Report of Project 3

18300290007 加兴华

Abstract

图像字幕 (Image Captioning) 是计算机视觉以及自然语言处理领域比较热门的一项任务，目标为让模型自动生成语言描述给定画面，根据PJ要求，我对其衍生任务，陌生目标的图像字幕（Novel Object Captioning) 进行研究，这项任务要求模型能够一定程度上为训练集中不存在的目标的图像生成字幕。

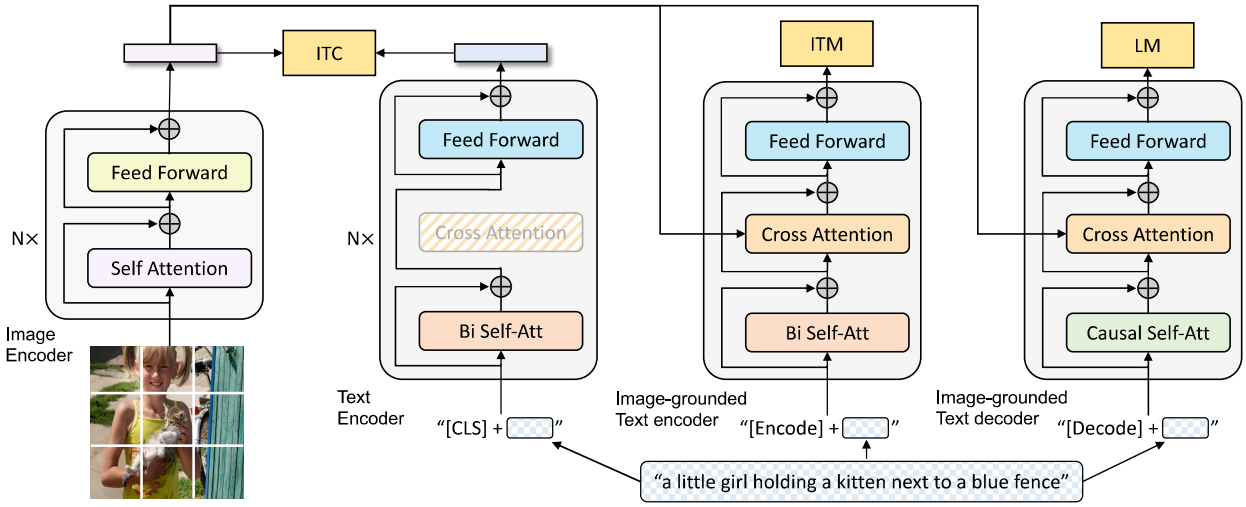
在本项目中，我对近期被提出的Bootstrapping Language-Image Pre-training（BLIP）模型[1]以及经典的Neural Baby Talk 的改进模型 Neural Twins Talk[2] 进行了复现，并在 MS COCO 数据集和 Hendrick 的 Novel划分上对它们进行了实验。

最终，预训练+微调的BLIP模型和从头训练的NTT模型在整个novel测试集上的CIDEr指标分别为126.8(%) 和93.4(%)，更具体的指标详见报告`Experimental Results` 一节。在计算完各类指标后，我借助具体图像的推断表现进一步探索了模型获得当前CIDEr和F1得分的原因。

