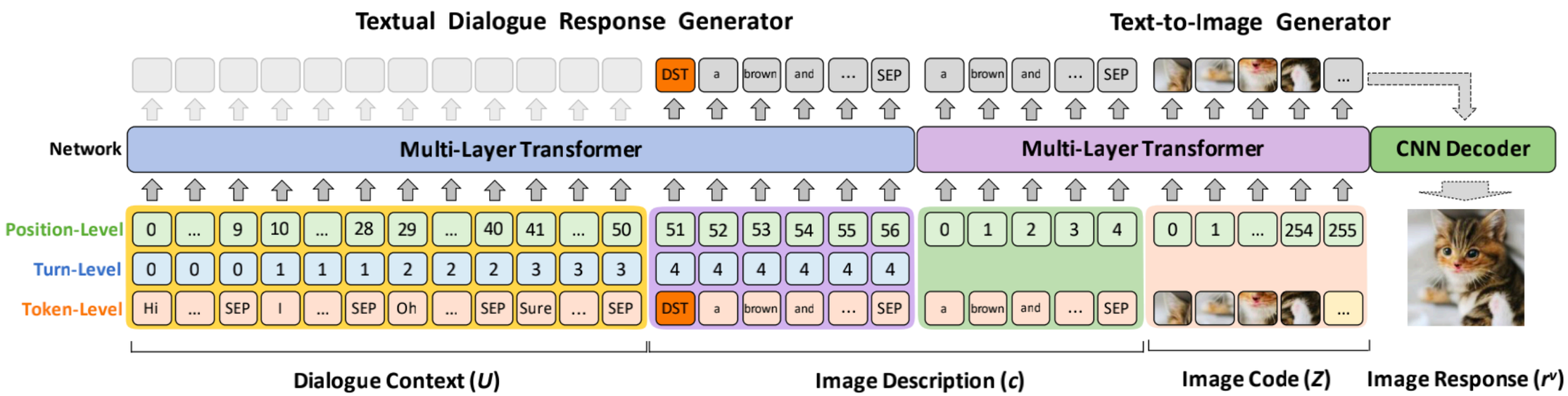
$$p(c) = p(w^v | U) \prod_t p(w_t | U, w^v, w_{1:t-1}) \quad (4)$$

- Text-to-Image Generator:
  - it consists of an image representation module and a text-to-image transformer module
  - image representation module :

    use a learned discrete codebook to represent the image

    take a convolutional model consists of an encoder and a decoder together to learn to represent images
  - text-to-image transformer： 24-layers Transformer with a hidden size of 1024 and 16 attention heads

    take description as input and generate its image representation, find closest codebook entries, then uses the image decoder to reconstruct it to an image



Experiments

- Dataset:

  PhotoChat dataset released by Zang et al. (2021)

  a multi-turn multimodal conversations consists of 10917 unique images and 12286 dialogues, each of which is paired with an user image that is shared during the conversation, and each image is paired with its text description

```
{
  "dialogue":[
    {
      "share_photo":"Boolean value denoting whether a photo is shared in this turn.",
      "user_id":"0 or 1. User id of this turn.",
      "message":"Text of one conversation turn. Empty when share_photo is true."
    }
  ],
  "dialogue_id":"Integer. Unique dialogue id.",
  "photo_description":"String. Photo description. It includes info about object labels in the photo.",
  "photo_url":"Photo url."
}
```

Experiments

- Implementation Details:

  1. Textual dialogue response generator uses DialoGPT as pre-trained model initialization which has been trained on Reddit.

  2. Image representation module uses Taming Transformers for High-Resolution Image Synthesis which has been trained on 10M ImageNet images.

  3. For text-to-image transformer, randomly selecting 5M pairs from ImageNet, and 10M pairs from YFCC100M as training data and use CLIP to score.

