图文表征的预训练

VL-BERT

Pre-training of Generic Visual-Linguistic Representations

何沧平

2019-9

动机与设计目标

前人成功经验

- 预训练在计算机视觉取得成功，ImageNet上预训练，通用性强，微调到处使用
- NLP领域，2017年出现的Transformer很成功

待解决的困难

- 图像、文本交叉任务上，缺少通用的预训练模型
- Image captioning图像描述，对给定的图片，生成一句描述它的文字
- Visual question answering(VQA)：视觉问答，根据图片选择正确答案
- Visual commonsense reasoning(VCR): 视觉常识推理
- 为单一任务设备设计的模型，容易过拟合，因为训练数据少

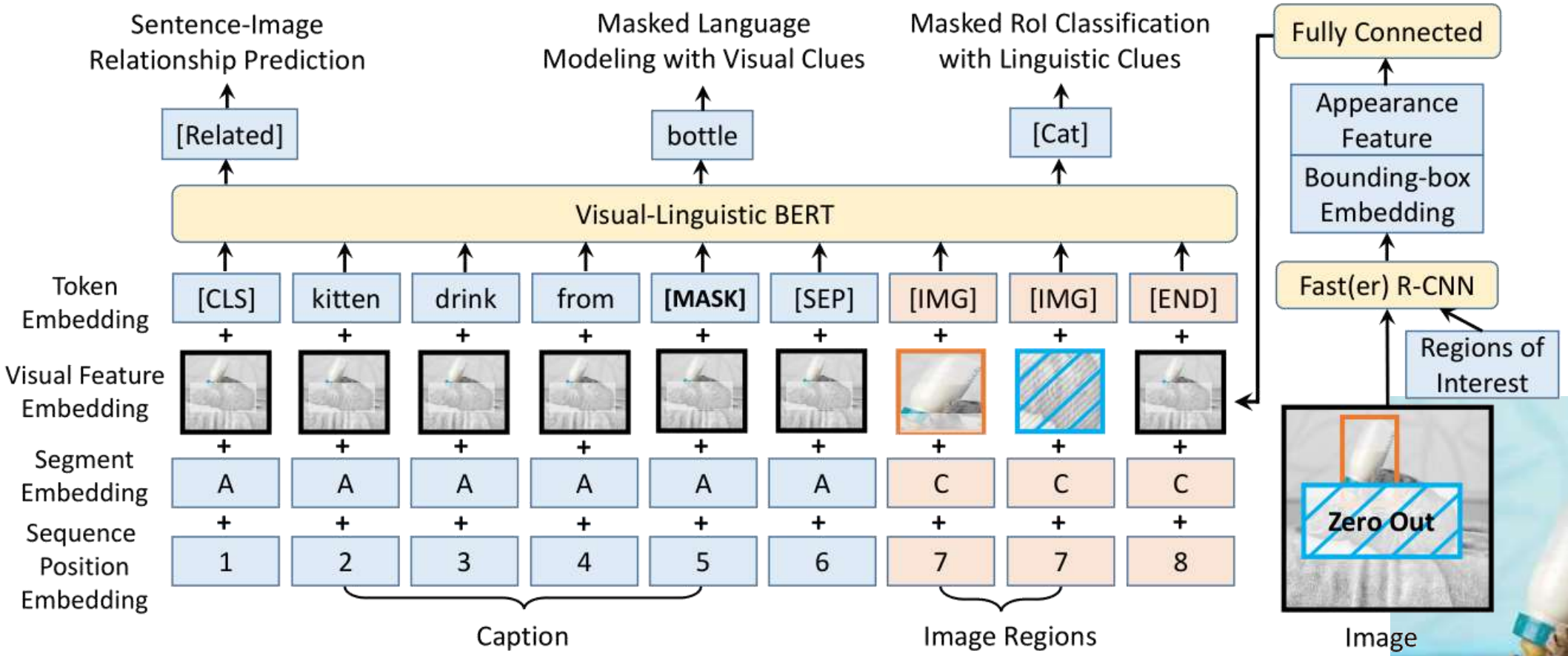
VL-BERT模型的设计目标

- 同时提取文本和图像的信息
- 文本含义和图像中的物体要对应起来

VL-BERT实现路径

- 借鉴transformer attention和BERT的masked language model(MLM)

VL-BERT模型架构



贡献在2种输入、3个目标函数

VL-BERT模型架构

骨架

- 双向transformer编码器，与BERT一样
- 文本输入：与BERT相同
- 图像输入：兴趣区域(RoI，region-of-interest)的特征向量，检测出来的区域、标注的区域都行

模型的输入

- 图像向量，文本向量，分隔符
- 输入开头是一个特殊的类别符[CLS]，接下来是文本、图像，最后是结束符[END]
- 文本的不同句子之间、文本与图像之间插入分隔符[SEP]
- 对每一个元素，其特征向量是四种向量的和：token向量、图像向量、片段向量、序列位置向量
- 图像向量是本文新加的，其它三种与BERT一样

图像向量

- 长相向量与和几何向量连接起来
- 长相特征用Fast R-CNN提取兴趣区域的向量，即输出层前面的那一层的向量。
- 对非图像的(文本)元素，对应的图像是整张图片
- 几何向量代表兴趣区域在整张图片中的位置，左上-右下坐标共4个数值，用Relation Networks转为长度为2048的向量

位置向量

- 图像区域之间没有顺序，因此所有图像元素的位置向量都一样

预训练

数据集

- Conceptual Captions dataset，2018年发布，330万个带标题的图片
- 网络收集，经过了自动清洗，语义清晰、流畅
- VL-BERT输入形式是<Caption, Image>

目标函数1：句子-图片关系预测

- 图像问答、基于标题的图片检索（Caption-based Image Retrieval)要求理解句子与图片的关系，例如图文是否相关
- 从数据集中随机选取标题和图片，50%的概率是正确对应，标记为[Related]；50%的概率不正确对应，标记为[NotRelated]，见架构图
- 在最终输出层的[CLS]元素上添加一个二分类器，sigmoid交叉熵损失

目标函数2：带视觉线索的遮挡语言模型

- 与BERT的遮挡语言模型几乎一样
- 只是图片向量也参加了训练
- 希望找到词与图像区域的对应关系
- kitten drinking from [MASK]中的遮挡位置，可以是任意的容器， bowl/spoon/bottle都可以。
- 添加bottle的图像信息，希望能找到bottle

目标函数3：带文本线索的遮挡图像模型

- 与函数2对偶
- 15%概率随机遮挡兴趣区域，根据其它信息去预测该区域的标签（标签有很多）
- 遮挡所用的图像向量也需要学习得到
- 为避免被遮挡区域的像素泄露，先将该区域的像素全部置0，再对其它兴趣区域应用Fast R-CNN
- 见架构图
- 这是个多分类问题

微调

- 给定恰当的输入，端到端微调网络参数就可以了
- 典型输入形式为<Caption, Image>， <Question, Answer, Image>
- 输出：[CLS]用来预测图文关系，词的向量、兴趣区域的向量用来做词预测和区域预测

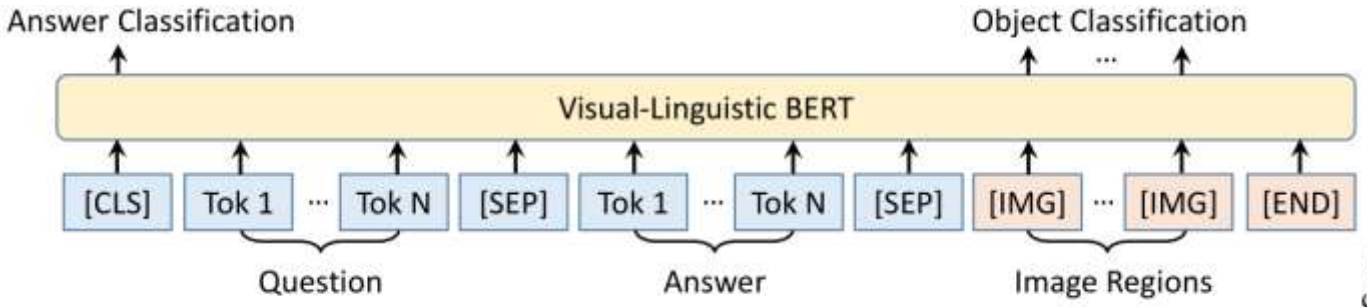
实验1/3

预训练

- Conceptual Captions dataset
- 参数初始值使用BERT的值。新添加的参数，用高斯分布初始化
- 每张图片至多有100个兴趣区域的检测得分超过0.5。无论得到的阈值是多少，每张图片至少拿出10个兴趣区域

