

# Project 3

xw-zeng

2023 年 6 月 14 日

## 1 Introduction

图像字幕生成 (Image Caption) 是计算机视觉领域中的一个重要任务，其主要目标是根据给定的图像内容生成与图像相关的自然语言描述，对于机器理解图像和语言之间的联系具有重要意义。此外，在目标检测任务中，基于已有的大量标注数据，如 Faster R-CNN 等成熟的目标检测技术不断涌现。于是，研究者 Lisa Anne [1] 提出了 Novel Object Captioning 的课题，即利用已有的目标检测信息并结合部分推理，以生成描述新目标的语言。

然而，Lisa Anne 提出的方法本质上并不是零样本学习 (zero-shot learning)。在实际的训练过程中，负责生成描述的 RNN 可以读取包含新单词的句子，而零样本学习是指在没有见过特定类别的样本的情况下，模型仍然能够对该类别进行正确分类或生成描述。此任务应该使用目标识别任务中的标签结果作为推理信息，关键在于语言模型 (LSTM 或 RNN) 中没有关于新物体的嵌入表示 (embedding)。因此，我们需要找到一种有效的方式将新物体的信息引入到模型中，并且使模型能够泛化到未见过的新物体上。

本项目将基于 MS COCO 数据集完成 Novel Object Captioning 的任务，在报告中依次进行文献综述、模型介绍和结果分析。本项目的代码基于 github 上相应模型的官方代码，根据已有环境对部分代码进行修改，以此解决其与 python、pytorch 和 tensorflow 的版本不匹配问题。

## 2 Literature Review

Image Caption 问题也可以看作是一种翻译问题，只不过输入的对象从序列化的 token 变为了图片，本质上没有区别，因此也可以使用 Encoder-Decoder 的框架去解决 (图1)。在 *Show and Tell* 这篇文献中 [2]，作者就借鉴了神经网络机器翻译的思想。Encoder 使用预训练好的 CNN 模型作为特征提取器，将图像转换为特征向量，然后输入到 Decoder 来生成对图像的描述 (如 RNN, LSTM, GRU, Transformer 等)。

Encoder-Decoder 的框架缺点也很明显，即 Encoder 把所有的输入都转换为了统一的特征向量 C，而 C 很容易受到输入句子长度的影响。当句子过长时，受到计算资源的限制，C 可能存不下所有信息，从而导致模型后续的精度大幅下降。

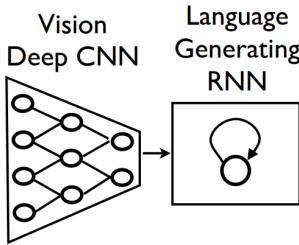


图 1: Encoder-Decoder

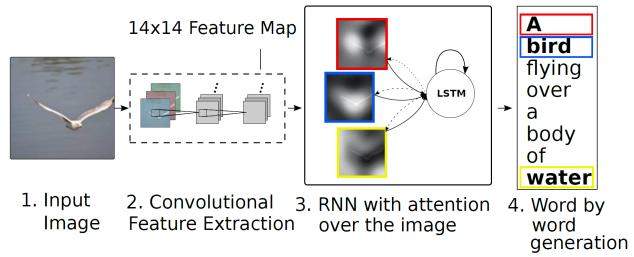


图 2: Attention-Based Encoder-Decoder

因此，有学者在 *Show, Attend and Tell* 中提出可以在原来框架中加入了 Attention 机制 [3]，如图2所示。Attention 机制在自然语言处理领域中应用非常广泛，它模仿了人脑思考问题的方式，会在不同时间注意不同位置的信息，如在翻译时针对性地看原句中的对应内容，而非平等地看待所有词语。实现方式是根据当前输出，对 Encoder 内容进行权重分配。图3展示了 Image Caption 任务中的 Attention 机制：对于 bird 单词，图片中鸟的区域关注程度很高，而对于 water 单词，水的区域关注程度很高。

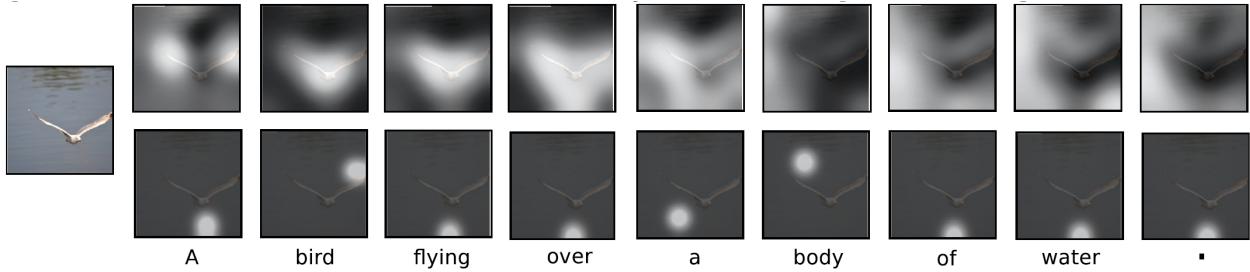


图 3: Image Caption 任务中 Attention 机制示意图

然而在解码时，有些词并不需要从图像中获取视觉信息，这类词被称为非视觉词 (non-visual words)，如 the, of 等虚词。而有些看似需要视觉信息的 words，依赖 Decoder 本身的语义信息就能预测，如 talking on a cell 之后的 phone 这类有意义的组合词。而且，在生成图像描述的过程中，非视觉词的梯度会误导或者降低视觉信息的有效性。因此，*Knowing When to Look* [4] 提出了带有视觉标记的自适应的 Attention 模型，在每一个 time step，模型会决定更关注 Encoder (CNN 产生的 feature map)，还是关注 Decoder (visual sentinel)。