- Baseline:

  BERT-base, T5-3B, SCAN

Experiments

- Evaluation Metrics
  - automatic metrics：
    1. Image Response Intent Prediction(F1)
    2. Text Description Generation(PPL, BLEU, Rouge, F1)
    3. Image Generation Quality(FID, IS)
    4. Text Response Generation(PPL, BLEU, Rouge, F1)
  - human evaluation:

    Three human annotators are asked to score the response quality on a scale of 0, 1, 2 from four aspects:

    1. Context Coherence
    2. Fluency
    3. Image Quality
    4. Background Consistency of Image

Models	Intent	Text Description Generation				Image Generation		Text Response Generation			
	F1	PPL	B-1	B-2	Rouge	FID ↓	IS ↑	PPL	B-1	B-2	Rouge
BERT-base	53.2	–	–	–	–	–	–	–	–	–	–
T5-3B	**58.9**	–	–	–	–	–	–	–	–	–	–
Divter	56.2	**5.12**	**15.08**	**11.42**	**15.81**	**29.16**	**15.8 ± 0.6**	59.63	**6.52**	**1.66**	**5.69**
Divter (*w/o* $\mathcal{G}$ pre-train)	47.3	122.56	1.99	0.23	2.60	29.78	15.5 ± 0.5	153.62	4.82	0.53	3.83
Divter (*w/o* $\mathcal{F}_\phi$ pre-train)	55.9	5.23	15.01	11.20	15.63	262.09	4.9 ± 0.7	63.76	6.28	1.51	5.40
Divter (*w/o* $\mathcal{G}$, $\mathcal{F}_\phi$ pre-train)	47.1	128.87	1.75	0.21	2.38	254.31	5.2 ± 0.6	163.85	4.53	0.48	3.55
Divter (*w/o* joint learning)	55.6	5.20	15.00	11.36	15.73	29.04	15.4 ± 0.6	**59.21**	6.47	1.58	5.63

Table 1: Automatic evaluation results of Divter and baselines on the test set. (*w/o* joint learning) means fine-tuning $\mathcal{G}$ and $\mathcal{F}_\phi$ respectively rather than using Eq. 11. Numbers in bold mean that the improvement to the best baseline is statistically significant (t-test with p-value < 0.01).

Models	Context Coherence	Fluency	Image Quality	Background Consistency	Kappa
SCAN(Retrieval)	–	–	1.95	0.96	0.65
Divter (*w/o* pre-train)	0.94	1.56	0.61	0.35	0.66
Divter	1.59	1.95	1.83	1.61	0.63

Table 2: Human evaluation results.