下游任务微调（VCR）

- 视频常识推理，给定一张图和一堆已经归类的兴趣区域，提出一个认识水平的问题。
- 模型需要选出正确的答案，并给出选择理由。每个问题，有4个备选答案和4个备选理由。



(a) Input and output format for Visual Commonsense Reasoning (VCR) dataset

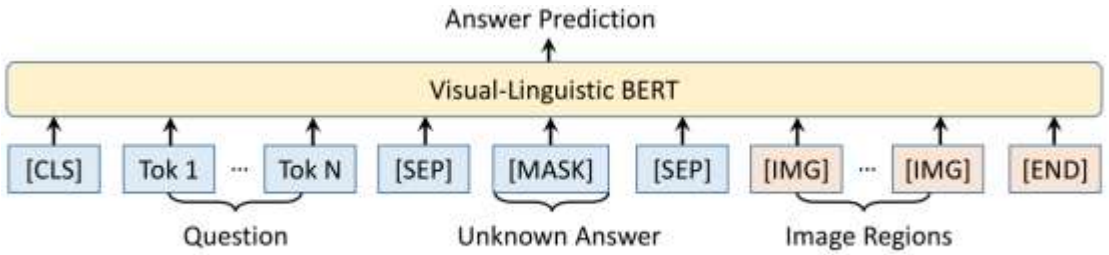
- 该任务（Q→AR）可分解为2个子任务：问答（Q→A）和答案判断（QA→R）
- VL-BERT输入格式<Question, Answer, Image>，对问答子任务，Q和A分别填入Question和Answer位置。对子答案判断子任务，将Q和A连接起来填入Question位置，R填入Answer位置。
- [IMG]处填入已经标出的兴趣区域的向量
- 将输出的[CLS]传入一个softmax分类器，用来判定答案是否正确
- 扔掉了另外2个目标函数
- 预训练仅提高0.2%。因为训练数据是认识实体，任务是推理，差别大

Model	Q → A		QA → R		Q → AR	
	val	test	val	test	val	test
R2C (Zellers et al., 2019)	63.8	65.1	67.2	67.3	43.1	44.0
ViLBERT (Lu et al., 2019)†	72.4	73.3	74.5	74.6	54.0	54.8
VisualBERT (Li et al., 2019b)†	70.8	71.6	73.2	73.2	52.2	52.4
B2T2 (Alberti et al., 2019)†	71.9	72.6	76.0	75.7	54.9	55.0
VL-BERT w/o pre-training	73.5	-	74.4	-	54.8	-
VL-BERT	73.7	74.0	74.5	74.8	55.0	55.5

2: Results compared to the state-of-the-art methods with single model on VCR dataset. cates concurrent works.

实验2/3 视觉问答

- 给定一张图片，问一个**感知**层次的问题，算法的任务是生成或选择一个答案
- VQA v2.0 dataset，有一个答案池子，只需选择答案
- 输入格式<Question, Answer, Image>，只在答案中随机添加[MASK]
- 在最终输出的[MASK]上建立一个多分类器，交叉熵损失



(b) Input and output format for Visual Question Answering (VQA) dataset

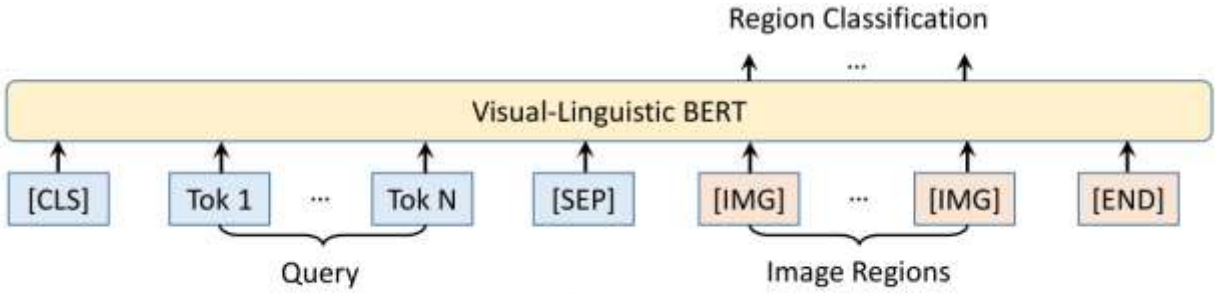
- 预训练版性能会高0.02%，没效果
- 比BUTD模型好很多，比同期模型差一些

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 (Hao Tan, 2019)†	72.42	72.54
VL-BERT w/o pre-training	69.58	-
VL-BERT	70.50	70.83

实验3/3 Referring Expression指示表达

- 描述图片中某一特定物体的一句自然语句。具有唯一性和区域性，比如「穿红色毛衣的女人」或「右边的男孩」，比图像描述更细致。
- 输入格式： <Query, Image>
- 目标函数：对每个兴趣区域做二分类，预测时，挑得分最高的那个兴趣区域

- 比MAttNet高出6%，比ViBERT差一些
- 原因：预训练就是用的相似任务，将文本与兴趣区域对应起来了



(c) Input and output format for Referring Expression task on RefCOCO+ dataset

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 w/o pre-training	73.96	76.65	67.64	65.92	72.30	55.55
VL-BERT	78.44	81.30	71.18	71.84	77.59	60.57

相关工作

	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 (Hao Tan, 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
Works in Progress	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)	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

‡ LXMERT is pre-trained on COCO Caption (Chen et al., 2015), VG Caption (Krishna et al., 2017), VG QA (Zhu et al., 2016), VQA (Antol et al., 2015) and GQA (Hudson & Manning, 2019).