前面的结构都是分 patch 输入，对全局信息的把握程度高，但对局部信息的关注程度不够（图4）。于是，有学者提出了新的 Bottom-Up and Top-Down Attention 结构 [5]。这是一种自下而上和自上而下的组合注意机制，在 Faster R-CNN 的基础上，使用 ROI-Align 方法，对于图片中的关键部分进行标注，以此得到较好的图片语义信息。这篇文献的作者还提出了双层 LSTM 结构（图5），上层的 LSTM 用于语言模型生成 (Bottom-Up 机制)，下层的 LSTM 用于图片的特征选择 (Top-Down 机制)。这种 Attention 结构充分利用了前面句子生成的信息，生成的图片描述更为准确。

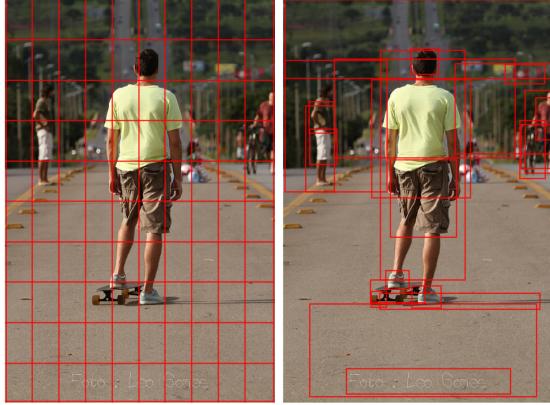


Figure 1. Typically, attention models operate on CNN features corresponding to a uniform grid of equally-sized image regions (left). Our approach enables attention to be calculated at the level of objects and other salient image regions (right).

图 4: Attention Structure

### 3 Models

本项目使用的模型有: Neural Baby Talk (NBT), Neural Twins Talk (NTT) 和 Bootstrapping Language-Image Pre-training (BLIP)。

#### 3.1 Neural Baby Talk

Neural Baby Talk [6] 的方法像是 Adaptive Attention 和 Bottom-Up and Top-Down Attention 的结合。在模型生成时, 利用 Bottom-Up Attention 机制, 结合句法结构和语义信息生成下一个位置可能出现的单词; 同时也借鉴了 Adaptive Attention 的做法, 生成一个隐变量  $s_t$ , 决定当前时刻应该采用 textual word 还是 visual word, 然后对 visual words 的概率进行缩放; 最后从极大似然的角度选取最可能出现的单词。

最大化似然函数:

$$\theta^* = \arg \max_{\theta} \sum_{(I,y)} \log p(y | I; \theta) \quad (1)$$

$$p(y | I) = \prod_{t=1}^T p(y_t | y_{1:t-1}, I) \quad (2)$$

$$p(y_t | y_{1:t-1}, I) = p(y_t | r_t, y_{1:t-1}, I) p(r_t | y_{1:t-1}, I) \quad (3)$$

其中  $y_t$  为生成的单词, 其类型由  $s_t$  进行控制。如果选取了 visual word, 那么还需要对词语进行变换使其适合当前文本上下文, 如单复数、时态等。作者总共考虑了两种变换, 分别为单复数 (如 dog 跟 dogs) 和类别的 fine-grained 标签 (如 dog 可以细分为 puppy 等), 单复数用二分类器, fine-grained 使用多分类器。

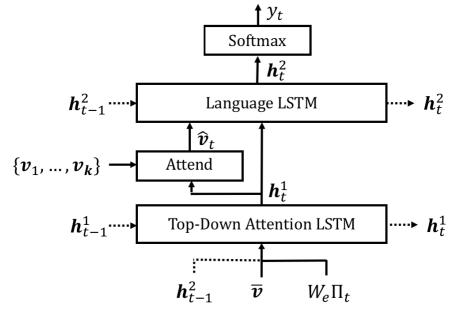


Figure 3. Overview of the proposed captioning model. Two LSTM layers are used to selectively attend to spatial image features  $\{v_1, \dots, v_k\}$ . These features can be defined as the spatial output of a CNN, or following our approach, generated using bottom-up attention.

图 5: Two LSTM Layers

训练目标为最小化交叉熵损失函数，该损失函数分为两部分：如果当前词语的 target 是 textual word，则用前半部分；若是 visual word，则用后半部分。要求一是 region（或者说 visual word）选取正确，二是词语的单复数与细分类别形态正确。

$$\begin{aligned}
 L(\theta) = & -\sum_{t=1}^T \log \underbrace{\overbrace{p(y_t^* | \tilde{r}, y_{1:t-1}^*) p(\tilde{r} | y_{1:t-1}^*) \mathbb{1}_{(y_t^* = y^{\text{txt}})}}^{\text{Textual word probability}}} \\
 & + \underbrace{p(b_t^*, s_t^* | r_t, y_{1:t-1}^*) (\frac{1}{m} \sum_{i=1}^m p(r_t^i | y_{1:t-1}^*)) \mathbb{1}_{(y_t^* = y^{\text{vis}})}}_{\text{Caption refinement}} \underbrace{\mathbb{1}_{(y_t^* = y^{\text{vis}})}}_{\text{Averaged target region probability}}
 \end{aligned} \quad (4)$$

文本词汇的产生来自于 LSTM+Attention 的语言模型结构，visual word 的产生来自于 Faster R-CNN 的结果（这部分与 Bottom-Up 机制相同）。然后，模型结合了 visual word 和之前语言模型信息  $h_t$ ，利用 LSTM 生成了新的  $s_t$  状态。 $s_t$  状态将会和图形词汇的分布一起决定最后的  $r_t$ 。 $r_t$  决定了最后选择从语言规则角度选择的单词，而 visual word 决定了这个地方使用 object detector 给出的单词，如图6所示。

