

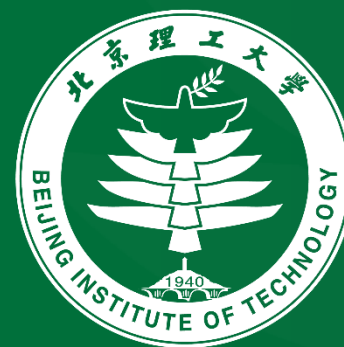


BIT



图像Caption-看图说话

指导教师：张华平 汇报人：蔡永奇、陈盈盈、郭镇远
马昊哲、史凯阅、张津铭



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

德以明理 学以精工



目录

- 一、ImageCaption 简介
- 二、Visual-Encoding 视觉编码
- 三、Language-Model 语言模型
- 四、Dataset 数据集和评级指标
- 五、Demo展示

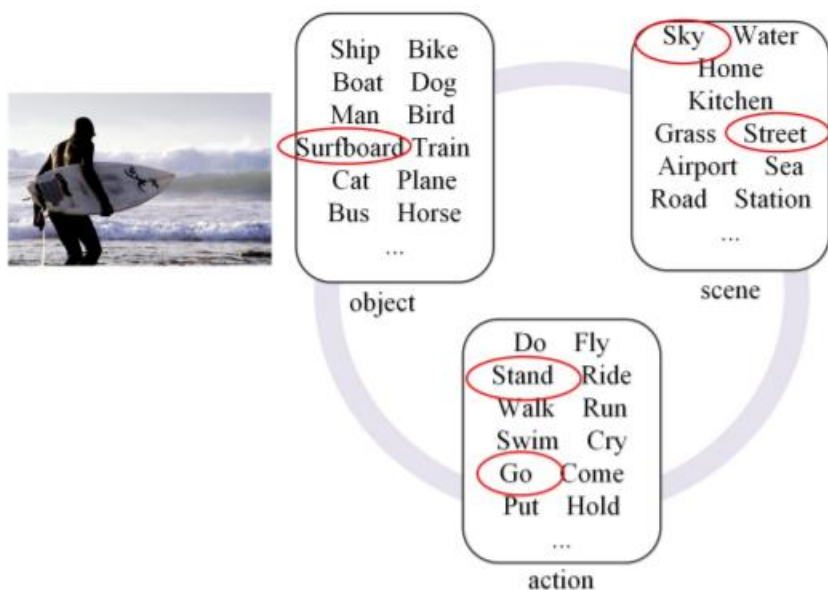


ImageCaption介绍

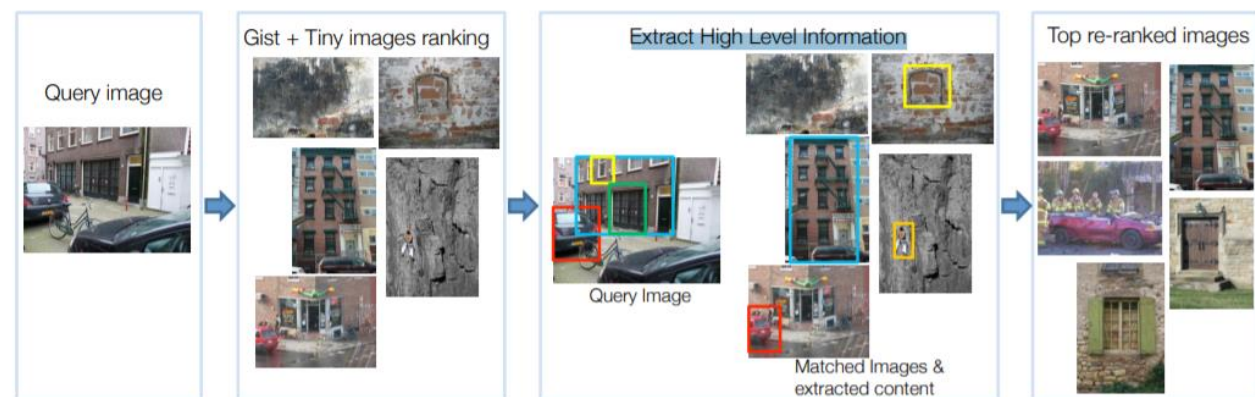


- 一个小女孩和她妈妈在屋子里做手工，桌子上有剪刀，书本贴纸瓶子。