这些模型思路类似。
这些作者相互熟悉，发表之前就获得别人的性能数据。

视频文本表征的联合模型

VideoBERT

A Joint Model for Video and Language Representation Learning

何沧平

2019-9

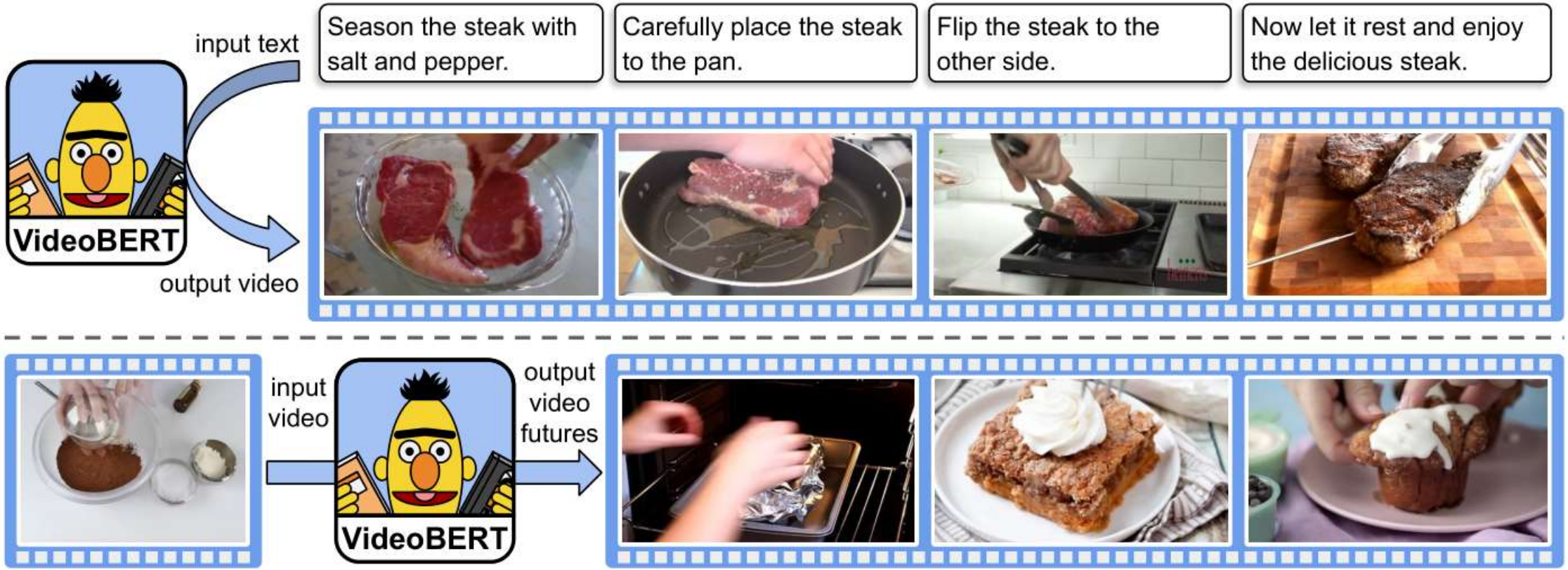
论文设计目标

前人成功经验

- BERT，自动语音识别
- 可以学习到文本、视频的低层特征（时长1秒以下）

目标

- 学习高层特征，例如大时长（几分钟）、不仅学习物品而且识别动作
- 文本与视频对应



VL-BERT模型架构

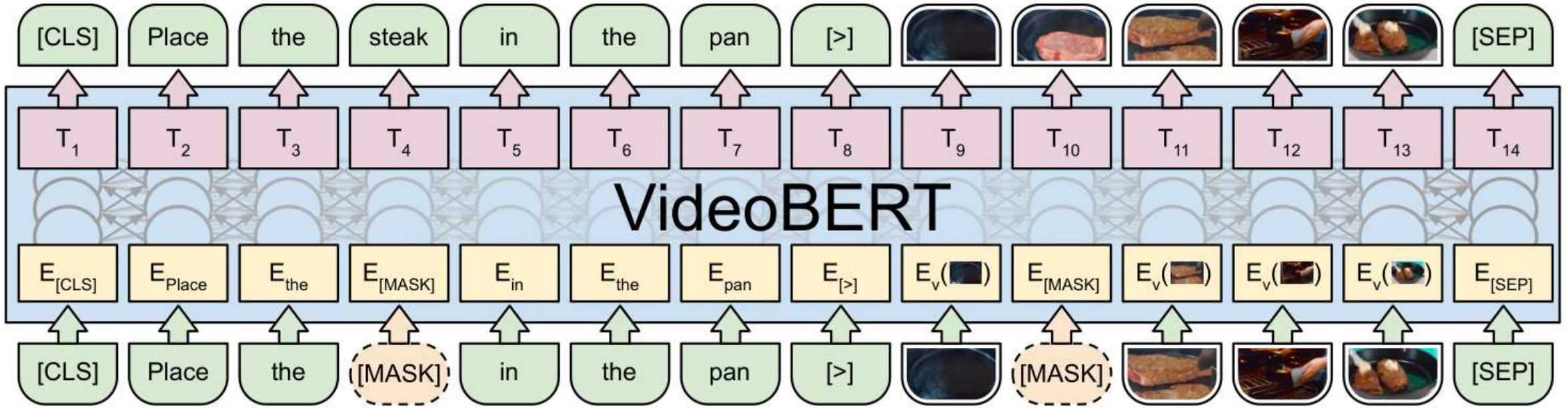


Figure 3: Illustration of VideoBERT in the context of a video and text masked token prediction, or *cloze*, task. This task also allows for training with text-only and video-only data, and VideoBERT can furthermore be trained using a linguistic-visual alignment classification objective (not shown here, see text for details).

贡献：如何制作视频输入

videoBERT模型架构

骨架

- 双向transformer编码器，与BERT一样
- 文本输入：与BERT相同
- 视频输入：视觉单词（详见后）

```
[CLS] orange chicken with [MASK]
sauce [>] v01 [MASK] v08 v72 [SEP]
```

- V01/ v08/ v72是视觉单词，[>]用来分隔文本和视频，[SEP]是为了保持和BERT一致
- [CLS]表示文本与视频是否对应。这个对应噪音多的情形：视频声音里提到的事物可能不在视频里。

训练模式

- 只用文本：使用BERT的遮挡目标函数
- 只用视频：使用BERT的遮挡目标函数
- 文本+视频：使用对齐任务的目标函数（[CLS]上的二分类）

实验与分析

数据集

- Youtube烹饪视频，带有cooking/ recipe标签的
- 时长小于15分钟，个数312K，累计时长23186小时（966天）
- 比第二大的YouCook II数据集大得多，它的数量2K，时长176小时
- 使用youtube的自动语音识别API提取文本，120K个视频有英文文本
- 仅视频训练时，用312K个视频，仅文本/文本+视频训练时，用120K个视频。在YouCook II上评估

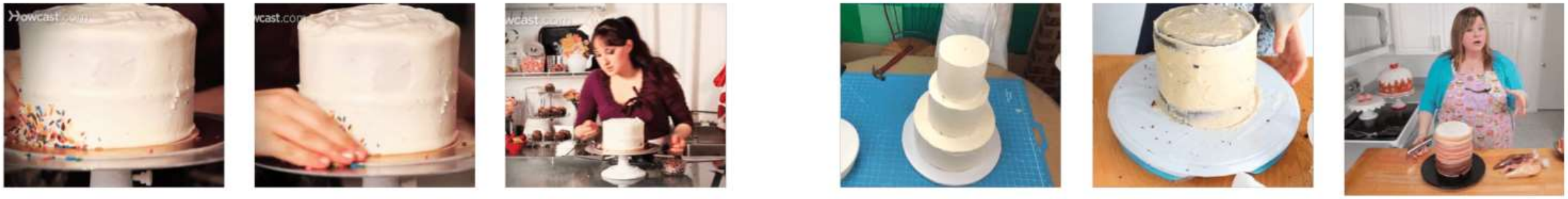
预训练

- BERT_large的权重进行初始化
- 字典中添加 20736个视频单词
- 4个TPU上训练2天

预处理

- 视频按20 fps取样，裁成长度30帧的片段（1.5秒）
- 使用S3D模型来提取视频片段的特征，1024维
- 对视频特征，用层次k-means聚类，共20736类。将类中心做为视频单词的向量
- 对语音识别得到的单词序列，使用一个现成的LSTM语言模型将其分割成一堆句子。句子中单词的处理，跟BERT一样（WordPiece）
- 视频句子：语音识别出来的每个句子都有时间戳，这段时间内的视频片段构成一个句子。没有语音的，16个片段一个句子

视频单词对应的类中心（右）



“but in the meantime, you're just kind of moving around your cake board and you can keep reusing make sure you're working on a clean service so you can just get these all out of your way but it's just a really fun thing to do especially for a birthday party.”

“apply a little bit of butter on one side and place a portion of the stuffing and spread evenly cover with another slice of the bread and apply some more butter on top since we're gonna grill the sandwiches.”

Figure 4: Examples of video sentence pairs from the pretraining videos. We quantize each video segment into a token, and then represent it by the corresponding visual centroid. For each row, we show the original frames (left) and visual centroids (right). We can see that the tokenization process preserves semantic information rather than low-level visual appearance.

实验与分析：zero-shot动作分类

- zero-shot：全新数据集，视频和相应的标签都没有用来训练
- 预测YouCook II中视频对应的动词和名词
- 文本句子模板：`now let me show you how to [MASK] the [MASK]`
- 为定量评估，使用别人标记好边界框的63个常见物品。从边界框的描述(caption)中抽取45个常见动词、100个常见名词。右图是3个预测出来动词名词，下表是定量性能。
- Top1动词性能比监督训练（S3D）差，top5相当

Top verbs: make, assemble, prepare
Top nouns: pizza, sauce, pasta

Top verbs: make, do, pour
Top nouns: cocktail, drink, glass

Top verbs: make, prepare, bake
Top nouns: cake, crust, dough

Method	Supervision	verb top-1 (%)	verb top-5 (%)	object top-1 (%)	object top-5 (%)
S3D [34]	yes	16.1	46.9	13.2	30.9
BERT (language prior)	no	0.0	0.0	0.0	0.0
VideoBERT (language prior)	no	0.4	6.9	7.7	15.3
VideoBERT (cross modal)	no	3.2	43.3	13.1	33.7

Table 1: Action classification performance on YouCook II dataset. See text for details.

视频描述 Transfer learning for captioning

- 将videoBERT用作特征抽取器
- 输入只有视频句子，凭空添加一个文本句子，文本句子用这个模板 `now let's [MASK] the [MASK] to the [MASK], and then [MASK] the [MASK]`
- 提取出视频对应的向量、被遮挡单词的向量，分别算平均，再将2个平均向量连接起来
- 用videoBERT提取的向量来代替某篇论文中的向量，任务是视频描述，然后对比效果
- S3D得到的向量用到对比基准。
- 还可以将videoBERT特征向量与S3D平均向量连接起来
- videoBERT将性能提高了一些

Method	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Zhou *et al.* [39]	-	1.42	11.20	-	-
S3D [34]	6.12	3.24	10.00	26.05	0.35
VideoBERT	6.80	4.07	10.99	27.51	0.50
VideoBERT + S3D	**7.81**	**4.52**	**11.85**	**28.78**	**0.55**

Caption示例，极少有完全正确的


GT: add some chopped basil leaves into it

VideoBERT: chop the basil and add to the bowl

S3D: cut the tomatoes into thin slices


GT: cut the top off of a french loaf

VideoBERT: cut the bread into thin slices

S3D: place the bread on the pan


GT: cut yu choy into diagonally medium pieces

VideoBERT: chop the cabbage

S3D: cut the roll into thin slices


GT: remove the calamari and set it on paper towel

VideoBERT: fry the squid in the pan

S3D: add the noodles to the pot

评价与展望

- 作者明确表示，这只是文本-视频联合训练的第一步，还有很多改进点
- 可能应用到微博美食频道上，添加视频描述，提取物品，聚类
- 经验：联合训练关键是如何得到高质量的视频特征