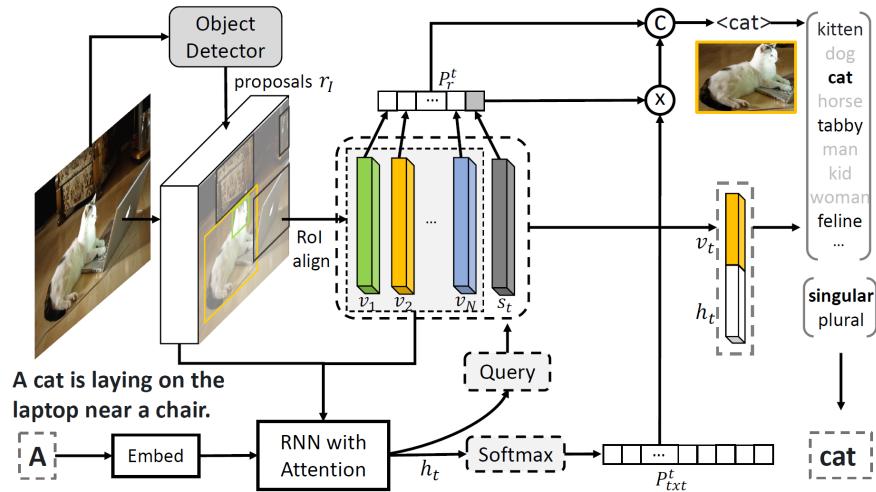


图 6: NBT 框架图

### 3.2 Neural Twins Talk

Neural Twins Talk [7] 是 NBT 的改进模型，不修改 Encoder (Bottom-Up Attention)，只修改了 Decoder 的 Top-Down Attention，其他与 NBT 完全相同，以表明双级联注意模型可以有效地使负责部署 LSTM 和注意机制的深层网络性能更好。NTT 中使用了 2 个注意力通道，这使得 joint LSTM and language LSTM 能够以集成的方式来细化 slots，如图7所示。

NTT 与 NBT 的主要不同之处在于 NTT 中的 language LSTM 从它们较低层次的 attention LSTM 接收假设和上下文向量，而不是通过自己拥有的向量；相同之处在于的 joint LSTM 都从低级语言 LSTM 接收假设和上下文向量。

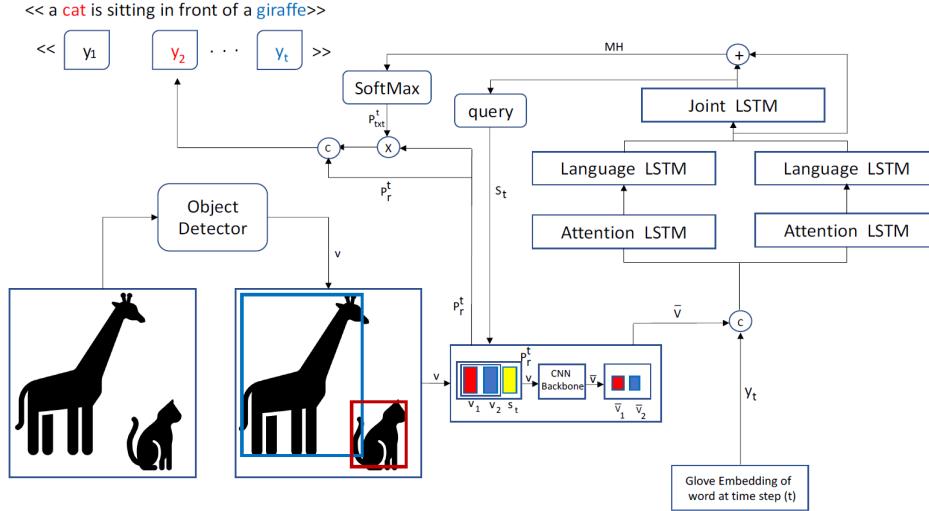


图 7: NTT 框架图

### 3.3 Bootstrapping Language-Image Pre-training

Bootstrapping Language-Image Pre-training [8] 是一个新的 VLP (Vision-language pre-training) 框架，与现有方法相比，它可以实现更广泛的下游任务。

为了预训练一个既有理解能力又有生成能力的统一模型，作者提出了多模态混合编码器-解码器 Multimodal mixture of Encoder-Decoder (MED) 框架，这是一个多任务模型，可以作为单模态编码器、基于图像的文本编码器或基于图像的文本解码器工作（图8）。

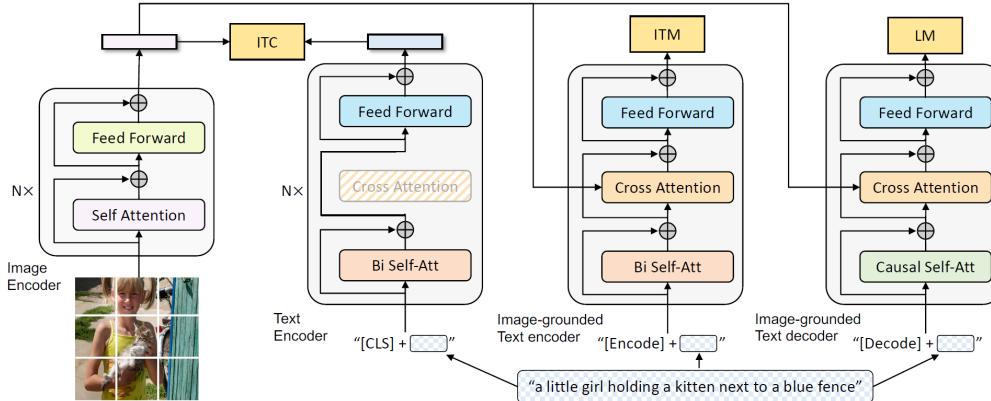


图 8: 预训练框架图

该模型与以下三个视觉语言目标联合进行预训练：

1. Image-Text Contrastive (ITC): 图像-文本对比任务，目的是通过促进正向的图像-文本对与负向的图像-文本对有相似表示，以此对齐 ViT 和 text Transformer 的特征空间。
2. Image-Text Matching (ITM): 图像-文本匹配任务，这是一个二元分类任务，模型根据多模态

