当NLPer爱上CV：后BERT时代生存指南之VL-BERT篇

原创 小鹿鹿lulu 夕小瑶的卖萌屋 3月16日

来自专辑

卖萌屋@自然语言处理

一只小狐狸带你解锁 炼丹术&NLP 秘籍

前言

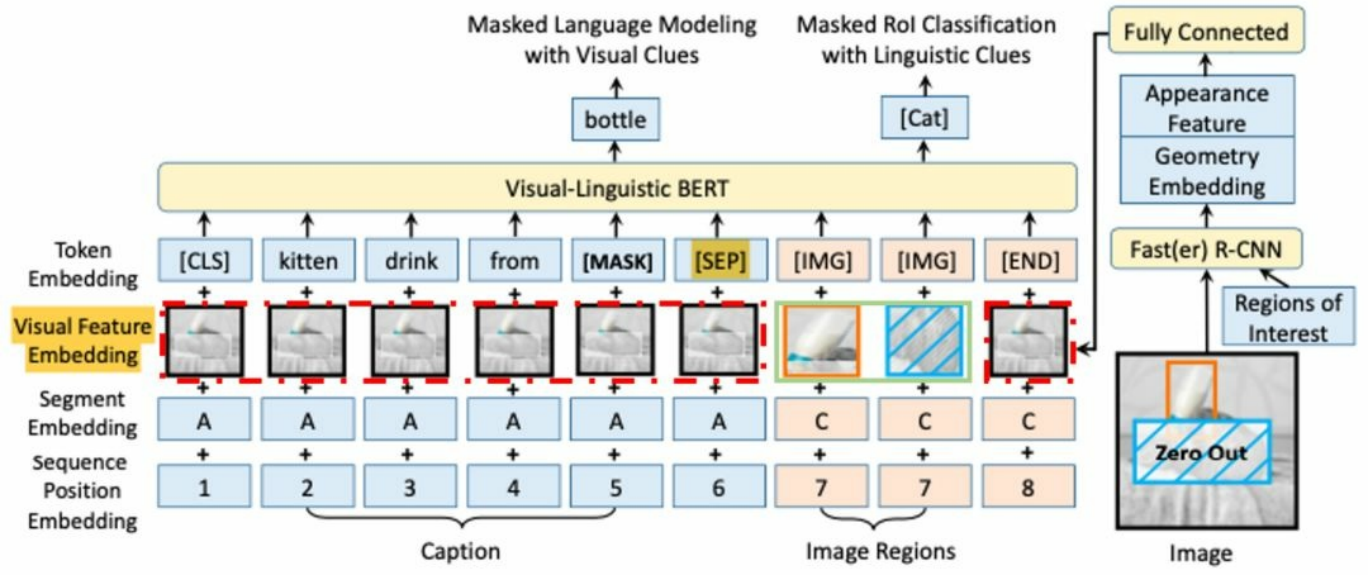
BERT的出现让NLP发展实现了一个大飞跃，甚至有大佬说NLP已经没有可以做的啦，后面就是拼机器拼money了。但是，我认为任何领域的进步之后都会有更苛刻的要求，科研没有尽头，需求也永远无法满足。**而多模态，要求机器拥有多维度的感知能力，就是一个更强的挑战。** 关于这个话题也逐渐成为另外一个新热点。从19年到现在的论文数量就可见一斑。