B2T2

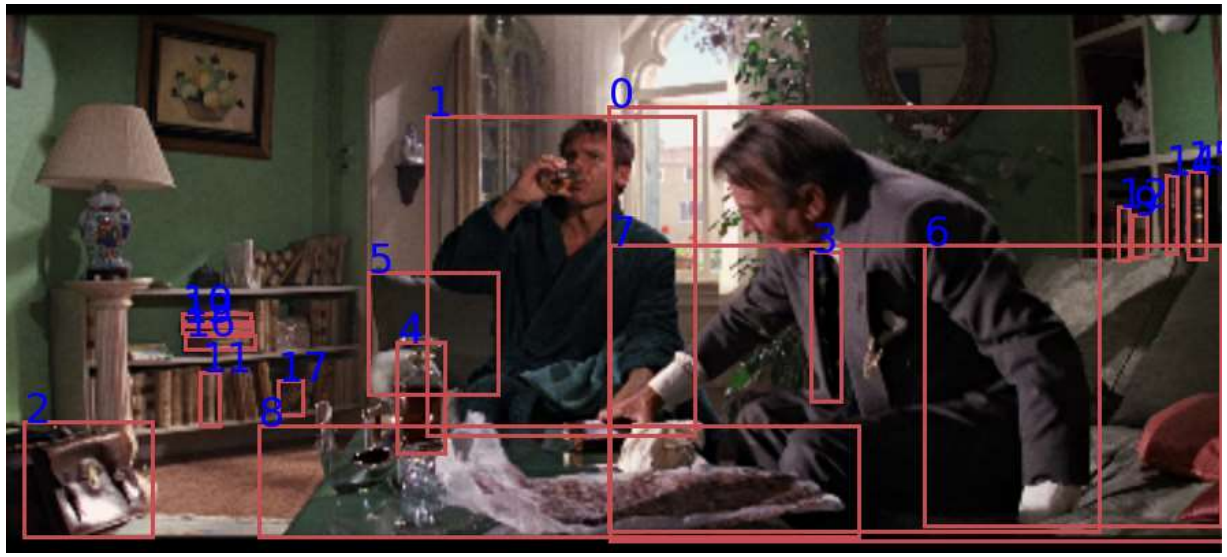
Fusion of Detected Objects in Text for Visual Question Answering

何沧平

2019-9

动机与设计目标

- 预训练模型，用于视觉常识推理（VCR）
- 借用BERT的成果



Q: What was [1] doing before he sat in his living room?

A_1: He was reading [10].
A_2: He was taking a shower. ✓
A_3: [0] was sleeping until the noise [1] was making woke him up.
A_4: He was sleeping in his bedroom.

R_1: His clothes are disheveled and his face is glistening like he's sweaty.
R_2: [0] does not look wet yet, but [0] looks like his hair is wet, and bathrobes are what you wear before or after a shower.
R_3: He is still wearing his bathrobe. ✓
R_4: His hair appears wet and there is clothing hanging in front of him on a line as if to dry.

双编码器

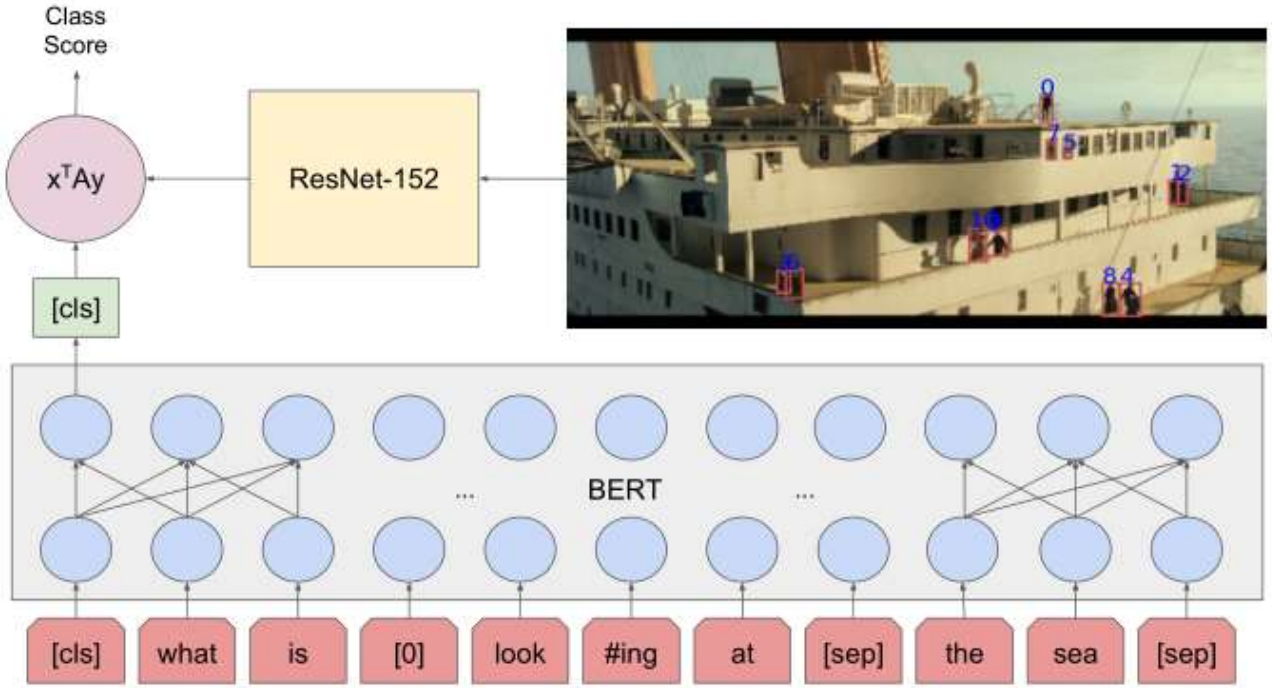


Figure 2: Dual Encoder architecture with late fusion. The model extracts a single visual feature vector from the entire image. Bounding boxes are ignored.

将BERT输出的[CLS]向量与图片向量相乘
不要求2个向量维度相同
矩阵A待学习
以图文是否匹配进行二分类
非本文原创，未提细节

B2T2

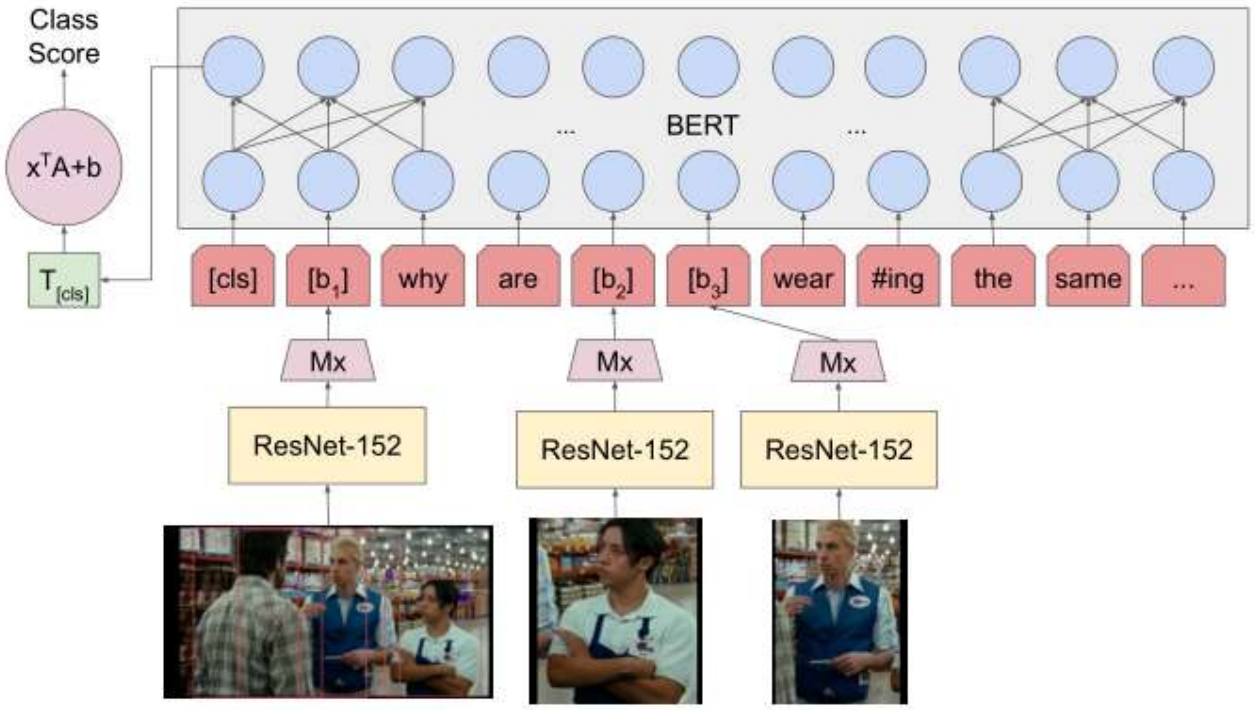


Figure 3: B2T2 architecture with early fusion. Bounding boxes are inserted where they are mentioned in the text and at the end of the input, as described in Sec. 4.