特征使用一个 ITM 头（线性层）来预测一个图像-文本对是否匹配。

3. Language Modeling Loss (LM): 语言建模损失，研究人员采用 0.1 的标签平滑度 (label smoothing) 来计算损失。与其他用于 VLP 的 MLM 损失相比，LM 损失使模型具有泛化能力，能够将视觉信息转换为连贯的标题。

作者还引入了 Captioning and Filtering (CapFilt) 作为一种新的数据集增强方法，用于从噪声图像-文本对中学习。作者将预先训练的 MED 分为两个模块，字幕器和过滤器。前者用于生成给定 web 图像的合成字幕，后者用于从原始 web 文本和合成文本中删除嘈杂的字幕。字幕器和过滤器都是从相同的预训练 MED 模型初始化的，并在 MS COCO 数据集上分别进行微调，训练流程如图9。

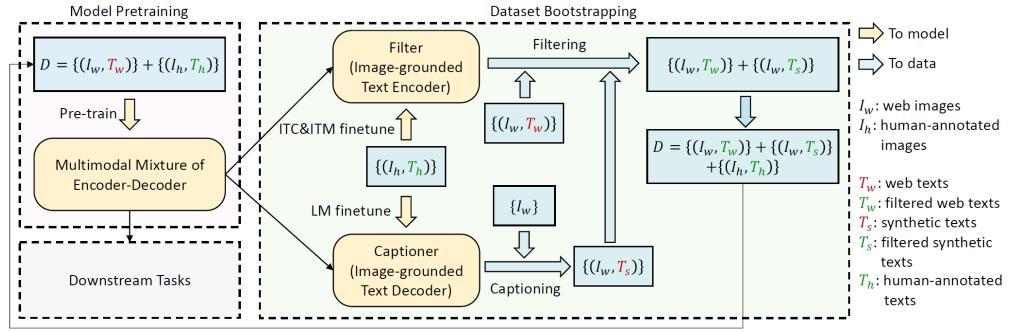


图 9: BLIP 学习框架图

## 4 Experiments

### 4.1 Dataset

本项目使用 MS COCO 2014 数据集 [9] 进行模型训练。训练集、验证集、测试集分别包含了 82783、40504、40775 个图像，每个图像都有大约 5 个人工 caption 作为 inference (参考答案)。

与 NBT 的数据预处理方式相同，本项目按照指导文件要求，以 8 个 novel object 对数据集进行划分：bus, bottle, couch, microwave, pizza, racket, suitcase, zebra，并将它们从数据集中抽取出来。然后我们对单词的词形进行还原（如单复数、时态等），然后对单词进行 embedding，得到 embedding 向量。这一部分借助 StandfordNLP 和 torchtext 包来实现。

### 4.2 Metrics

本项目使用 BLEU-4, METEOR, CIDEr-D, SPICE, F1-Score 作为模型的评价指标，以 CIDEr-D 作为指标保存最优模型参数。BLEU-4, METEOR, CIDEr-D, SPICE 这 4 个指标都直接调用了 pycocotools 中内置的函数进行计算。F1-Score 的计算公式为：

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = \frac{\text{TP}}{\text{TP} + 1/2(\text{FP} + \text{FN})} \quad (5)$$

在第5节中，所有指标都将以百分数的形式展示。

### 4.3 Code

代码详见 code 文件夹：NBT (TopDownModel) [10], NTT (NewTopDownModel) [11], BLIP [12]。

### 4.4 Training

NBT 和 NTT 在训练时都使用在 ImageNet 上预训练过的 ResNet101 作为特征提取器 (Bottom-Up Attention)，进行了一层 Fine Tune，以更好地配合下游任务；使用 GLOVEs 预训练的 Gloves.6B.300d 作为词 embedding；其余部分则从头开始训练。这 2 个模型的超参数都相同，如表1所示。

表 1: NBT&NTT 模型超参数

Optimizer	Initial LR	LR Decay	Epoch	Batch Size	Weight Decay
Adam	5e-4	0.8	15	16	0

训练中学习率每 2 个 epoch 衰减 0.8，NTT 模型的损失曲线和学习率曲线如图10、图11所示。

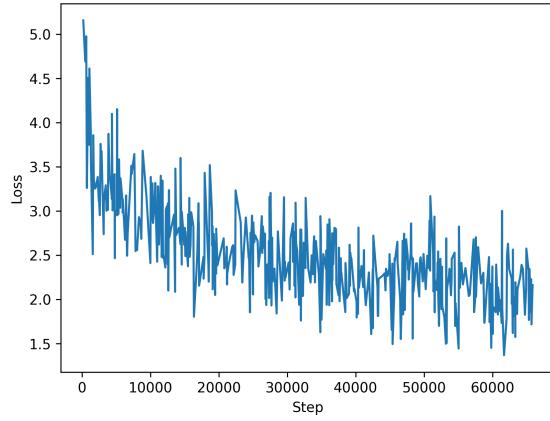


图 10: NTT 损失曲线

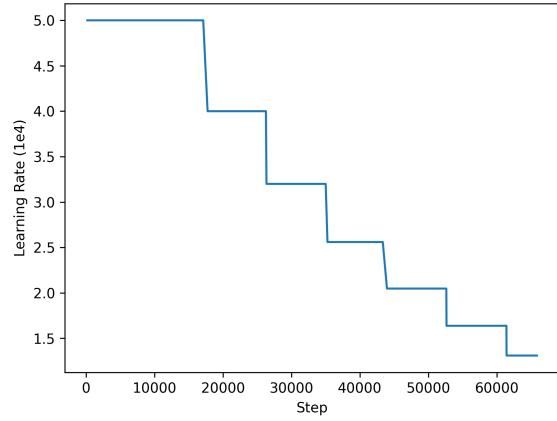


图 11: NTT 学习率曲线

BLIP 模型在训练时使用作者提供的 *BLIP w/ ViT-B and CapFilt-L* 作为预训练模型 [13]。原作者使用的 batch size 为 32，初始学习率为 1e-5，受到算力限制，我将 batch size 减少至 8，初始学习率降低为 5e-6。模型参数如表2所示。