	Method	Architecture	Visual Token	Pre-train Datasets	Pre-train Tasks	Downstream Tasks
Published Works	VideoBERT (Sun et al., 2019b)	single cross-modal Transformer	video frame	Cooking312K (Sun et al., 2019b)	1) sentence-image alignment 2) masked language modeling 3) masked visual-words prediction	1) zero-shot action classification 2) video captioning
Works Under Review / Just Got Accepted	CBT (Sun et al., 2019a)	two single-modal Transformer (vision & language respectively) + one cross-modal Transformer	video frame	Cooking312K (Sun et al., 2019b)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature regression	1) action anticipation 2) video captioning
	ViLBERT (Lu et al., 2019)	one single-modal Transformer (language) + one cross-modal Transformer (with restricted attention pattern)	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification	1) visual question answering 2) visual commonsense reasoning 3) grounding referring expressions 4) image retrieval 5) zero-shot image retrieval
	B2T2 (Alberti et al., 2019)	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling	1) visual commonsense reasoning
	LXMERT (Tan & Bansal, 2019)	two single-modal Transformer (vision & language respectively) + one cross-modal Transformer	image RoI	‡ COCO Caption + VG Caption + VG QA + VQA + GQA	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification 4) masked visual-feature regression 5) visual question answering	1) visual question answering 2) natural language visual reasoning
	VisualBERT (Li et al., 2019b)	single cross-modal Transformer	image RoI	COCO Caption (Chen et al., 2015)	1) sentence-image alignment 2) masked language modeling	1) visual question answering 2) visual commonsense reasoning 3) natural language visual reasoning 4) grounding phrases
	Unicoder-VL (Li et al., 2019a)	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification	1) image-text retrieval 2) zero-shot image-text retrieval
	Our VL-BERT	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018) + BooksCorpus (Zhu et al., 2015) + English Wikipedia	1) masked language modeling 2) masked visual-feature classification	1) visual question answering 2) visual commonsense reasoning 3) grounding referring expressions

所以，为了迎上发展的势头，在继videoBERT之后又调研了一番image和BERT结合的工作。下文将介绍**MSRA出品的VL-BERT**，通过这个模型来一览现阶段 **image+BERT** 的研究现状吧。

后台回复【VL-BERT】下载论文原文~~

模型介绍



VL-BERT模型以transformer为骨干，将BERT的输入扩展为文本+图像。那么问题来了，怎样将两者花式融合呢？让我们揣测一下作者的炼丹思路：

1. **图片和文本没法直接对齐，暴力输入整张图**

于是就有了图中用红色虚线框起来的部分，直接将图像、文本、segment和position embeding加和输入。这样做MLM任务是没问题了，但怎样确定模型能准确提取图像信息呢？

2. **提取图像中的重要部分，增加无文本的图像输入**

由于整张图片的粒度远大于文本token，一次性输入整张图片显然不利于图像和文本信息的交互。**所以使用了目标检测工具对图片进行分块，提取图像中感兴趣的核心部分RoI（region-of-interest），加上[IMG]标识，输入到模型中（图中浅绿色实线框起来的部分）。** 为了不失掉全局信息，在[END]对应的位置又加上了整张图像。另外，我们假设图片的不同区域是没有顺序可言的，即position embedding是一样的。

> 类比文本输入，模型实际上接受的是文本token（subword）对应的word embedding，所以我们会对所有图像输入（不管是整张图片还是局部RoIs）使用pre-trained R-CNN提取2048维的visual feature embedding输入到模型中。