Bounding Boxes in Text Transformer

- 在句子里提到的图像位置处，插入边界框(bounding box)向量+框内图像的向量
- 边界框形式：左下坐标-高-宽，嵌入矩M阵待学习
- 二分类，交叉熵损失

预训练

- 任务1：遮挡语言模型，像BERT一样
- 任务2：真假图像，使用假图像构造分类负样本
- 数据用image-caption对

预训练图示

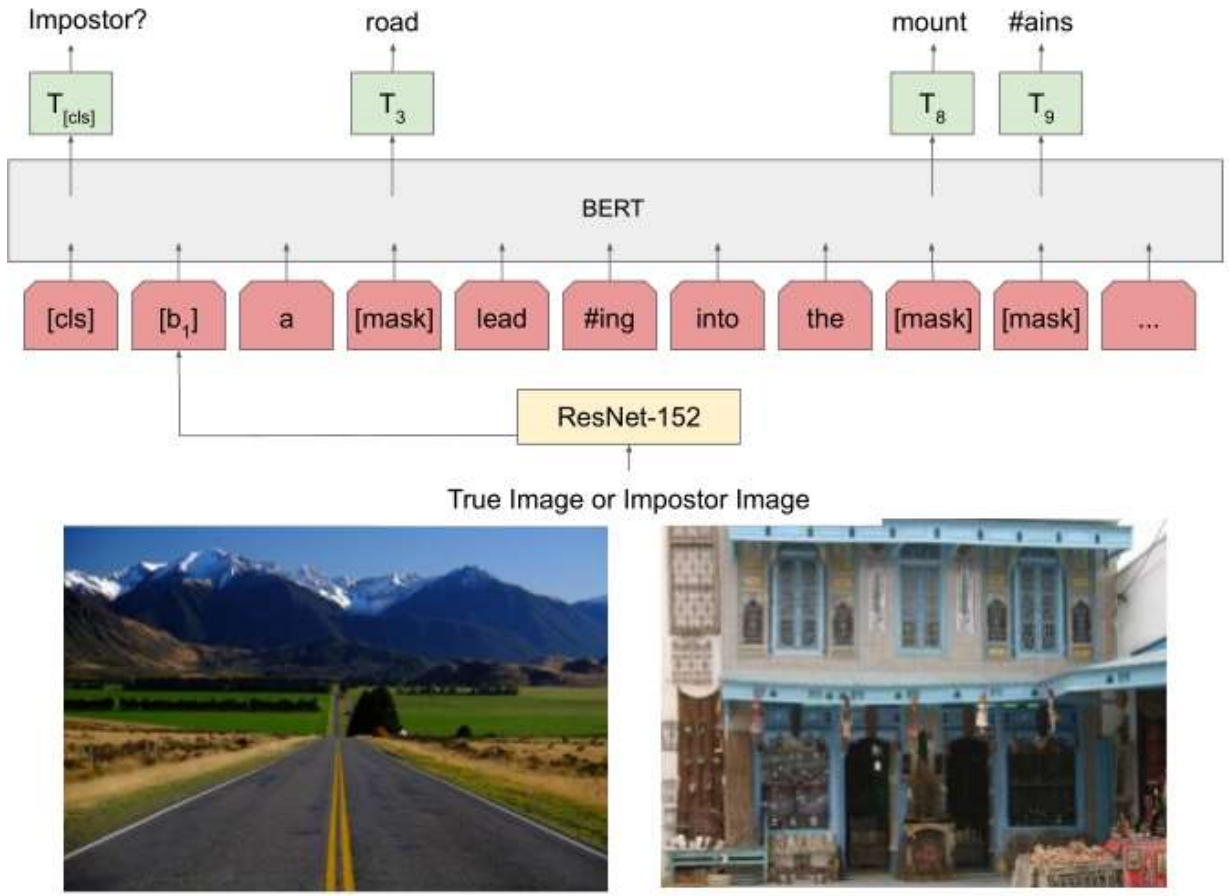


Figure 5: Mask-LM pretraining for B2T2.

性能及消融实验

Model	$Q \rightarrow A$ Val	$Q \rightarrow A$ Test	$QA \rightarrow R$ Val	$QA \rightarrow R$ Test	$Q \rightarrow AR$ Val	$Q \rightarrow AR$ Test
Chance	25.0	25.0	25.0	25.0	6.2	6.2
Text-Only BERT (Zellers et al.)	53.8	53.9	64.1	64.5	34.8	35.0
R2C (Zellers et al.)	63.8	65.1	67.2	67.3	43.1	44.0
MUGRN (subm. NeurIPS)	-	68.2	-	69.4	-	47.5
SNU MRCNet (unpub.)	-	68.4	-	70.5	-	48.4
UTS CCD (unpub.)	-	68.5	-	70.5	-	48.4
ResNet50-BERT (unpub.)	-	70.1	-	70.8	-	49.8
B-VCR (unpub.)	-	70.4	-	71.5	-	50.2
Text-Only BERT (ours)	59.5	-	65.6	-	39.3	-
Dual Encoder (ours)	66.8	-	67.7	-	45.3	-
B2T2 (ours)	71.9	72.6	76.0	75.7	54.9	55.0
B2T2 Ensemble (ours)	**73.2**	**74.0**	**77.1**	**77.1**	**56.6**	**57.1**
Human		91.0		93.0		85.0

Table 2: Experimental results on VCR, incorporating those reported by Zellers et al. (2019). The proposed B2T2 model and the B2T2 ensemble outperform published and unpublished/undocumented results found on the VCR leaderboard at visualcommonsense.com/leaderboard as of 5/21/2019.

	$Q \rightarrow A$
Dual Encoder	66.8
No bboxes	67.5
Late fusion	68.6
BERT-Base	69.0
ResNet-50	70.4
No bbox class labels	70.9
Fewer appended bboxes ($p = 4$)	71.0
No bbox position embeddings	71.6
Full B2T2	71.9

Table 3: Ablations for B2T2 on VCR dev. The Dual Encoder and the full B2T2 models are the main models discussed in this work. All other models represent ablations from the full B2T2 model.

- 插入边界框最重要，在输入层融合次重要
- BERT-large比BERT-base高2.9%
- 图像嵌入方法，ResNet-152比ResNet-50高1.5%
- 边界框的位置向量影响0.3%

学习跨模态编码表征

LXMERT:
Learning Cross-Modality Encoder Representations from Transformers

何沧平

2019-9

摘要

设计目标与成绩

- 理解视觉概念，理解文本语义，将二者对应起来
- 通用模型，预训练 - 微调 路线
- 下游任务：图像问答
- 成绩：比当时最好的模型提高22%

设计思路

- 三个编码器，文本一个，图像一个，交叉一个
- 将文本与图像嵌入同一个向量空间，并像BERT那样的[CLS]来对齐文本与图像
- 5个任务做目标函数，比BERT多3个

LXMERT模型架构

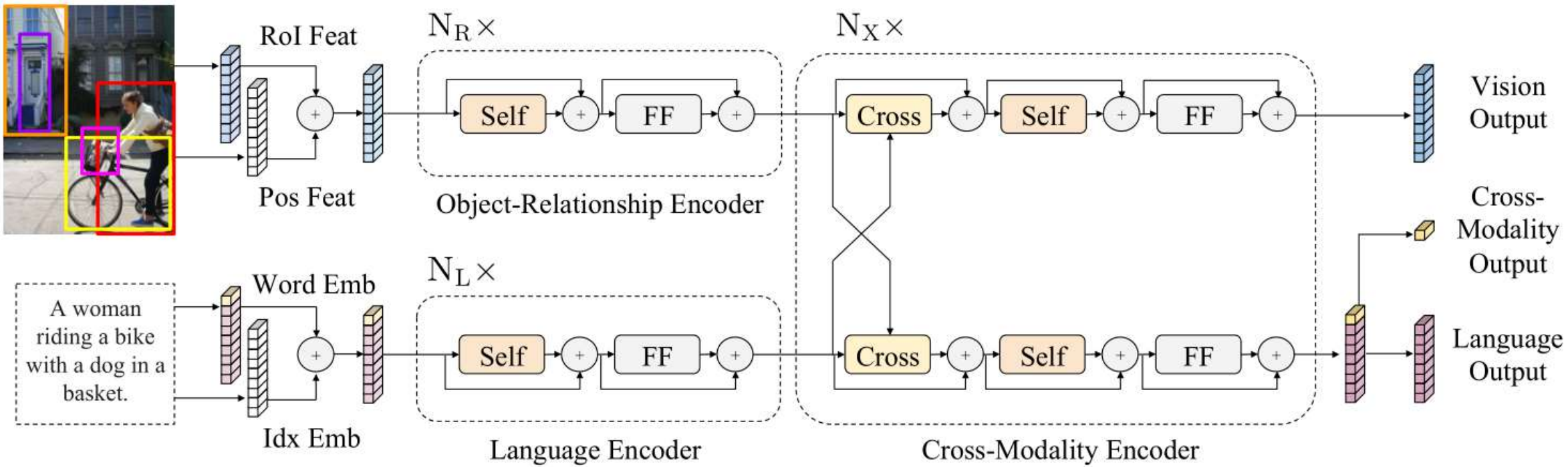


Figure 1: The LXMERT model for learning vision-and-language cross-modality representations. 'Self' and 'Cross' are abbreviations for self-attention sub-layers and cross-attention sub-layers, respectively. 'FF' denotes a feed-forward sub-layer.