Models

BLIP

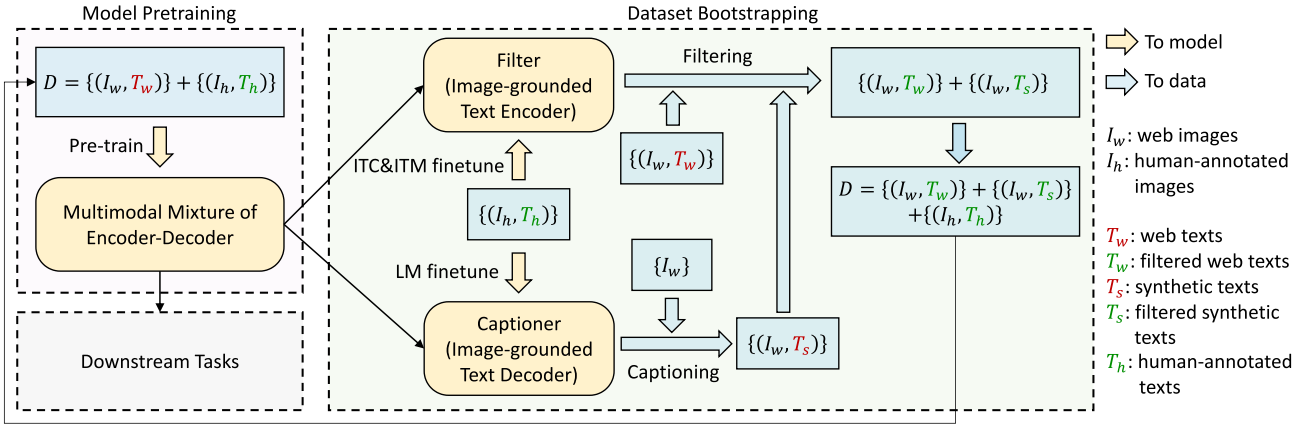
BLIP[1] 的模型框架如下所示：



BLIP提出了多模态混合编码器-解码器（multimodal mixture of encoder-decoder），这是一种统一的视觉语言模型，可以实现以下三种功能之一：

- 单峰编码器通过图像-文本对比（ITC）损失进行训练，以对齐视觉和语言表述。
- 基于图像的文本编码器使用传统的交叉注意层来模拟视觉语言交互，并使用图像-文本匹配（ITM）损失来训练，以区分正负图像-文本对。
- 基于图像的文本解码器将双向自注意层替换为因果自我注意层，并与编码器共享相同的交叉注意层和前馈网络。解码器使用语言建模（LM）损失进行训练，以生成给定图像的字幕。

BLIP的学习框架如下所示：



可以看到预训练和数据增强占比非常之大。BLIP引入一个标题生成器来生成web图像的标题，以及一个过滤器来去除不匹配的图像文本对，它们从相同的预训练模型初始化，并在一个小规模的带注释的数据集上分别进行微调。使用这种方法（CapFilt），能够改善预训练模型所用的数据从而提升预训练效果。

更具体而言，训练过程中同时优化三个目标函数，分别为基于理解的目标 (ITC,ITM) 和基于生成的目标 (LM)。每个图像-文本对只需要一次 ViT 前向计算，以及三次 Text Transformer 前向计算，并使用不同的函数来计算上述的三个损失。