Ablation study

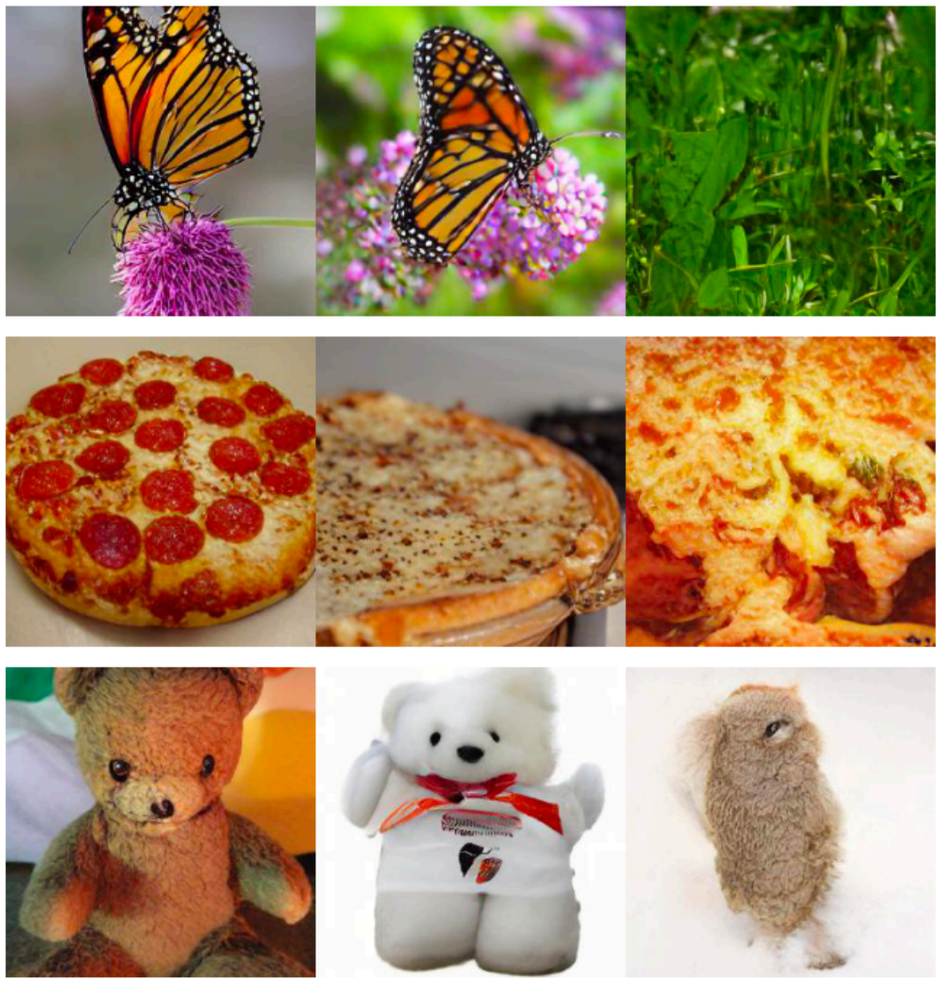


Table 3: Qualitative assessment of various model and variants for image response generation with the same textual dialogue context as input in PhotoChat test set. 1st column: Divter . 2nd column: Divter *w/o* $\mathcal{G}$ pre-train. 3rd column: Divter *w/o* $\mathcal{F}_\phi$ pre-train.

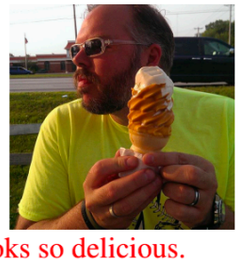
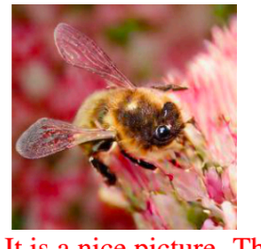
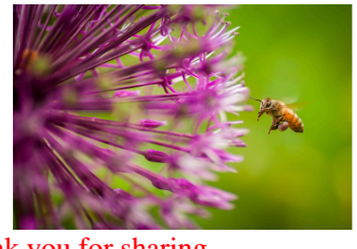
Example 1	Example 2
A: OMG...the new ice cream shop is amazing. ..... **A:** I had the twist chocolate and vanilla but it was so fresh tasiting. like you just made it. like you just made it. **B:** I call it the malado gilato. **A:** Sam wouldn't let me have another lick bc he thought I'd eat it all. **D:** That sounds interesting. **D:** Yes, could you please share it with me? **D:** Objects in the photo: Chocolate Ice cream, Dairy, Drink. **D:**   **D:** Wow! The ice cream looks so delicious. **D:** Sure, it tastes pretty good.	**A:** Have you been out in nature lately? **B:** Yes. ..... **A:** I'm sitting at home now looking through some old photographs. **B:** I see. than. **A:** Would you like to see one of my favorites It's a cool shot of a honey bee near a beautiful flower. **D:** Objects in the photo: Honey bee, Insect, Animal, Flower. **D:**   **D:** It is a nice picture. Thank you for sharing. **D:** Haha, just enjoy the beautiful scenery. **D:** Yeah, definitely.

Table 4: Examples of PhotoChat test set. In each example, the turns with the prefix of "A"/"B" are the given context; the blue text is the text description generated by Divter; the **left** image and the red response are generated by Divter, the **right** image is the ground-truth image.