表 2: BLIP 模型超参数

ViT	Optimizer	Initial LR	Epoch	Batch Size	Weight Decay
base	AdamW	5e-6	30	8	5e-2

## 5 Results

### 5.1 Neural Baby Talk

由于 NTT 是基于 NBT 进行改进的模型，训练效果更优，因此这里就简单放一下 NBT 模型的相关评价指标（表3、表4），在 NTT 的结果部分再进行仔细分析。

表 3: NBT 训练结果 (beam size=3)

	BLEU-4	METEOR	CIDEr-D	SPICE	F1-Score
This Project	29.62	24.05	92.82	18.41	86.56
Paper	30.79	-	93.83	18.17	-

表 4: NBT Novel Objects 训练结果 (beam size=3)

	bus	bottle	couch	microwave	pizza	racket	suitcase	zebra	average
BLEU-4	23.49	27.37	<b>35.37</b>	34.53	26.19	25.28	24.14	30.47	28.36
METEOR	22.05	22.92	<b>26.81</b>	25.51	23.06	26.69	20.96	25.51	24.19
CIDEr-D	45.74	<b>77.74</b>	66.48	48.57	44.95	32.28	57.49	48.08	52.66
SPICE	17.14	18.82	20.86	16.79	17.50	<b>21.65</b>	15.20	18.14	18.26
F1-Score	89.07	63.35	76.58	90.98	93.02	<b>96.91</b>	86.25	96.32	86.56

### 5.2 Neural Twins Talk

NTT 模型结果见表5、表6，可以看到模型结果相比于 NBT 模型有所提升。

表 5: NTT 训练结果 (beam size=3)

	BLEU-4	METEOR	CIDEr-D	SPICE	F1-Score
This Project	30.49	24.35	93.75	18.53	89.91
Paper	30.82	-	94.01	18.26	-

#### 5.2.1 Constrained Beam Search

Constrained Beam Search (CBS) 在使模型获得泛化能力方面有着极为重要的作用 [14]。Beam Search 的思想是每一步求解选择 k 个最优片段，每次基于上一步的选择范围进行剪枝，在最后一步选择概率最大的文本。Constrained Beam Search 则在此基础上增加了一个约束的单词集合，能够对生成文本施加控制。

该方法建立了一个有限状态机 (Finite-state machine)，每次向候选集合中加入一个约束单词，然后

表 6: NTT Novel Objects 训练结果 (beam size=3)

	bus	bottle	couch	microwave	pizza	racket	suitcase	zebra	average
BLEU-4	26.94	28.56	<b>35.68</b>	31.06	27.35	24.53	25.32	29.70	28.64
METEOR	23.27	23.48	<b>26.98</b>	24.55	22.96	26.87	21.50	24.95	24.32
CIDEr-D	53.31	<b>77.00</b>	69.80	41.41	45.29	30.21	62.17	47.65	53.35
SPICE	18.55	18.74	21.27	16.27	17.38	<b>22.11</b>	15.18	16.19	18.21
F1-Score	91.75	64.18	90.64	90.32	94.98	<b>99.43</b>	91.06	96.89	89.91

排列组合出最有可能的  $b$  条序列（长度为原本句子中单词数量 +1），直到句子结束后发现有满足约束条件的句子，就返回这些句子中损失函数最小的文本。这样能使得生成的 caption，既符合合适的语法结构，又包含所需的 tag。CBS 具体结构如图12所示。

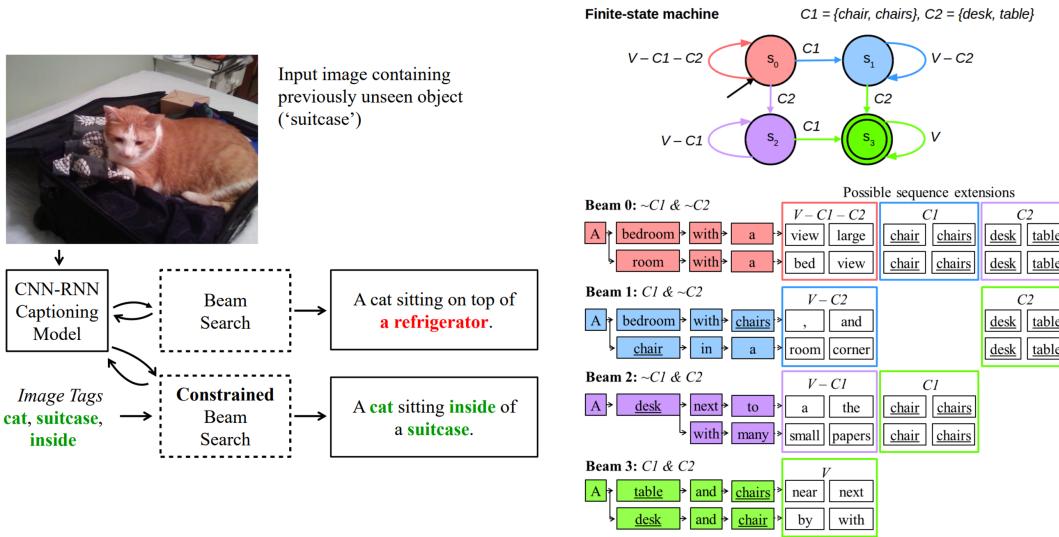


图 12: CBS 结构示意图

根据经验法则，beam size（也就是 tag 数量）为 3 时，搜索效率和结果较好。表7展示了是否使用 CBS 对模型结果带来的影响。

表 7: NTT with or without CBS (beam size=3)