自监督学习任务（pretrain）

结合上文介绍的模型结构，再强调一下两个预训练任务：

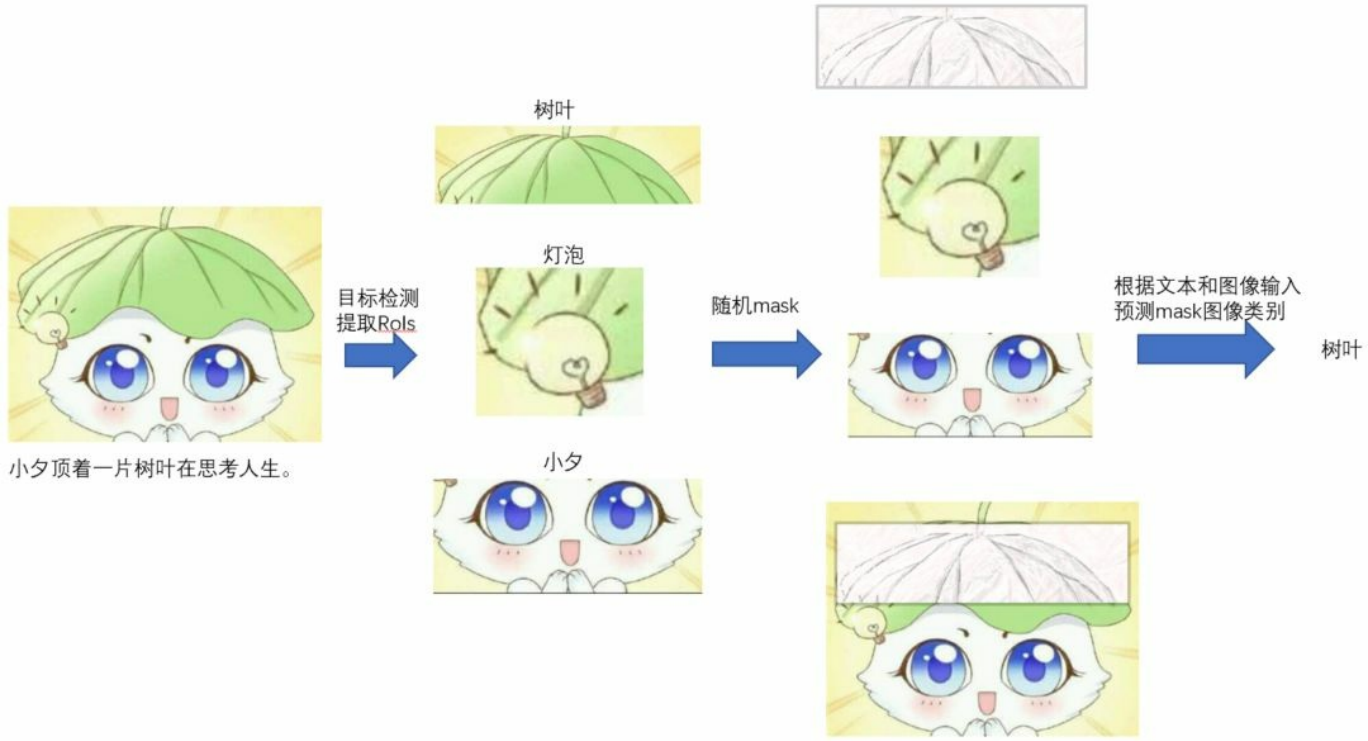
1. **Masked Language Model with visual Clues**

根据文本+图像信息预测文本token，升级版的MLM。 唯一不同的是被mask的word除了根据没被mask的文本来预测还可以根据视觉的信息来辅助。比如上图中的例子，被mask后的word sequence是kitten drinking from [MASK]，如果没有图片给我们的视觉信息是无法预测出被mask的词是bottle。

2. **Masked RoI Classification with LinguisticClues**

根据文本+图像信息预测RoIs的类别，针对图像的“MLM”。 以下图为例，首先对图片使用目标检测工具提取RoIs并获得所属类别，然后随机mask局部区域（树叶部分）。需要注意的是，由于模型会接收整张图片的输入，为了避免信息泄露，整张图片对应