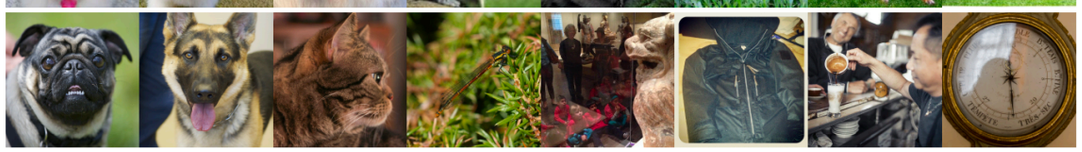
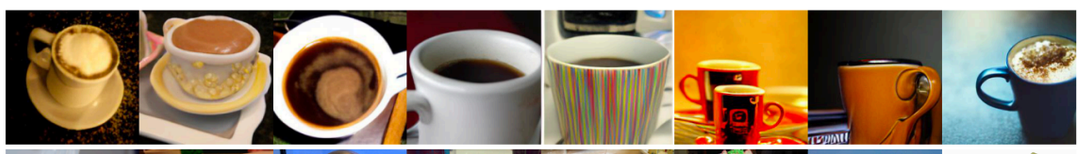
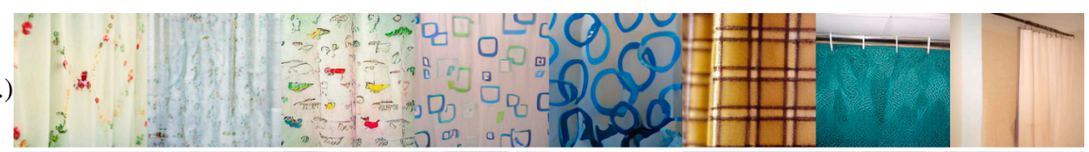
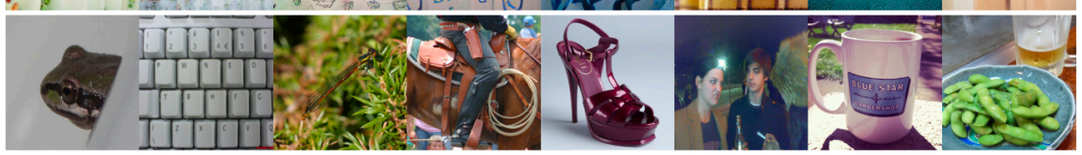
	Generated / Retrieved Images
Generation (Generated Description: objects in the photo: animal, dog, carnivore, grassland)	
Retrieval (Should contain "dog".)	
Generation (Generated Description: objects in the photo: coffee cup, drink, bottle, mug, tea.)	
Retrieval (Should contain "coffee cup".)	
Generation (Generated Description: objects in the photo: curtain.)	
Retrieval (Should contain "curtain".)	

Table 5: Examples of the images generated by Divter and the images retrieved by SCAN.

Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts

多粒度视觉语言预训练：将文本与视觉概念对齐

Yan Zeng, Xinsong Zhang, Hang Li
ByteDance AI Lab

Introduction

- Vision language pre-training aims to learn vision language alignments from a large number of imagetext pairs.A pre-trained Vision Language Model(VLM) fine-tuned with a small amount of labeled data has shown the SoTA performances in many Vision Language tasks

- Existing methods:fine-grained & coarse-grained

- Problem & Motivation:

  Both have drawbacks.

  existing methods either depend on object-centric features or overall features.

  want the VLM to represent and learn multi-grained alignments between the images and texts.

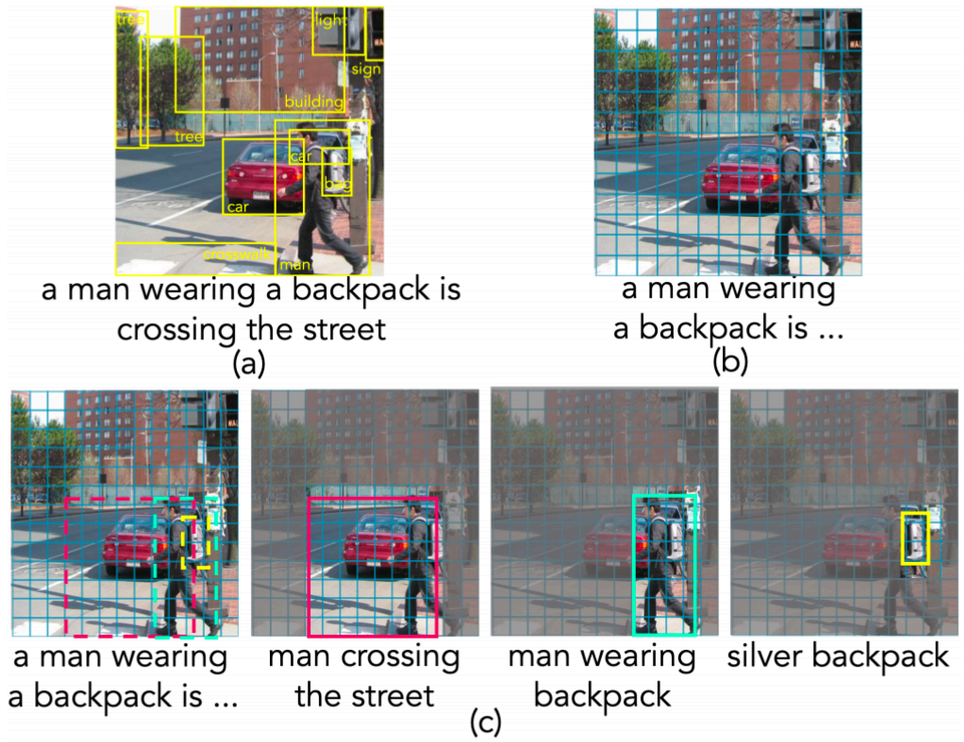


Figure 1: A comparison of (a) the existing methods relying on object detection, (b) the methods aligning the text with the whole image, and (c) our method.