	bus	bottle	couch	microwave	pizza	racket	suitcase	zebra	average
No CBS	63.94	3.14	49.05	49.38	28.53	13.19	52.51	92.31	44.01
With CBS	91.75	64.18	90.64	90.32	94.98	99.43	91.06	96.89	89.91

结果显示，在使用 CBS 以后，所有指标都显著提高，这是由 CBS 的选择能力导致的。因此，计算所有模型的评价指标时，本项目都使用 CBS (beam size=3) 预测得到的结果。

### 5.2.2 Sentences in Each Epoch

研究随着训练的 epoch 数增多, caption 的变化情况 (无 CBS)。选取训练集 dataloader 的前 3 张图片作为例子 (图13) 进行分析。

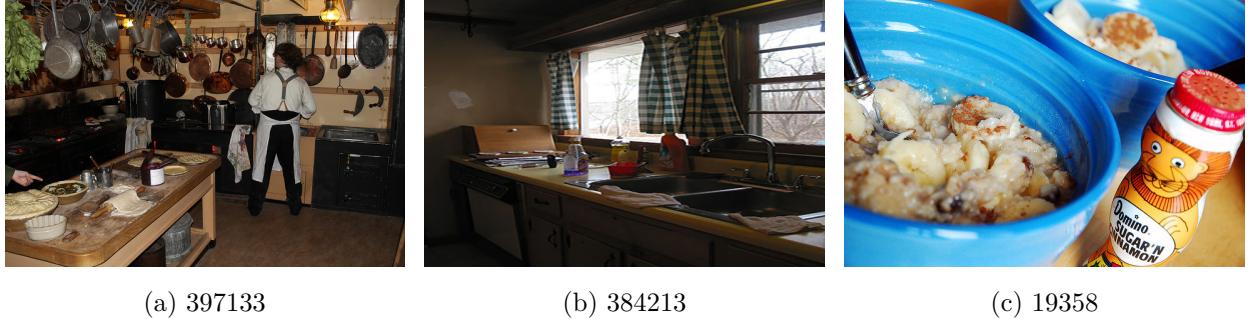


图 13: 示例图片

结果如表8所示。随着模型训练轮数的增多, 模型识别到的物品也增多, 如 397133 从 kitchen 到 kitchen 和 food, 384213 从 sink 到 sink 和 window, 19358 从 food 到 food 和 bootle。

表 8: Caption 示例

Image ID	Epoch	Caption
397133	0	a man is standing in front of a kitchen
	3	a man standing in front of a kitchen counter
	9	a man standing in a kitchen preparing food
	CBS	a man standing in a kitchen preparing food
384213	0	a kitchen with a sink and a sink
	3	a kitchen with a sink and a window
	CBS	a kitchen with a sink a bottle and a window
19358	0	a bowl of food that is on a table
	3	a bowl of food with a bottle of coffee
	6	a blue bowl filled with food on a table
	12	a bowl of cereal and a bottle of coffee
	CBS	a bowl of food with a bottle of coffee

大部分的句子在 10 个 epoch 后就保持不变了, 相应的 CIDEr-D 指标也基本保持不变, 然而即使模型收敛, 有些图片中的新物体也没有被发现。例如, 在 397133 图片中, 厨房桌面上有许多 pizza, 但是生成的 caption 里并不包括 pizza; 在 384213 图片中, 没有 CBS 的情况下 caption 里也没有包括新物品 bottle。这可能是因为模型从图像中识别到的物品非常多 (厨房中的物品本来就比较杂乱且多), 在没有给定约束的情况下, 模型很难选择出符合要求的单词。

CBS 在给定约束之后，能有更大概率选择出我们想要识别出的物体，如 384213 图片就通过 CBS 找到了 bottle。然而，CBS 也不能完全解决这种问题，如 397133 图片中即使给定 pizza 约束也没能选择 pizza，目标检测技术也没能识别出桌面上的是 pizza。这一方面可能是因为 man 和 pizza 的相关关系其实不是很大，二者的距离也比较远；另一方面桌面上的还没做好的 pizza 看上去就是一张面皮，确实挺难识别出来（我自己肉眼看图也没看出来这是 pizza），所以不对 food 做细化而是直接用 preparing food 是合理的。

### 5.3 Bootstrapping Language-Image Pre-training

BLIP 模型结果见表9、表10。

表 9: BLIP 训练结果 (beam size=3)

	BLEU-4	METEOR	CIDEr-D	SPICE	F1-Score
	38.54	29.66	125.72	22.90	91.60

表 10: BLIP Novel Objects 训练结果 (beam size=3)

	bus	bottle	couch	microwave	pizza	racket	suitcase	zebra	average
BLEU-4	28.22	29.68	<b>36.89</b>	33.71	28.52	25.81	26.64	31.08	30.07
METEOR	24.59	24.60	<b>28.22</b>	25.69	24.26	27.94	22.87	26.27	25.56
CIDEr-D	59.76	<b>83.46</b>	75.89	48.12	51.60	36.36	68.22	54.12	59.69
SPICE	19.56	20.05	22.30	17.48	18.83	<b>23.14</b>	16.26	18.31	19.49
F1-Score	92.81	68.43	92.85	92.56	96.10	<b>99.77</b>	92.11	98.16	91.60