的部分也要mask。最后，模型根据文本信息和被mask的图片信息预测mask区域所属类别。



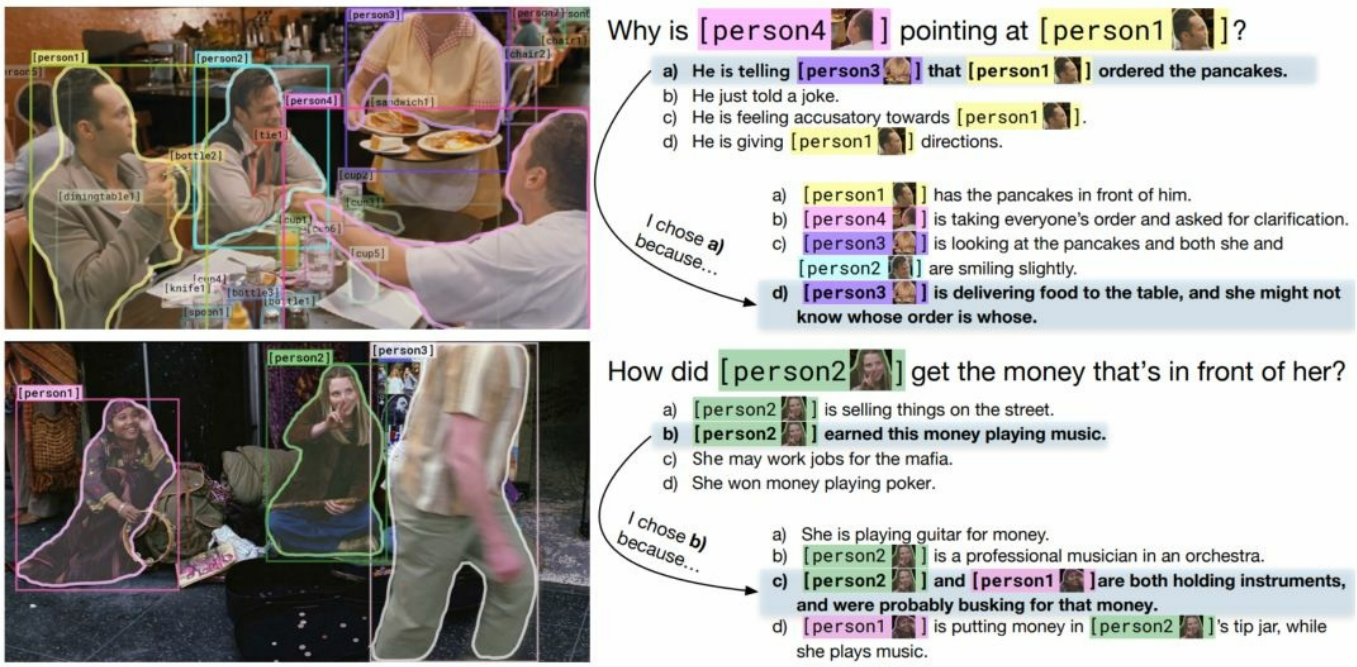
下游任务（finetune）

模型通过接收<text, image>输入，通过自监督任务学习到general跨模态表示后，可以很自然的应用很多跨模态的任务中。延续原始BERT的设定，**[CLS]最后输出的feature可以预测文本和图片的关系（sentence-image-relation），被mask的text token或者RoI的输出用来做word-level或者RoI-level的预测。**

下面来看看不同的下游任务是怎么实现的叭~

1. 视觉常识推理(VCR)

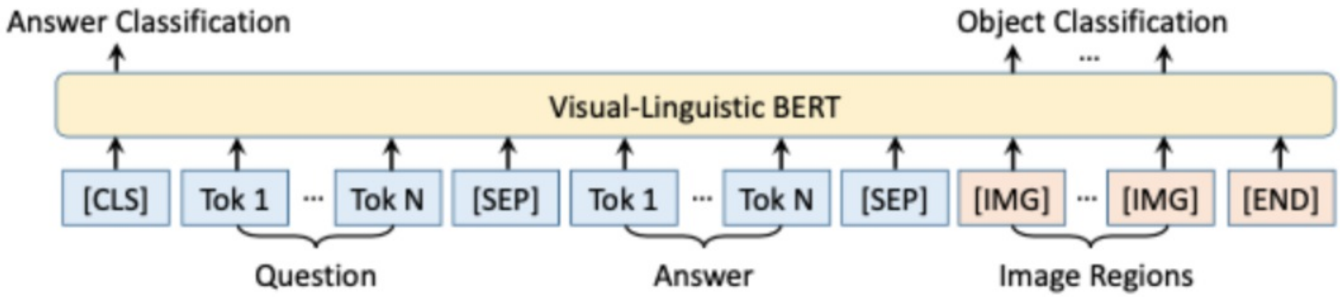
给定一张图片中的多个RoIs和问题（Q），需要选出答案（A）并解释为什么（R）。VCR任务超越目标检测（object detection），是需要结合认知层面的复杂推理任务。下图展示了数据中的两个例子[1]，确实很难很复杂 。