Related Work

- Fine-grained细粒度的:
  - Method: An image is represented by dozens of object-centric features of the identified regions relying on object detection
  - Problem:cannot represent relations among multiple objects in multiple regions
- Coarse-grained粗粒度的:
  - Method:extracting and encoding overall image features with convolutional network or vision transformer
  - Problem:the fine-grained features seem critical for learning VLMs.Coarse-grained performances are usually not as good as fine-grained approach

Method & Model

- **Overview:** X-VLM consists of an image encoder (I-trans),a text encoder (T-trans), and a cross-modal encoder (X-trans). All encoders are based on Transformer.The cross-modal encoder fuses the vision features with the language features through cross-attention at each layer.

  an image may have multiple regions enclosed by bounding boxes, and each of them is associated with a text that describes an object or a region, denoted as

$$(I, T, \{(V^j, T^j)\}^N)$$

Vj is an object or region in the image I associated with a bounding box

$$b^j = (cx, cy, w, h)$$

represented by the normalized center coordinates, width, and height of the box.

Method & Model

- **Image Encoder:**

① The encoder first splits an image into non-overlapping 16x16 patches and linearly embeds all patches, yielding $\{v_1^0, \ldots, v_{N^I}^0\}$, [cls] denotes v_{cls}^0.

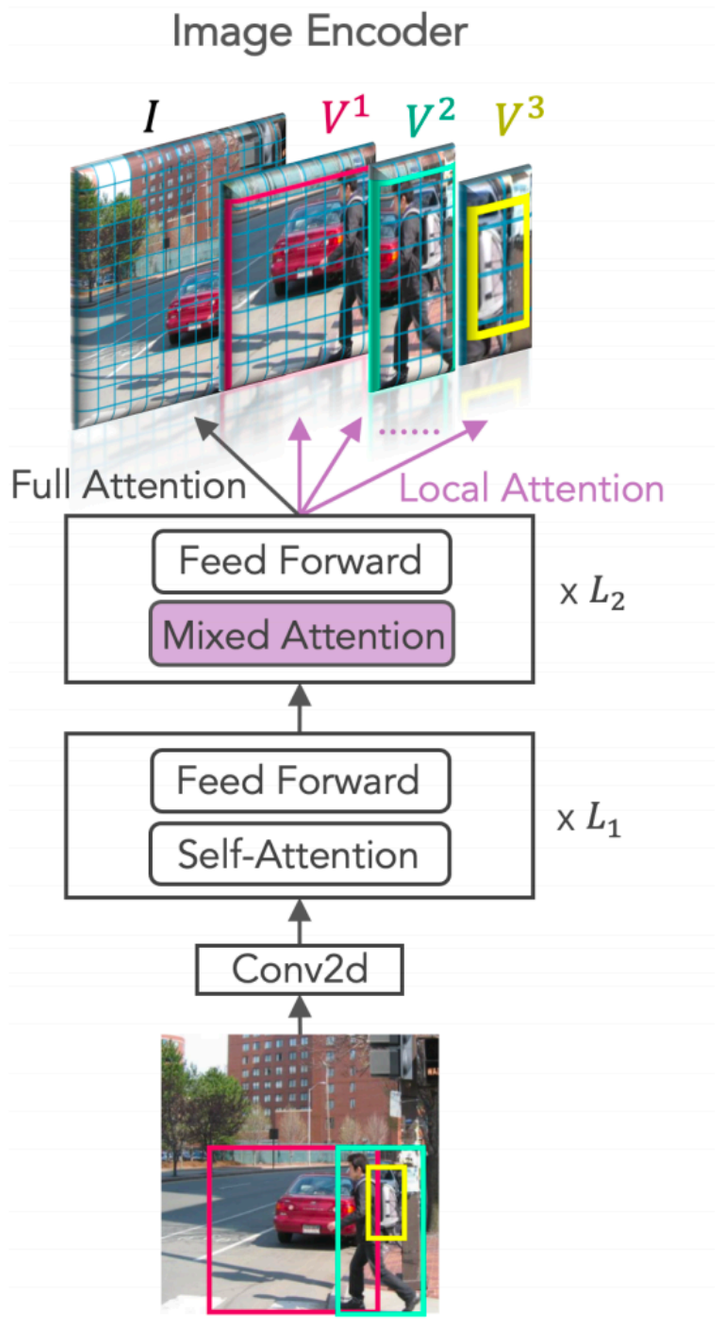
② At the first L1 layer, apply bi-directional attention to all patches.