+：残差和layernorm

交叉编码器接受2种输入

Cross-modality output: 对齐标记

LXMERT模型架构

输入：文本句子向量

- 使用WordPiece分词，将词向量和位置相加，layernorm变换，得到每个词的向量

$$\hat{w}_i = \text{WordEmbed}\left(w_i\right)$$
$$\hat{u}_i = \text{IdxEmbed}\left(i\right)$$
$$h_i = \text{LayerNorm}\left(\hat{w}_i + \hat{u}_i\right)$$

输入：图像向量

- Faster R-CNN检测到的兴趣区域(region-of-interest, RoI)，位置特征p_j，兴趣区域特征f_j，变换加和，得到单个区域的特征
- 按Faster R-CNN的输出顺序，组成图像句子
- Layernorm是为了平衡两种信息，防止一大一小

$$\hat{f}_j = \text{LayerNorm}\left(W_{\text{F}} f_j + b_{\text{F}}\right)$$
$$\hat{p}_j = \text{LayerNorm}\left(W_{\text{P}} p_j + b_{\text{P}}\right)$$
$$v_j = \left(\hat{f}_j + \hat{p}_j\right)/2$$

单模态编码器

- 文本编码器和图像编码器各用一个transformer
- 全连接层有2层

跨模态编码器

- 图象向量和文本向量，交替用作“上下文向量”和“查询向量”

$$\hat{h}_i^k = \text{CrossAtt}_{\text{L}\rightarrow\text{R}}\left(h_i^{k-1}, \{v_1^{k-1}, \ldots, v_m^{k-1}\}\right)$$
$$\hat{v}_j^k = \text{CrossAtt}_{\text{R}\rightarrow\text{L}}\left(v_j^{k-1}, \{h_1^{k-1}, \ldots, h_n^{k-1}\}\right)$$

输出：3种

- 文本输出，图像输出
- 图中小黄块 [CLS]是跨模态输出

预训练任务

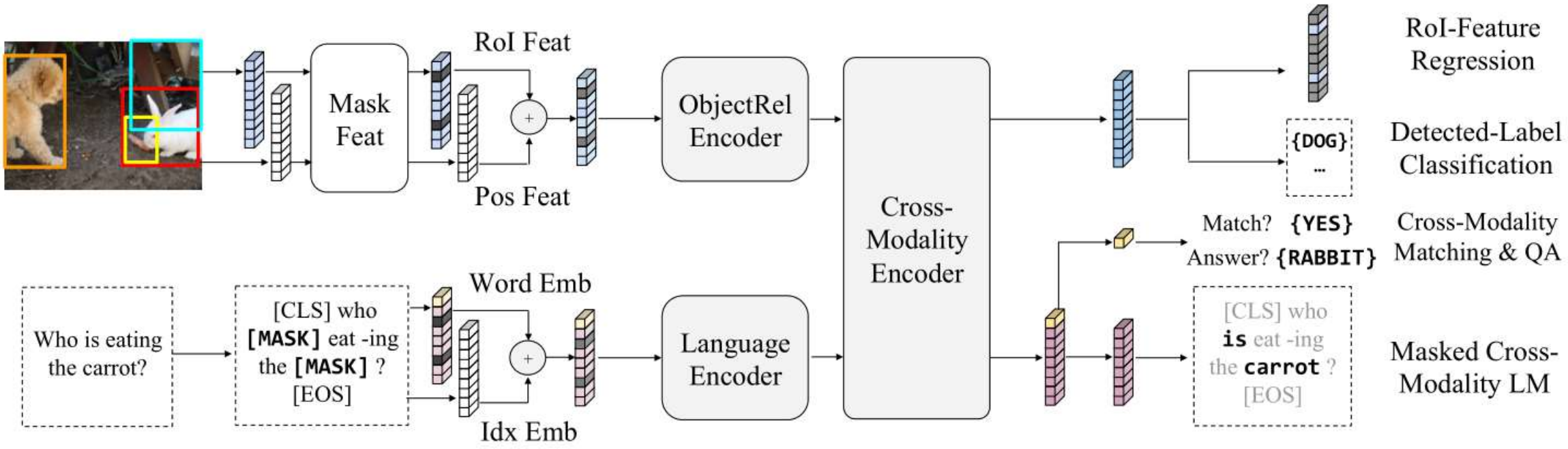


Figure 2: Pre-training in LXMERT. The object RoI features and word tokens are masked. Our five pre-training tasks learn the feature representations based on these masked inputs. Special tokens are in brackets and classification labels are in braces.

预训练

语言任务：跨模态遮挡语言模型

- 遮挡方法跟BERT一样
- 不能用BERT的参数初始化，否则性能下降

视视觉任务：遮挡物体

- 跟BERT类似，遮挡物体，只是输入没有了[CLS]
- 兴趣区域特征回归，用L2损失
- 用目标检测的标签做分类，交叉熵损失

跨模态任务

- 匹配：将图象句子和文本句子，以0.5概率不匹配，训练[CLS]，二分类，交叉熵。与BERT下一句任务类似
- 图像问答：为扩大预训练数据集，三分之一的句子是图像问答。当图像和问题匹配时，让模型去预测答案。

所用数据集

- MS COCO, Visual Genome, VQA v2.0, GQA balanced version, VG-QA

Image Split	Images	Sentences (or Questions)					
		COCO-Cap	VG-Cap	VQA	GQA	VG-QA	All
MS COCO - VG	72K	361K	-	387K	-	-	0.75M
MS COCO ∩ VG	51K	256K	2.54M	271K	515K	724K	4.30M
VG - MS COCO	57K	-	2.85M	-	556K	718K	4.13M
All	180K	617K	5.39M	658K	1.07M	1.44M	9.18M

Table 1: Amount of data for pre-training. Each image has multiple sentences/questions. 'Cap' is caption. 'VG' is Visual Genome. Since MS COCO and VG share 51K images, we list it separately to ensure disjoint image splits.

预训练过程与性能

- 文本句子用WordPieces分割
- Faster R-CNN提取36个兴趣区域
- N_L = 9, N_X = 5, N_R=5
- 隐层大小为768，与BERT_BASE相同
- 从头训练，参数随机初始化或置为0
- 训练20轮，在4块 Titan Xp上跑10天

- 在图像问答任务上性能很好
- 其它任务上性能，论文未提

Method	VQA				GQA			$NLVR^2$	
	Binary	Number	Other	**Accu**	Binary	Open	**Accu**	Cons	**Accu**
Human	-	-	-	-	91.2	87.4	89.3	-	96.3
Image Only	-	-	-	-	36.1	1.74	17.8	7.40	51.9
Language Only	66.8	31.8	27.6	44.3	61.9	22.7	41.1	4.20	51.1
State-of-the-Art	85.8	53.7	60.7	70.4	76.0	40.4	57.1	12.0	53.5
LXMERT	**88.2**	**54.2**	**63.1**	**72.5**	**77.8**	**45.0**	**60.3**	**42.1**	**76.2**

Table 2: Test-set results. VQA/GQA results are reported on the 'test-standard' splits and $NLVR^2$ results are reported on the unreleased test set ('Test-U'). The highest method results are in bold. Our LXMERT framework outperforms previous (comparable) state-of-the-art methods on all three datasets w.r.t. all metrics.