整体任务{Q->AR}可以拆解成两个子任务 **{Q->A}（根据问题Q预测答案A）** 和 **{QA->R}（根据QA推理原因R）**。而这两个子任务

都是选择题，模型只需要从候选答案中挑选认为最正确的选项就好。如下图文本输入由两部分组成Question（已知信息）和Answer（候选答案），图像输入为人工标注的RoIs。**针对{Q->A}任务**，已知的文本信息为问题Q的文本描述。**对{QA->R}任务**，已知的文本信息为问题Q加上一个任务预测的答案A。两个任务都根据最后一层[CLS]的输出预测该候选答案（A/R）是否正确。

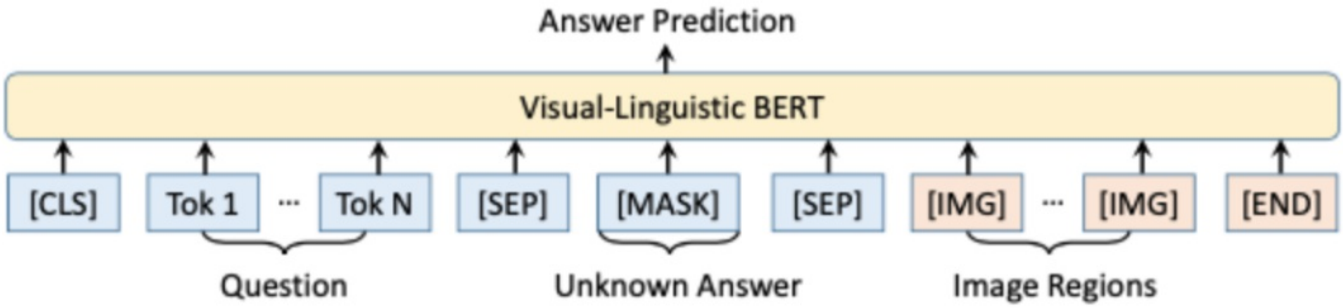
> 有一个不合理的地方是，正常人的思维模式是先有一个靠谱的理论依据R得出正确答案A。但是上面模型的逻辑是先有正确答案，再去找合理的原因。因果颠倒。



最终结果，不管是对比task-specific模型R2C还是其他多模态模型，VL-BERT都有非常明显的优势的。

2. **视觉问答 (VQA)**

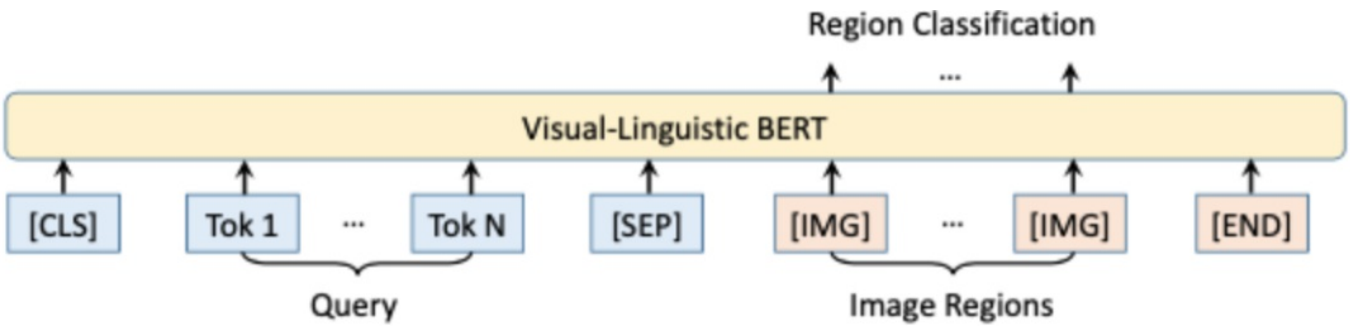
文章follow了一个专门针对VQA任务设计的模型BUTD实验设定，把VQA任务转化为一个有3k+候选答案的多分类问题，根据被masked Answer token最后一层的输出预测。