③ If there are N regions, duplicate the $\{v_{cls}^{L_1}, v_1^{L_1}, \ldots, v_{N^I}^{L_1}\}$ into N+1 copies. At L2 layer, apply full attention to one copy and local attention to the others.

A region Vj is image patches $\{p_1^j, \ldots, p_M^j\}$ and incorporate self-attention mask in the L1+1 to L1+L2 Transformer layers to implement local attention: $H = softmax(\frac{QK^T}{\sqrt{d}} + M)V$

self-attention mask matrix M ∈ R(NI+1)×(NI+1) (with Mij ∈ {0, −∞}) determines whether a position can attend to the other positions.

Then we get N+1 concept representations in different granularities: $V^j = \{v_{cls}^j, v_{p_1^j}^j, \ldots, v_{p_M^j}^j\}, j \in [0, N]$



Method & Model

- **Cross-modal modeling:**

1. Bonding box prediction:

   predict the bj of visual concept Vj given the image representation with full attention and the text representation: $\hat{b}^j(I, T^j) = Sigmoid(MLP(x^j_{cls}))$

   Loss: IoU:

   $L_{bbox} = E_{(V^j,T^j)\sim I;I\sim D}[L_{iou}(b^j, \hat{b}^j) + ||b^j - \hat{b}^j||l]$

2. Contrastive learning:

   predict (visual concept, text) pairs, use cosine similarity $s(V, T) = g_v(v_{cls})^T g_w(w_{cls})$, and Loss:

   $L_{cl} = \frac{1}{2}E_{(V,T)\sim D}[H(y^{v2t}(V), p^{v2t}(V)) + H(y^{t2v}(V), p^{t2v}(V))]$

3. Matching:

   determine whether a pair of visual concept and text is matched, and Loss:

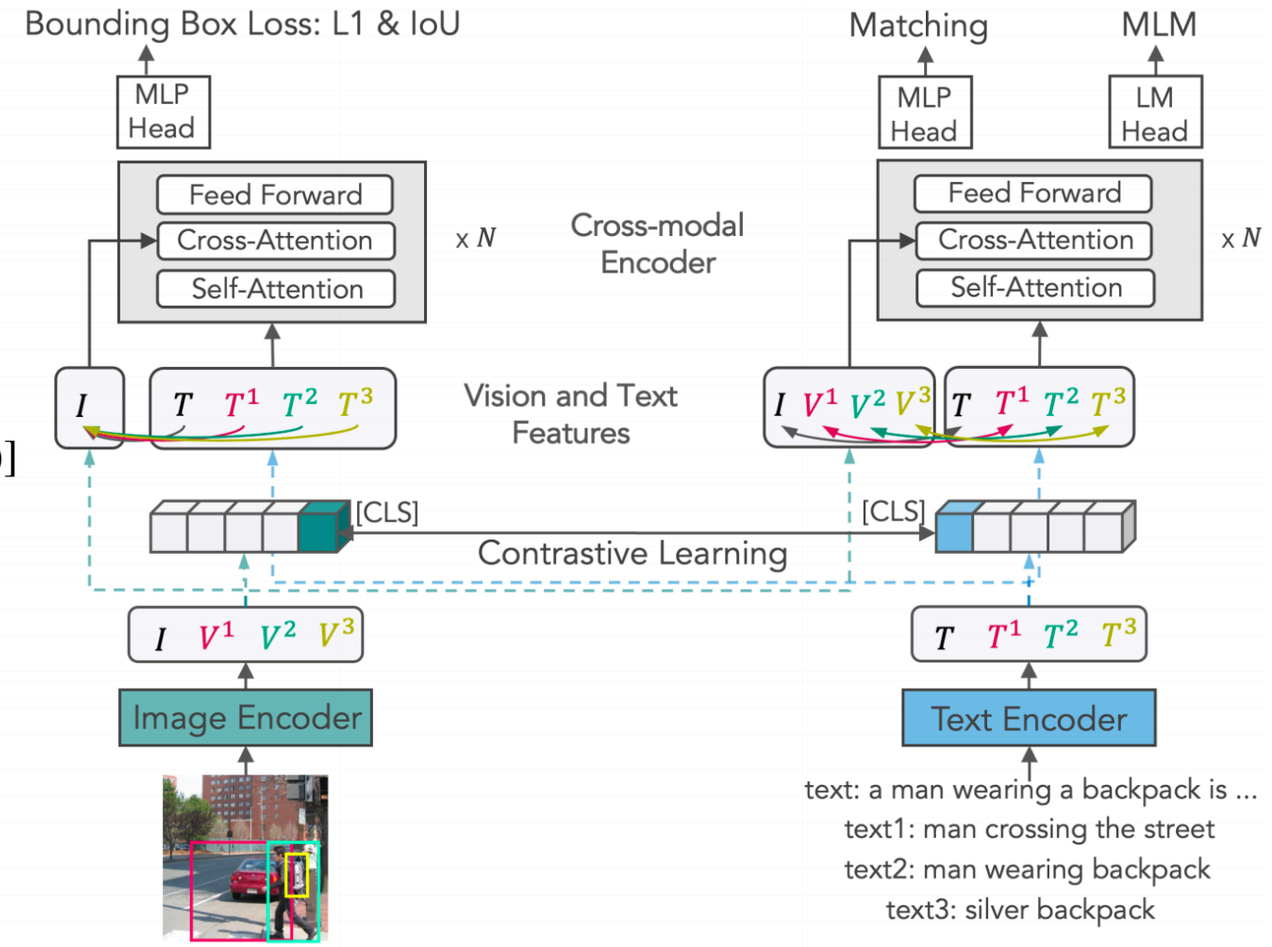
   $L_{match} = E_{(V,T)\sim D}[H(y^{match}, p^{match}(V, T))]$

4. Masked Language Modeling:

   predict the masked words in the text based on the visual concept. Mask rate:25%. 10% random, 10% unchanged, 80% [mask]. Loss:

   $L_{mlm} = E_{t_j\sim\hat{T}:(V,\hat{T})\sim D}H(y^j, p^j(V, \hat{T}))$

   Total loss: $L = L_{bbox} + L_{cl} + L_{match} + L_{mlm}$



Experiment

- Pre-training datasets:

  In-domain datasets: **coco** & **Visual genome(VG)**

  Out-of-domain datasets: **SBU Captions** & **Conceptual Captions (CC)**

- Implementation details:

  Image encoder: **12-layer ViT-base** which is initialized with weights pre-trained on ImageNet-1k

  Text encoder: **6-layer Bert-base** which is initialized using the first six layers of BERTbase

  Cross-modal encoder: **6-layer Bert-base** which is initialized using the last six layers of BERTbase

Experiment

- Downstream tasks:

  Image-Text Retrieval: text retrieval (TR) and image retrieval (IR) on datasets MSCOCO & Flickr30K

  Visual Question Answering: predict an answer given an image and a question. use a six-layer Transformer decoder to generate answers based on the outputs of the cross-modal encoder

  Natural Language for Visual Reasoning:determine whether a text describes the relations between two images

  Visual Grounding:locate the region in an image that corresponds to a specifi c text description on RefCOCO+ in both supervised(bounding box annotations) and weakly-supervised(Grad-CAM) settings