#### 5.3.1 Metrics

比较所有的模型的结果，可以看出基本所有的指标都是 BLIP 模型最好。这是因为 BLIP 的作者在实验中表示过，加载预训练模型后对网络进行微调之后就能达到很不错的效果 [8]；而且 BLIP 模型的重点在于用 MED 来提升效果，单模块的部分并不完善，如果像 NBT 和 NTT 一样从头开始训练，最终效果其实是不如 NBT 和 NTT 的。

另外，从 Novel Objects 的结果中看，发现在所有模型中指标最大的都是同一个新物体：couch 的 BLEU-4, METEOR; bottle 的 CIDEr-D; racket 的 SPICE 和 F1-Score。

**1. BLEU-4:** 该指标衡量的是 4-gram 的精确率，即生成的 caption 中与 reference 相同的 4-gram 个数与 reference 总 4-gram 个数的比值。couch 的 BLEU-4 最大可能是因为 couch 总是以 on a couch 等与介词组合的形式出现，更多依赖于语义信息而非图片信息，故无论图片什么样，只要识别出 couch，BLEU-4 指标就会大一些。

**2. METEOR:** 该指标衡量的是生成 caption 与 inference 的匹配程度, 考虑了单词、词干、同义词和其他语言变体的匹配精确率。该指标在各类新物体中的方差并不大, 就不探索 couch 的 METEOR 较高的原因了。

**3. CIDEr-D:** CIDEr-D 比较的是生成 caption 与 inference 之间的相似程度。racket 在该指标的得分非常低, 挑选几张对应类别的图片查看情况, 发现 racket 图像生成的 caption 多以 racket 为主语, 即使画面中有人也没有体现在字幕中 (图14、图15)。



[NTT] a tennis **racket** that is on a metal pole  
[BLIP] a tennis **racket** hanging from a metal fence

[Inference]

1. a **man** with a tennis **racket** is on a tennis court
2. a **woman** hitting a tennis ball with her tennis **racket**
3. a **person** with a **racket** jumps to hit a ball
4. a **person** on a tennis court hitting a ball with a **racket**

图 14: 416267 (racket)



[NTT] a tennis **racket** is being played on a field  
[BLIP] a tennis **player** is ready to hit the tennis **racket**

[Inference]

1. a tennis **player** with his leg up in the air and a **racket** in one hand
2. a **man** with his leg up looking in the air
3. **photographers** taking pictures of a tennis **player** during a game

图 15: 121744 (racket)

我本来猜测这是因为加入 CBS 约束后, 选择 racket 的概率提升, 造成了一定程度的主次颠倒。于是我又查看了 BLIP 模型的 CIDEr-D 指标在 CBS 前后是否有变化, 发现在采用 CBS 时所有新物体的 CIDEr-D 都降低了 (表11), 这说明 racket 得分低有着其他原因。

表 11: BLIP CIDEr-D 指标比较

Method	bus	bottle	couch	microwave	pizza	racket	suitcase	zebra	average
No CBS	60.19	88.79	80.21	65.81	53.17	39.06	71.14	54.84	64.03
With CBS	59.76	83.46	75.89	48.12	51.60	36.36	68.22	54.12	59.69

我又仔细查看了一下 CIDEr-D 的计算方式, 发现它是基于 TF-IDF 构建的指标, 即它既考虑生成的 caption 与 inference 重合的词语个数, 又考虑这些词语在所有图像的 inference 中出现的频率。前者越高, 后者越低, 则 CIDEr-D 越高。因此在 Image Caption 任务中, CIDEr-D 衡量的不仅仅是相似程度, 还包含了独特程度的信息。这就能解释为什么 racket, microwave 的 CIDEr-D 特别低

而 bottle 的 CIDEr-D 特别高了：racket, microwave 一般分别出现在网球场上和厨房里，而 bottle 不但能出现在这些地方，还可能出现在其他多种多样的场景下。CBS 能降低 CIDEr-D 也是因为加入约束会降低 caption 的独特性。

**4. SPICE**: 该指标相比于 BLEU 指标考虑了更为详细和复杂的语义信息，通过将 caption 转换为 scene graph，再用谓词逻辑 (predicate logic) 将其转换为元组计算相似度。该指标在各类新物体中的方差也不大，就不探索 racket 的 SPICE 较高的原因了。

**5. F1-Score**: 该指标在4.2节已经介绍过，对于新目标对象与数据集中对象比较相似的类别，效果都比较好，比如 bus 对应 car, pizza 对应 food, zebra 对应 horse 等；对于总是与 tennis 一起出现的 racket (且词库中没有与之相近的词) 也不难生成。然而，bottle 存在许多相近概念的物体但实物相差较大（如 bottle 和 glass, cup 等），模型容易将其混淆，F1-Score 就比较低（图）。好在 CBS 能够帮助 caption 成功选择对应的单词。