相比special design的网络结构（BUTD），VL-BERT在准确率上提升了5%，和其他多模态pretrain model效果相当。

Model	test-dev	test-std
BUTD (Anderson et al., 2018)	65.32	65.67
ViLBERT (Lu et al., 2019)†	70.55	70.92
VisualBERT (Li et al., 2019b)†	70.80	71.00
LXMERT (Tan & Bansal, 2019)†	72.42	72.54
VL-BERT$_{BASE}$ w/o pre-training	69.58	-
VL-BERT$_{BASE}$	71.16	-
VL-BERT$_{LARGE}$	71.79	72.22

3. **Referenceing Expression Comprehension（visual grounding）**



这个任务呢，是**根据一句自然语言的描述，定位图片中的具体区域，即判断这句描述讲的是图片的哪个位置。** 因为我们已经对图片划分出了RoIs，所以只需要将每个RoIs最后的输出，接一个Regin classification（二分类），判断Query是否是描述这个区域即可。

Model	Ground-truth Regions			Detected Regions		
	val	testA	testB	val	testA	testB
MAttNet (Yu et al., 2018)	71.01	75.13	66.17	65.33	71.62	56.02
ViLBERT (Lu et al., 2019)†	-	-	-	72.34	78.52	62.61
VL-BERT$_{BASE}$ w/o pre-training	74.41	77.28	67.52	66.03	71.87	56.13
VL-BERT$_{BASE}$	79.88	82.40	75.01	71.60	77.72	60.99
VL-BERT$_{LARGE}$	80.31	83.62	75.45	72.59	78.57	62.30

分析

VL-BERT模型以transformer为骨干，将BERT扩展可以同时接受文本和图片型输入，学习跨模态的表示，在三个下游任务上远超task specific的SOTA模型，并取得和其他pretrain模型comparable或者略胜一筹的结果。

其主要的优势在于 **文本和图片的深度交互**。对比同期工作LXMERT[2]，对text和image输入分别使用single-modal Transformer，然后再接一个cross-modal Transformer，VL-BERT使用一个single cross-modal Transformer，让文本和图片信息能更早更多的交互。

但是这个工作我认为还是有一个需要打问号，或者进一步深入研究的地方。

文章使用的两个自监督任务都是由MLM衍生而来，没有判断文本和图片是否一致（Sentence-Image Relation Prediction）的这个典型任务。

Settings	Masked Language Modeling with Visual Clues	Masked RoI Classification with Linguistic Clues	Sentence-Image Relationship Prediction	with Text-only Corpus	Tuning Fast R-CNN	VCR Q→A val	VCR QA→R val	VQA test-dev	RefCOCO+ Detected Regions val
w/o pre-training						72.9	73.0	69.5	62.7
(a)	✓					72.9	73.1	71.0	69.1
(b)	✓	✓				73.0	73.1	71.1	70.7
(c)	✓	✓	✓			72.2	72.4	70.3	69.5
(d)	✓	✓		✓		73.4	73.8	71.1	70.7
VL-BERT$_{BASE}$	✓	✓		✓	✓	73.8	73.9	71.2	71.1

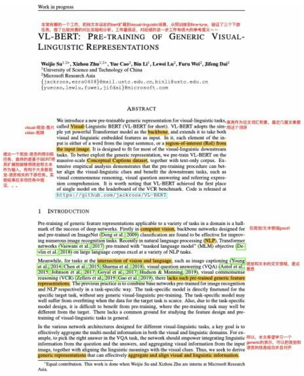
文章在对比实验分析中，提到**加入Sentence-Image Relation Prediction任务进行预训练会导致下游任务效果下降，原因分析是由于数据质量问题，sentence-image对应信号噪声较大。** 但是直觉上文本和图片的对应关系是一个很强的学习跨模态表示的信号，并且在ViBERT[3]和LXMERT上该任务是有正向收益的。