评价与展望

- 新点子，图像向量与文本向量直接运算，嵌入同一个空间
- 跨模态交叉作用机理没提，不清楚

- 图像兴趣区域的种类受限于Faster R-CNN，难以直接应用到微博业务
- 将文本与图像对齐的思路，都面临图像种类太小的问题，扩大数量的话，需要先扩大目标检测模型

ViLBERT

Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

何沧平

2019-9

动机与设计目标

- 得到一个统一的预训练模型，以便用于下游任务（视觉问答，视频常识推理，指示表达referring expressions， 基于描述的图像检索caption-based image retrieval）
- 改造BERT模型，理解图像、图文对齐

模型架构1/2

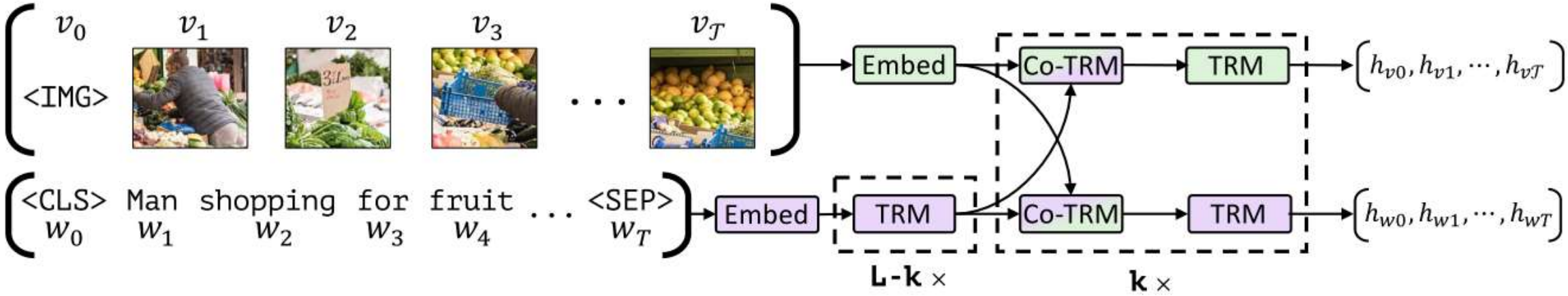
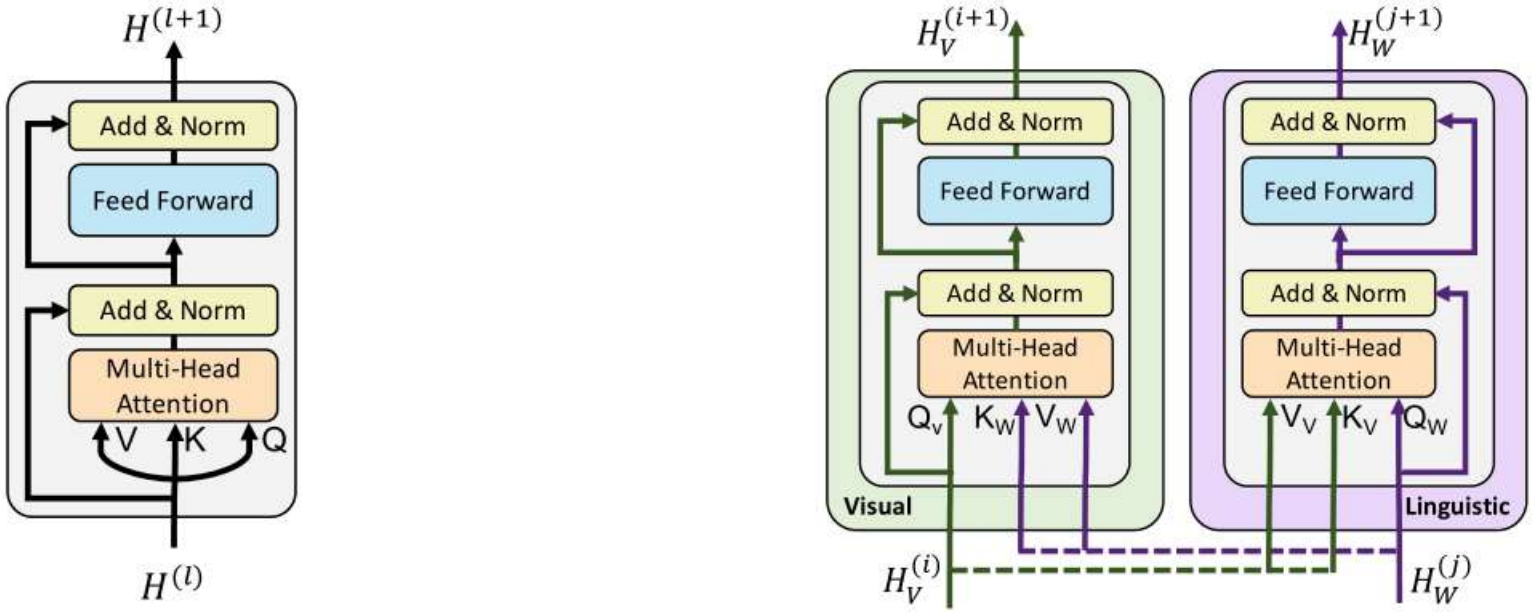


Figure 1: Our ViLBERT model consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers. This structure allows for variable depths for each modality and enables sparse interaction through co-attention. Dashed boxes with multiplier subscripts denote repeated blocks of layers.

TRM是标准的transformer
Co-TRM是本文核心部件（不是原创），细节见下页

模型架构2/2



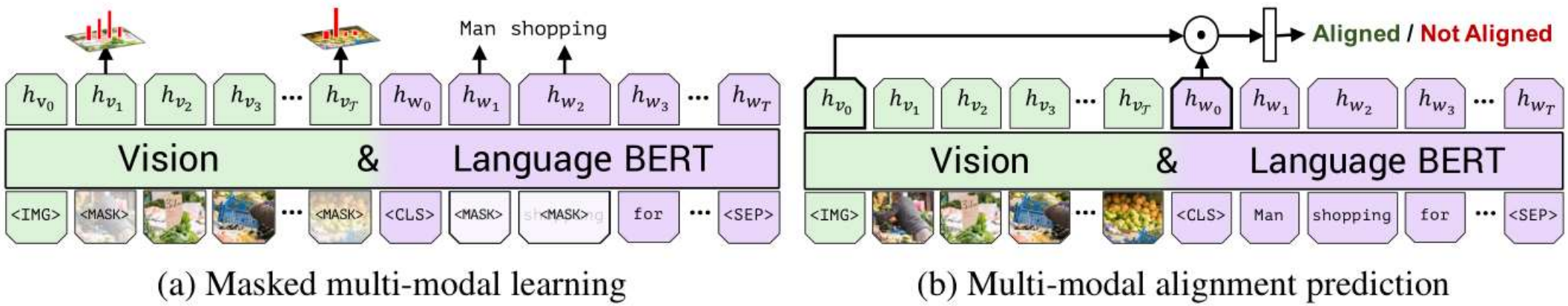
(a) Standard encoder transformer block

(b) Our co-attention transformer layer

Figure 2: We introduce a novel co-attention mechanism based on the transformer architecture. By exchanging key-value pairs in multi-headed attention, this structure enables vision-attended language features to be incorporated into visual representations (and vice versa).

Transformer对输入的向量，会计算出value/key/query
Co-TRM接收视觉向量H_V和文本向量H_W，将向量value/key给对方使用

预训练（2个任务）



(a) Masked multi-modal learning (b) Multi-modal alignment prediction

Figure 3: We train ViLBERT on the Conceptual Captions [24] dataset under two training tasks to learn visual grounding. In masked multi-modal learning, the model must reconstruct image region categories or words for masked inputs given the observed inputs. In multi-modal alignment prediction, the model must predict whether or not the caption describes the image content.