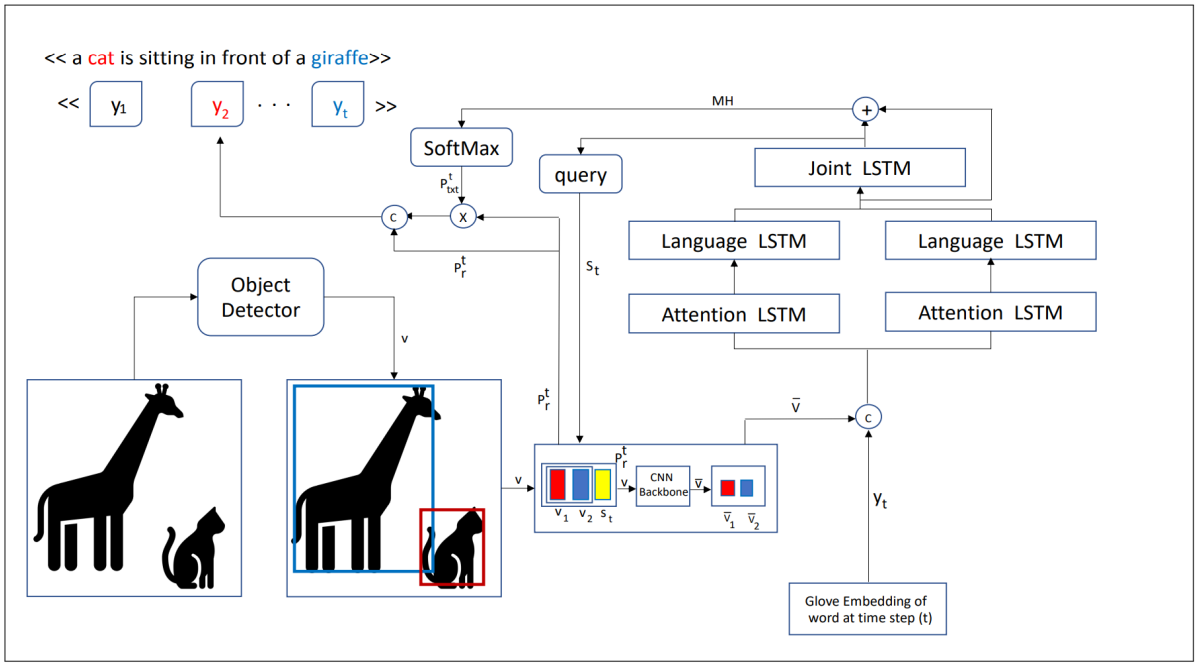
Objective

对于下游任务Image Captioning，数据流只需经过图像编码器与基于图像的文本解码器（Image-grounded Text Decoder），因此任务目标为前文提到的功能(3)：最小化交叉熵语言建模（LM）损失。

在原文中作者实验证明，只需要加载预训练模型，并对这两部分网络在任务数据集上进行微调就能获得相当不错的效果；而由于BLIP致力于使用多模态提升效果，其单一模块不如其他工作复杂和完善，因此如果用BLIP从头训练Image Catption效果不如其他工作。

Neural Twins Talk

NTT [2] 的模型框架如下所示：



在PJ指导文件中引用了Neural Baby Talk(NBT)[3] 作为参考文献，它在生成句子的每个时刻t同时进行视觉词语（visual word）和文本词语（textual word）的生成，在上图中 P_r^t 和 P_{txt}^t 分别表示视觉词语集合和文本词语集合中各元素的概率，最终选择概率最大者嵌入生成的句子，另外如果词语来自视觉，则对其单复数形式和概念粒度进行调整。

NTT实际上是NBT的进一步研究，只修改了NBT的解码器部分（top-down attention），以表明双级联注意模型能够有效地使负责部署LSTM和注意机制的深层网络性能更好。解码器中与NBT的不同在于NTT中的语言LSTM从他们较低层次的注意力LSTM接收他们的假设和上下文向量, 而不是拥有独自的向量；相同之处在于的联合LSTM都从低级语言LSTM接收假设和上下文向量。

Objective

NTT的训练目标与NBT一样为最小化下面的交叉熵损失函数，其中 '*' 表示ground truth：

$$L(\boldsymbol{\theta}) = -\sum_{t=1}^{T} \log \left(\underbrace{p\left(y_t^* \mid \tilde{r}, \boldsymbol{y}_{1:t-1}^*\right) p\left(\tilde{r} \mid \boldsymbol{y}_{1:t-1}^*\right) \mathbb{1}_{(y_t^*=y^{\text{txt}})}}_{\text{Textual word probability}} + \underbrace{p\left(b_t^*, s_t^* \mid \boldsymbol{r}_t, \boldsymbol{y}_{1:t-1}^*\right)}_{\text{Caption refinement}} \underbrace{\left(\frac{1}{m} \sum_{i=1}^{m} p\left(r_t^i \mid \boldsymbol{y}_{1:t-1}^*\right) \right) \mathbb{1}_{(y_t^*=y^{\text{vis}})}}_{\text{Averaged target region probability}} \right)$$