Method	# Pre-train Images	MSCOCO (5K test set)						Flickr30K (1K test set)					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UNITER	4M	64.4	87.4	93.1	50.3	78.5	87.2	85.9	97.1	98.8	72.5	92.4	96.1
OSCAR	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
ALBEF	4M	73.1	91.4	96.0	56.8	81.5	89.2	94.3	99.4	99.8	82.8	96.7	**98.4**
VinVL	5.6M	74.6	92.6	96.3	58.1	83.2	90.1	-	-	-	-	-	-
X-VLM	4M	**78.8**	**94.3**	**97.5**	**60.6**	**84.2**	**90.5**	**96.0**	**99.7**	**99.9**	**84.1**	**96.9**	**98.4**
ALIGN	1.8B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	**99.8**	**100.0**	84.9	97.4	98.6
ALBEF	14M	77.6	94.3	97.2	60.7	84.3	90.5	95.9	**99.8**	**100.0**	85.6	**97.5**	**98.9**
X-VLM	14M	**79.7**	**95.3**	**97.6**	**62.5**	**85.3**	**91.0**	**96.8**	99.7	99.9	**86.0**	97.2	98.7

Table 1: Image-text retrieval results on MSCOCO and Flickr30K datasets. IR: Image Retrieval and TR: Text Retrieval.

Method	VQA		NLVR^2		RefCOCO+		
	test-dev	test-std	dev	test-P	val^d	testA^d	testB^d
VisualBERT (Li et al., 2019)	70.80	71.00	67.40	67.00	-	-	-
ViLBERT (Lu et al., 2019)	70.55	70.92	-	-	72.34	78.52	62.61
VL-BERT (Su et al., 2019)	71.16	-	-	-	72.59	78.57	62.30
LXMERT (Tan and Bansal, 2019)	72.42	72.54	74.90	74.50	-	-	-
UNITER (Chen et al., 2020)	72.70	72.91	77.18	77.85	75.31	81.30	65.58
12-in-1 (Lu et al., 2020)	73.15	-	-	78.87	-	-	-
OSCAR (Li et al., 2020c)	73.16	73.44	78.07	78.36	-	-	-
Pixel-BERT (Huang et al., 2020)	74.45	74.55	76.50	77.20	-	-	-
VILLA (Gan et al., 2020)	73.59	73.67	78.39	79.30	76.05	81.65	65.70
SOHO (Huang et al., 2021)	73.25	73.47	76.37	77.32	-	-	-
ViLT (Kim et al., 2021)	70.94	-	75.24	76.21	-	-	-
ALBEF (Li et al., 2021)	74.54	74.70	80.24	80.50	57.85*	65.89*	46.43*
ALBEF (14M)	75.84	76.04	82.55	83.14	58.46*	65.89*	46.25*
VinVL (5.6M) (Zhang et al., 2021)	75.95	76.12	82.05	83.08	-	-	-
X-VLM	76.20	76.23	82.40	82.42	75.81/**64.95***	82.13/72.07*	**68.26**/54.84*
X-VLM (14M)	**76.77**	**76.89**	**83.40**	**83.84**	**76.31**/64.79*	**82.78/72.46***	67.40/**55.08***

Table 2: Comparison on downstream V+L tasks. RefCOCO+ scores with * are in the weakly-supervised setting.

Figure 3: Grad-CAM visualization and bounding box prediction on unseen images.

Ablation study

	Meta-Sum	MSCOCO		Flickr30K		VQA	NLVR^2	RefCOCO+	
		TR	IR	TR	IR	test-dev	test-P	testA^d	testB^d
X-VLM	**755.4/605.0***	**78.8**	**60.6**	**96.0**	**84.1**	76.20	82.42	**82.13**/72.07*	**68.26**/54.84*
w/o object	749.7/603.5*	77.4	60.4	95.0	83.7	75.87	82.10	81.19/**73.37***	64.94/**55.69***
w/o region	740.2/596.0*	76.8	60.2	**96.0**	83.6	75.84	82.20	80.11/70.73*	64.12/50.60*
w/o bbox loss	-/594.9*	77.5	60.2	95.7	83.5	**76.77**	81.49	-/69.32*	-/50.38*
w/o mixed attn	749.9/600.8*	77.0	60.1	95.4	83.2	75.87	**82.55**	82.08/72.16*	67.05/54.51*
w/o all	-/580.6*	74.5	57.9	95.6	82.8	74.90	80.70	-/67.79*	-/46.43*

Table 3: Ablation study on X-VLM: w/o object is training without concepts of object; w/o region is training without concepts of region; w/o bbox loss is without bounding box prediction; w/o mixed attn is applying completely local attention to obtain fine-grained concept representations; w/o all represents removing all above components.