- 一个遮挡任务，一个对齐任务
- 预训练数据集是图片描述

预训练

图片输入

- 图片向量：Faster R-CNN(骨干ResNet- 101)抽取图片的兴趣区域向量，1024维。设计一下阈值，选出10~36个
- 位置：兴趣区域的左上-右下坐标，和占整个图像的面积比例，5个数字，映射到1024维向量空间(映射方法没提)，然后与兴趣区域向量加起来

目标任务1：遮挡多模态模型

- 文本的遮挡方法与BERT相同
- 图片的遮挡：图片中被遮挡区域全部置0.
- 目标函数是KL散度，而不是回归。具体做法没提

目标任务2：图文对齐

- 2个特殊输入[IMG]和[CLS]，分别图像和文本 $\{\mathrm{IMG}, v_1, \ldots, v_T, \mathrm{CLS}, w_1, \ldots, w_T, \mathrm{SEP}\}$
- 对应输出向量为h_IMG和h_CLS，按元素相乘，得到的向量用于二分类，指示图文是否匹配
- 负样本构造：随机替换成不匹配的图片或者文本（caption）

超参数

- Batchsize 512, 数据跑10轮，Adm优化器
- 8 块 Titan X GPU
- 两个损失函数权重相同

visualBERT

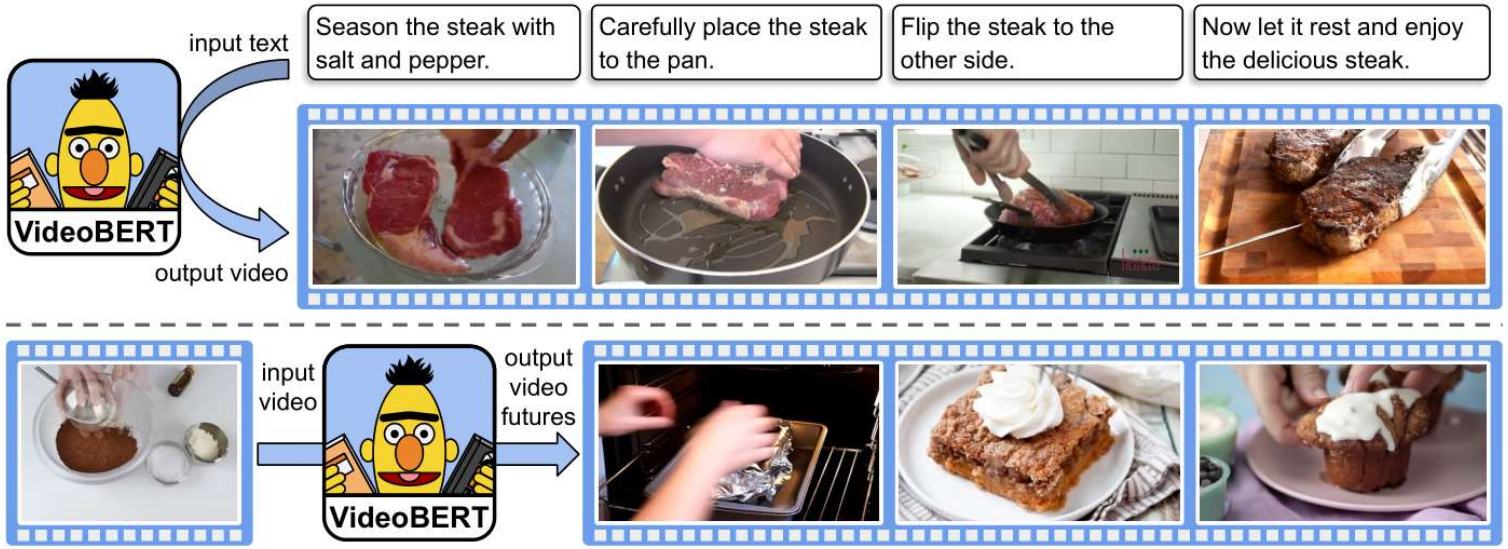
A Simple and Performant Baseline for Vision and Language

何沧平

2019-9

论文设计目标

- 通用的图文预训练模型，用于下游任务 VQA/VCR/NLV2/Flickr30K
- VQA视频问答，VCR视觉常识推理，VLV2图文联合推理， Flickr30K边界框内的图像描述
- 与videoBERT模型类似，只是应用范围不再局限于烹饪



模型架构

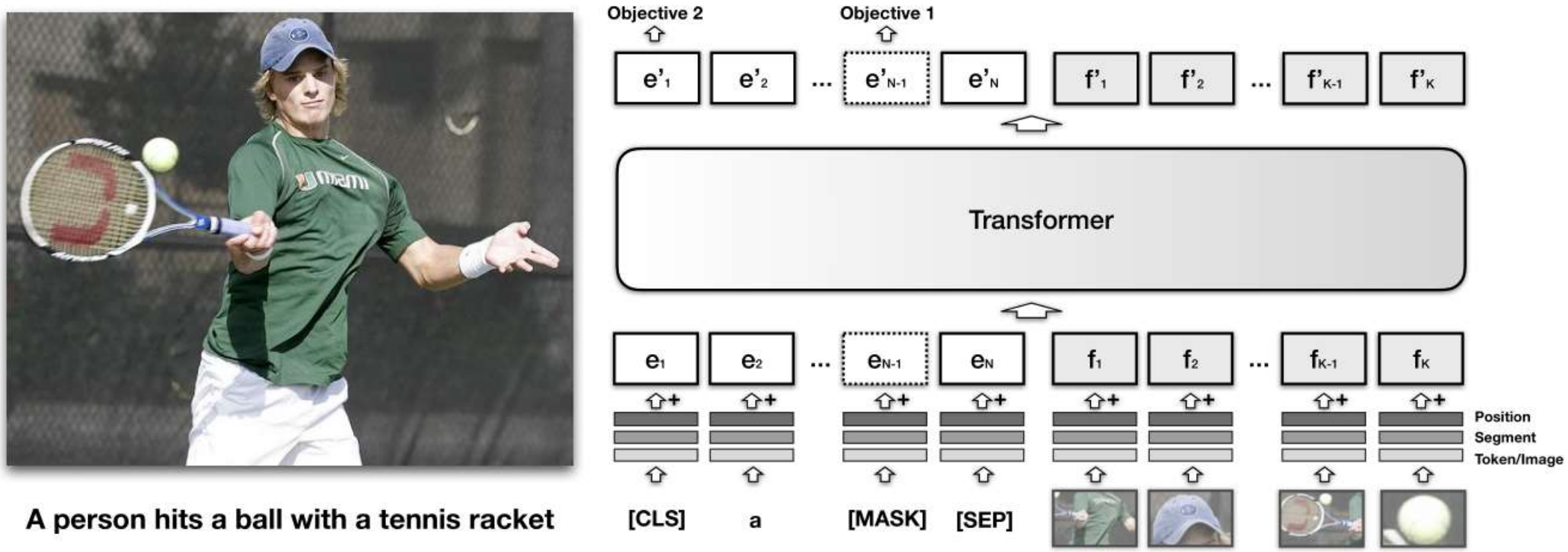


Figure 2: The architecture of VisualBERT. Image regions and language are combined with a Transformer to allow the self-attention to discover implicit alignments between language and vision. It is pre-trained with a masked language modeling (Objective 1), and sentence-image prediction task (Objective 2), on caption data and then fine-tuned for different tasks. See §3.3 for more details.

- 最简单的图文融合模型，只是把BERT中的下句换成兴趣区域

模型架构

输入：图像向量

- 是3个向量的和
- 兴趣区域向量，由Faster R-CNN或其它卷积网格提供（随数据集而变）
- Segment embedding: 表明是图像而不是文本。
- 位置向量：论文中没说怎么处理

输入：文本向量

- 与BERT一样，只是遮挡训练的时候要带上图像向量

目标任务1：遮挡语言模型

- 遮挡方法与BERT一样，此时不遮挡图像

目标任务2：句子-图像预测

- 对COCO数据集训练，1个图片有5个描述
- 对一张图片，一次使用2个描述，其中一个描述是正确的，另一个描述有50%是正确的，50%是随机的。类似BERT的上下句训练方法。

评价

- 思路简单
- 实验中，部分性能挺好
- 论文中缺少一些关键细节（例如位置向量）

- 凡是使用Faster R-CNN等工具提取图像向量的方法，都面临着1000类的限制，无法直接应用到实际业务

图文融合微博业务中的应用场景

打标签

- 标签之间相关性较强。例如：化妆品、美妆爱用物、口红，这3个标签的范围大小不一致，还有包含关系。
- 人工打标签。二级标签850余个，三级标签约50万个。人工判断，标签可能不准确、不全面，标签的使用频率不均衡。
- 当前推荐系统中同时使用标签、关键字作为特征，关键字粒度比标签小，未直接使用图片内容
- 一个应用方法：将图片中物品映射到关键字上，这样能在不改动推荐系统的情况下使用图片内容
- 然后用于打标签、博文聚类或直接用于召回排序

挑战

- Faster R-CNN等目标检测模型常用ImageNet训练，常用物品只1000类，而且这1000类与微博常见物品有差异
- 预训练所用的图像描述是英文，不是中文，构建中文数据集花费巨大

可能的图文融合尝试

- 将图片中的人物对映射名星、大V的名字上
- 按频道构建训练集，每个频道涉及的物品较少
- 寻找无监督的目标检测模型，摆脱对训练集的依赖

图文融合微博业务中的应用场景

向量化特征

- 推荐系统在排序使用的是大规模FM模型，据了解，还没有用上文本向量或者图像向量
- 根据FM的算法原理，FM中也难以使用向量特征

图文相似度

- 将图片、微博分别向量化，用2个向量计算图文相似度
- 论文B2T2中考虑过个任务，仍然依赖于图像特征抽取器，受限于1000类、图像描述训练集缺少中文描述

谢 谢