Textual word probability　文本词语概率，根据 $y_{1:t-1}^*$ 来计算出要选择的区域特征 r 的可能性，然后根据这一部分预测出的 $y_t = y_t^*$ 的概率；

Averaged target probability　对象落在各目标区域的平均概率，是视觉词语概率的因子；

Caption refinement　文字描述调整的概率，根据所有特征区域和之前的所有上下文单词，计算t时刻单词单复数形式以及概念粒度各种组合的可能性，同样是视觉词语概率的因子。

Experimental Results

Dataset　我使用了MS COCO [4] 数据集进行模型的训练与实验。COCO2014 包含82783、40504和40775张图像，分别用于培训、验证和测试。每个图像都有大约5个人工标注的标题作为ground truth。

Split　按照PJ要求，我是用Hendirck等人提出的Novel split [5] 对训练集、验证集和测试集进行划分，这种划分选中bottle、bus、couch、microwave、pizza、racket、suitcase、zebra这8个对象作为novel objects，将它们从标准训练集中除外作为新的训练集；将标准验证集切成两半分别作为新的验证集和新的测试集，并且还可以按含novel object与否进一步细分。

Metrics　在实验中我使用了BLEU4 [6] 、Meteor [7] 、CIDEr [8] 、SPICE [9] 、F1分数 [5] 共5种指标对模型进行评价。

> 注：下文中的指标均以百分数的形式展示。

Results of BLIP

BLIP在训练时使用paper作者提供的“BLIP w/ ViT-B and CapFilt-L”作为预训练模型，并在训练集上fine-tuning 1个epcoh，其余的超参数如下所示：

hyperparameters				
ViT	batch size	optimizer	init LR	weigtht decay
base	6	AdamW	5e-6	0.05

训练结束后，BLIP模型对整个测试集以Beam size=3进行预测，结果的评价指标如下：

BLIP Results on COCO2014 and Hendrick's Novel split			
Metircs			
Bleu_4	METEOR	CIDEr	SPICE
38.8	30.3	126.8	23.1

可以看到预训练+fine-tune后的BLIP模型整体性能相当好；

从测试集中提取出只含有单一novel object的8个子集，并分别进行预测的指标评价，结果如下：

	bottle	bus	couch	microwave	pizza	racket	suitcase	zebra
BlEU4	35.5	34.5	**36.3**	32.3	30.2	33.7	33.7	23.5
CIDEr	**113.8**	93.1	94	75.2	81	61.1	96.7	50.9
METEOR	28.4	27.6	29	27.4	25.2	**29.6**	26.7	22.5
SPICE	20.3	**22.5**	21.6	18.9	18.4	22	19.9	15.3
F1	0	**71.2**	6.25	20	16.62	0	0	42.76

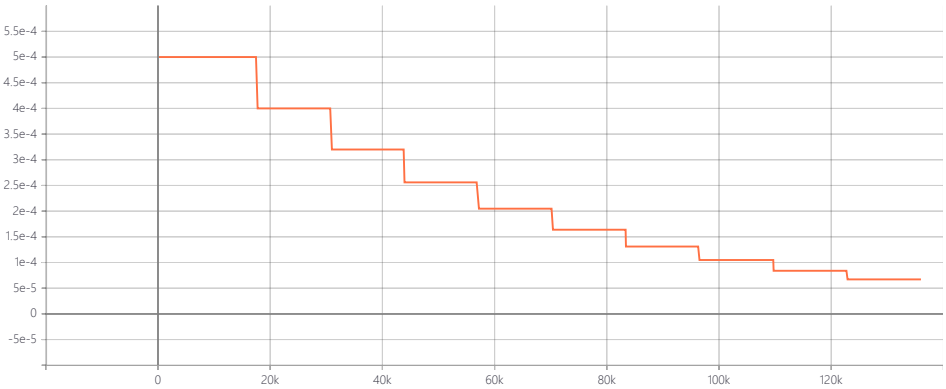
其中，各组的CIDEr和F1指标波动性相当大，其余指标除了少数离群外整体一致。

Results of NTT

NTT在训练时使用在ImageNet上预训练过的ResNet101作为特征提取器（bottom-up attention），ResNet101的RoI head部分随着训练进行fine-tune; 此外，使用GLOVEs预训练的Gloves.6B.300d作为词嵌入；NTT的其余网络部分则从头开始学习。训练的超参数如下表所示：