### 5.3.2 Object Detection

结合图片，进一步研究 Image Caption 与 Object Detection 之间的关系。

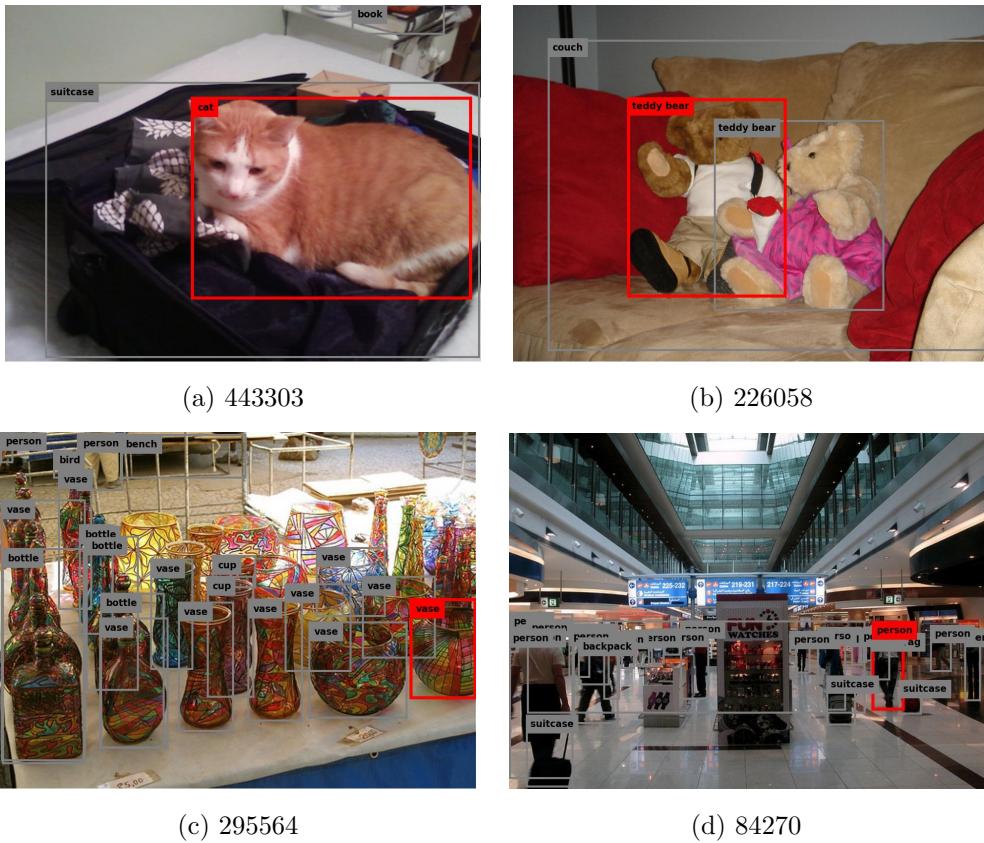


图 16: 目标检测示例图片

图16a、图16b的结构比较清晰，目标检测技术能清晰地识别出重要物体，从而模型结果也比较好。CBS 的结果相比于没有 CBS 的结果更好，这是因为 CBS 提供了额外的有效信息。

后两张图片的结构比较复杂，含有许多物体，目标检测检测出了许多对象，caption 时不知道选择什么作为主体。此时 Attention 机制就会导致模型偏向于生成训练集中出现过的对象，从而忽略新物体（如图16c的 vase 与 bottle、图16d的 suitcase）。具体的 caption 见表12。

表 12: Caption 示例

Image ID	Method	Caption
443303	No CBS	a <b>cat</b> laying on top of a <b>bag</b> on a <b>table</b>
	With CBS	a <b>cat</b> laying on top of a <b>suitcase</b> next to a <b>book</b>
226058	No/With CBS	two <b>teddy bears</b> sitting on a <b>couch</b> in a bedroom
295564	No CBS	a bunch of different types of <b>vase</b> on display
	With CBS	a bunch of <b>bottles</b> that are on a table
84270	No CBS	a group of <b>people</b> standing around a subway station
	With CBS	a group of <b>people</b> standing around <b>suitcases</b>

当然，结构复杂的图片也不一定就效果不好，模型很擅长通过使用 a set of, a group of 等表示量多的词汇对图片中识别出的相似物体进行概括，如 295564 中的 a bunch of bottles 和 84270 中的 a group of people。不过前提是目标检测必须要能识别出相关物体，否则就灾难了（如图17中将 pizza 识别为了许多 sports ball 和 tennis racket）。

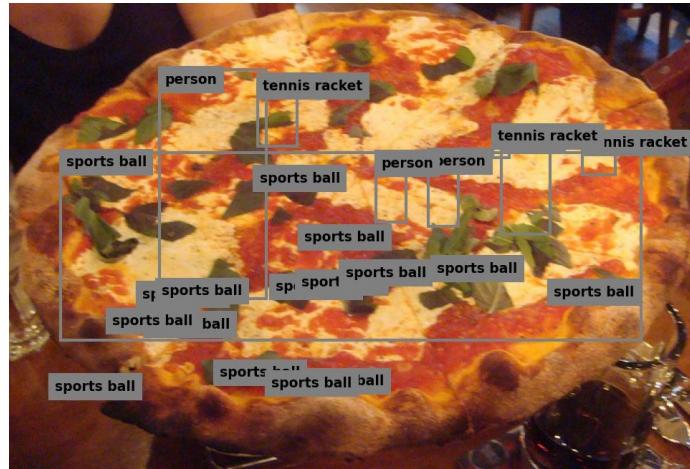


图 17: 397 (pizza)

## 参考文献

- [1] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [2] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator, 2015.
- [3] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2016.
- [4] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning, 2017.
- [5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering, 2018.
- [6] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk, 2018.
- [7] Zanyar Zohourianshahzadi and Jugal K. Kalita. Neural twins talk and alternative calculations. *International Journal of Semantic Computing*, 2021.
- [8] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.
- [9] <https://cocodataset.org/>.
- [10] <https://github.com/jiasenlu/NeuralBabyTalk>.
- [11] <https://github.com/zanyarz/NeuralTwinsTalk>.
- [12] <https://github.com/salesforce/BLIP>.
- [13] [https://storage.googleapis.com/sfr-vision-language-research/BLIP/models/model\\_base\\_capfilt\\_large.pth](https://storage.googleapis.com/sfr-vision-language-research/BLIP/models/model_base_capfilt_large.pth).
- [14] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search, 2017.