如果优化数据质量，减少sentence-imgage对应信号的噪声，是否可以优化VL-BERT的效果？

如果仍然是负收益，是否是另外两个自监督任务已经涵盖了sentence-image对应信息，增加这个任务唯一的作用就是带来了数据的噪声？

这三个自监督任务是否存在冲突或者矛盾的地方？其关系是什么？值得进一步的研究和探索。

后台回复【VL-BERT】下载带笔记的论文原文~~



Work in progress

VL-BERT: PRE-TRAINING OF GENERIC VISUAL-LINGUISTIC REPRESENTATIONS

Weijie Su[1,2*], Xizhou Zhu[1,2*], Yue Cao[2], Bin Li[1], Lewei Lu[2], Furu Wei[2], Jifeng Dai[2]
[1]University of Science and Technology of China
[2]Microsoft Research Asia
{jackroos,ezra0408}@mail.ustc.edu.cn, binli@ustc.edu.cn
{yuecao,lewlu,fuwei,jifdai}@microsoft.com

ABSTRACT

We introduce a new pre-trainable generic representation for visual-linguistic tasks, called Visual-Linguistic BERT (VL-BERT for short). VL-BERT adopts the simple yet powerful Transformer model as the backbone, and extends it to take both visual and linguistic embedded features as input. In it, each element of the input is either of a word from the input sentence, or a region-of-interest (RoI) from the input image. It is designed to fit for most of the visual-linguistic downstream tasks. To better exploit the generic representation, we pre-train VL-BERT on the massive-scale Conceptual Captions dataset, together with text-only corpus. Extensive empirical analysis demonstrates that the pre-training procedure can better align the visual-linguistic clues and benefit the downstream tasks, such as visual commonsense reasoning, visual question answering and referring expression comprehension. It is worth noting that VL-BERT achieved the first place of single model on the leaderboard of the VCR benchmark. Code is released at https://github.com/jackroos/VL-BERT

1 INTRODUCTION

Pre-training of generic feature representations applicable to a variety of tasks in a domain is a hallmark of the success of deep networks. Firstly in computer vision, backbone networks designed for and pre-trained on ImageNet (Deng et al., 2009) classification are found to be effective for improving numerous image recognition tasks. Recently in natural language processing (NLP), Transformer networks (Vaswani et al., 2017) pre-trained with "masked language model" (MLM) objective (Devlin et al., 2018) on large language corpus excel at a variety of NLP tasks.

Meanwhile, for tasks at the intersection of vision and language, such as image captioning (Young et al., 2014; Chen et al., 2015; Sharma et al., 2018), visual question answering (VQA) (Antol et al., 2015; Johnson et al., 2017; Goyal et al., 2017; Hudson & Manning, 2019), visual commonsense reasoning (VCR) (Zellers et al., 2019; Gao et al., 2019), there lacks such pre-trained generic feature representations. The previous practice is to combine base networks pre-trained for image recognition and NLP respectively in a task-specific way. The task-specific model is directly finetuned for the specific target task, without any generic visual-linguistic pre-training. The task-specific model may well suffer from overfitting when the data for the target task is scarce. Also, due to the task-specific model design, it is difficult to benefit from pre-training, where the pre-training task may well be different from the target. There lacks a common ground for studying the feature design and pre-training of visual-linguistic tasks in general.

In the various network architectures designed for different visual-linguistic tasks, a key goal is to effectively aggregate the multi-modal information in both the visual and linguistic domains. For example, to pick the right answer in the VQA task, the network should empower integrating linguistic information from the question and the answers, and aggregating visual information from the input image, together with aligning the linguistic meanings with the visual clues. Thus, we seek to derive generic representations that can effectively aggregate and align visual and linguistic information.

*Equal contribution. This work is done when Weijie Su and Xizhou Zhu are interns at Microsoft Research Asia.

参考文献

[1] VCR: *https://arxiv.org/pdf/1811.10830.pdf*

[2] LXMERT: *https://arxiv.org/pdf/1908.07490.pdf*

[3] ViBERT: *https://arxiv.org/pdf/1908.02265.pdf*