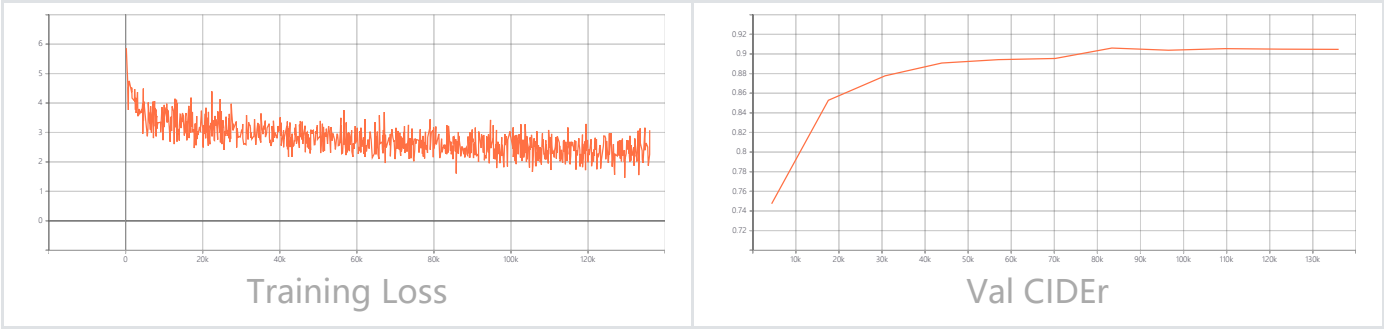
hyperparameters				
batch size	max epoch	optimizer	init LR	optim betas
16	31	Adam	5e-4	(0.8,0.999)

另外，训练中采用step-down的学习率策略，如下图所示，每3个epcoh学习率衰减为原先的0.8。



Learning Rate Schedule in NTT

训练中的损失曲线和验证集上的CIDEr曲线如下图所示：



其中CIDEr指标是对Beam size=1的预测进行评价，会让预测的效果变差但是更加快速。由于GPU和时间上的限制我没有训练更多的轮数，但可以看出曲线的后段都趋于水平，说明模型已经基本收敛。

训练结束后，NTT模型对整个测试集以Beam size=3进行预测，结果的评价指标如下：

NTT Results on COCO2014 and Novel split [Hendricks_2016_CVPR]				
	Metrics			
	Bleu_4	METEOR	CIDEr	SPICE
Mine	30.6	24.9	93.4	18.2
Paper	30.8	-	94.0	18.2

可以看到我的实验结果与paper中的结果基本吻合，各个指标都稍微偏低，这是因为paper中设置的batch size和max epoch都比我更大，从而效果更好；

从测试集中提取出只含有单一novel object的8个子集，并分别进行预测的指标评价，结果如下：

	bottle	**bus**	**couch**	**microwave**	**pizza**	**racket**	**suitcase**	**zebra**
BlEU4	28.9	22.1	33.8	**34.4**	27.5	22.2	23.5	23.1
CIDEr	**83.3**	45.9	68	61.3	52.1	27.8	55.6	36.7
METEOR	22.9	20.8	**25.7**	25	21.8	23.6	20.3	22.9
SPICE	16.1	14.6	**18.3**	16.2	16.4	14.8	13.1	16.6
F1	20.0	63.7	21.3	36.6	41.9	9.8	9.1	**72.8**

与BLIP的实验结果相似，各组的CIDEr和F1指标波动性相当大而其余指标整体一致；而不同在于，NTT的CIDEr和F1波动性更小一些，BLIP甚至有的组别F1指标为0。

Inference Comparison

对比上面两种模型的实验结果，综合上预训练微调的BLIP模型性能会高于从头训练的NTT模型；另一方面，在更细的测试集上，BLIP模型和NTT模型的CIDEr和F1指标都表现出了很大的波动性，甚至体现出了相似的分布，如下表所示。

	bottle	bus	couch	microwave	pizza	racket	suitcase	zebra
CIDEr(NTT)	83.3	45.9	68	61.3	52.1	27.8	55.6	36.7
CIDEr(BLIP)	113.8	93.1	94	75.2	81	61.1	96.7	50.9
F1(NTT)	20.0	63.7	21.3	36.6	41.9	9.8	9.1	72.8
F1(BLIP)	0	71.2	6.25	20	16.62	0	0	42.76

为了进一步明确这两种指标反映的特征及出现上述分布的原因，我进行了一些抽样分析：

F1得分在推断中的体现　两个模型在bottle测试集上的F1得分都很低而zebra测试集上的F1得分都很高，我从两个测试集中各随机抽取了一张图像进行推断，结果如下所示。

Novel Object: bottle

Val ID：495125

[BLIP]
a group of wine glasses sitting on top of a wooden table

[NTT]
a table topped with lots of glasses of wine

Novel Object: zebra

Val ID：100661

[BLIP]
a herd of zebras grazing in a grassy field

[NTT]
a group of zebras running in a field

观察发现，左图中虽然出现了bottle，但只是充当配角，画面的主体是glass；而右图中的主角是zebra。这说明，一方面模型可能对图像的主体更加敏感；另一方面可能是当图像或模型词库中存在概念相似的目标时（bottle和glass），模型更容易把novel object搞混淆。

CIDEr得分在推断中的体现　两个模型在zebra测试集上的CIDEr得分都很低，我从测试集中选取了一张图像进行推断，如下所示。

Val ID：82740

[BLIP]
a man riding a horse over an obstacle

[NTT]
a man riding on the back of a brown horse

[Ground Truth]
A black and white photo of a person riding a horse jumping over obstacles
A person jumping a horse over an obstacle
Zebra themed jumping rails with horse and rider performing
A black and white photo of aperson jumping on a horse
A horse and rider jumping a gate during a ride

CIDEr是一种基于词频-逆文本频率（TF-IDF）的指标，当对一张图像的预测中含有ground truth中出现频率高的词语且该词语出现在其他图像的ground truth中的频率很低时，TF-IDF就高，CIDEr得分正是用于衡量所有图像推断的TF-IDF的整体表现，反之像上面图像的推断中没有“jumping”，TF-IDF会很低，从而拉低CIDEr得分。

直观上来看，CIDEr得分越高，说明模型对于图像有别于其他图像的“个性”更加敏感。两个模型对于不同novel object的CIDEr差异较大，一方面则可能是不同novel object变现个性的能力也不同，比如zebra能够表达出的无外乎“吃跑跳”，而bottle则能够和环境形成多种交互；另一方面，测试集中相似的图片过多也会降低模型的CIDEr得分，这会让不同测试集上CIDEr大小比较说服力减弱，因此CIDEr得分还是更适合用于模型与模型在同样的测试集上进行比较。

Reference

1. **BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation** ↩ ↩
2. **Neural Twins Talk & Alternative Calculations** ↩ ↩
3. **Neural Baby Talk** ↩
4. **Microsoft COCO: Common Objects in Context** ↩
5. **Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data** ↩ ↩
6. **BLEU: a Method for Automatic Evaluation of Machine Translation** ↩
7. **Meteor Universal: Language Specific Translation Evaluation for Any Target Language** ↩
8. **CIDEr: Consensus-based Image Description Evaluation** ↩
9. **SPICE: Semantic Propositional Image Caption Evaluation** ↩