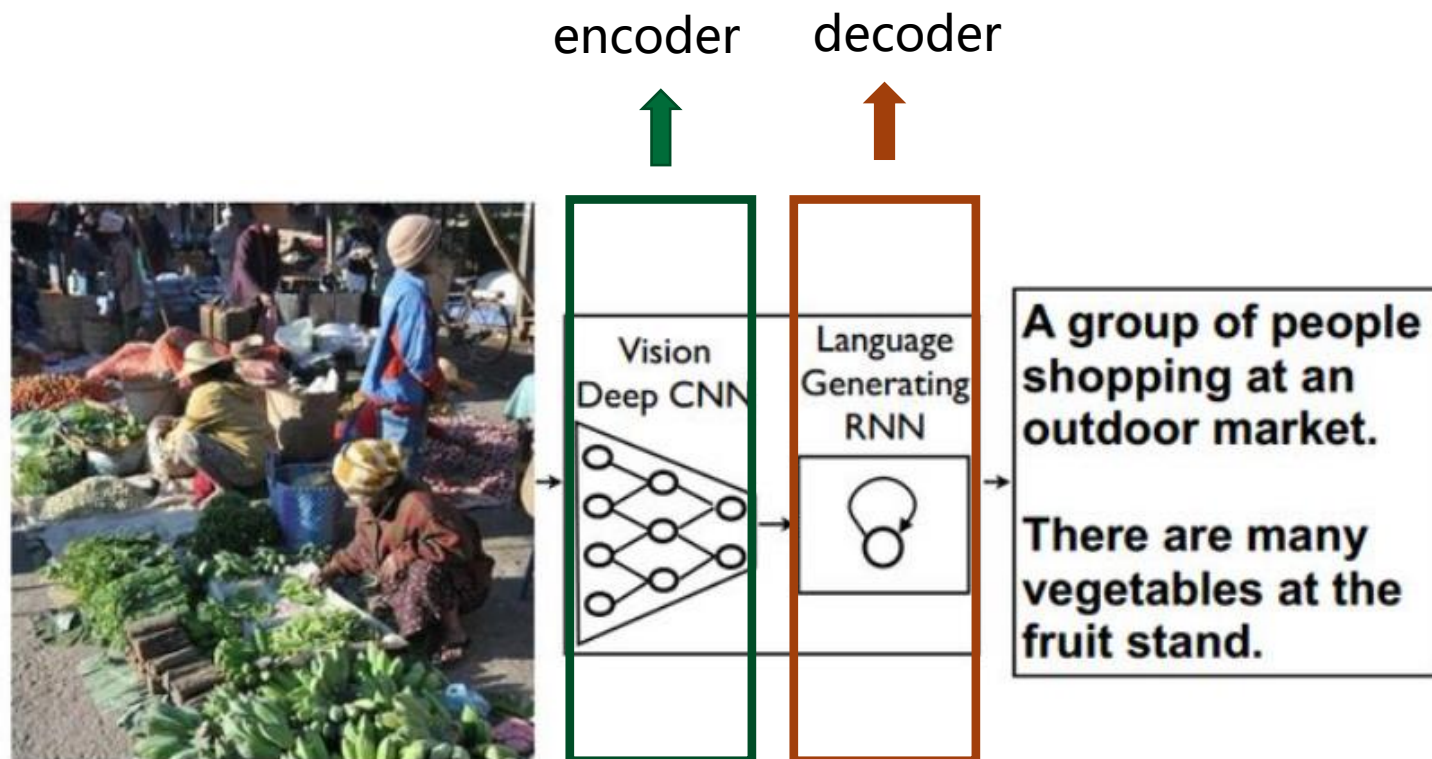
基于模板的图像描述方法



基于检索的图像描述方法

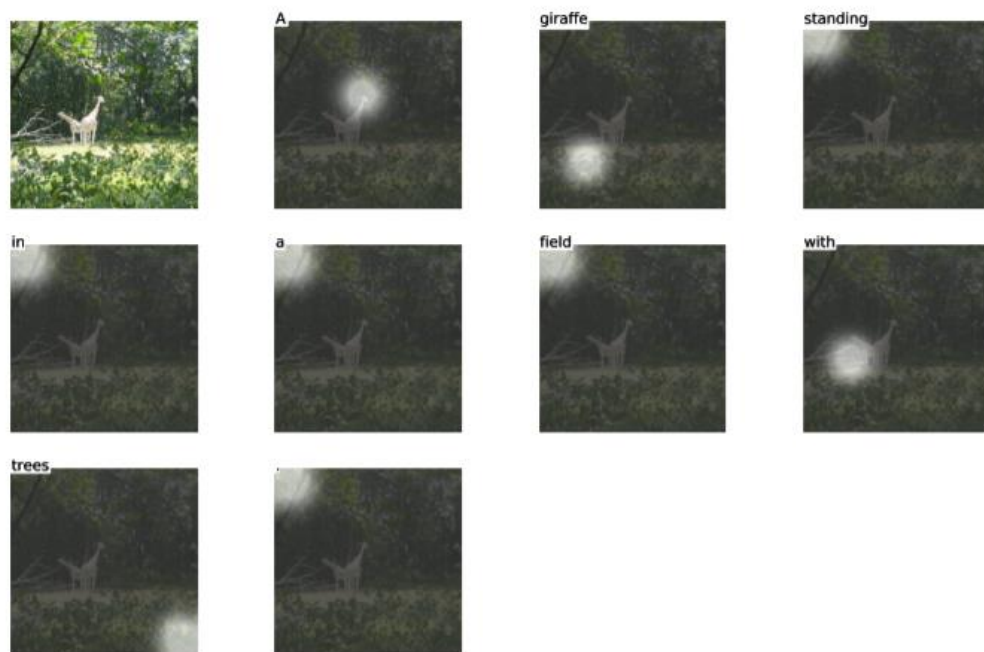


- 首次将深度学习引入image caption任务，提出了encoder-decoder框架，之后的大部分image caption任务都是基于这个框架进行优化



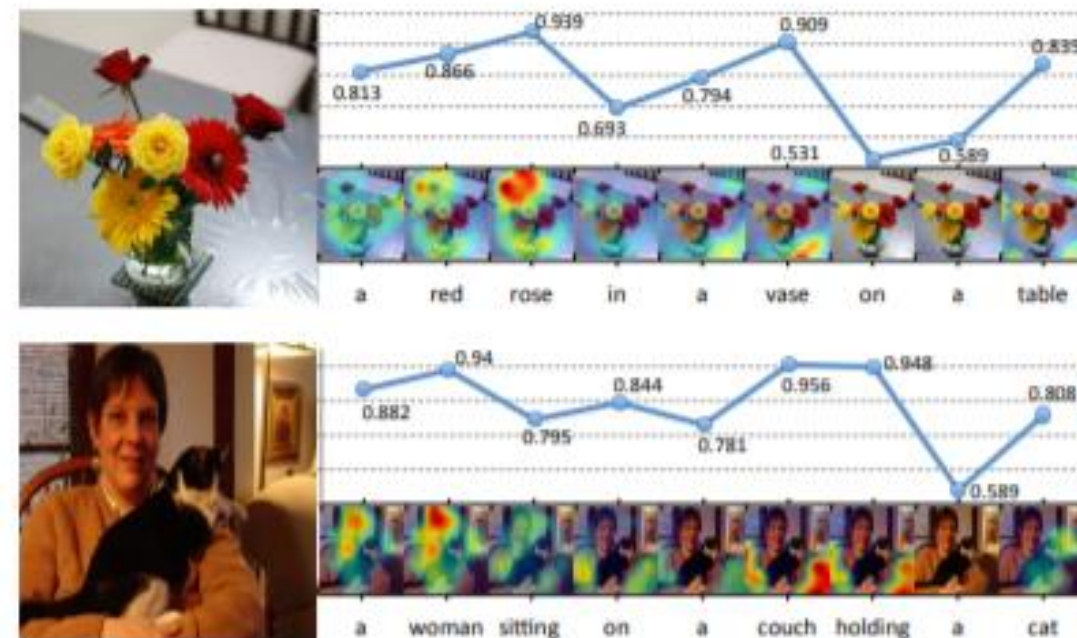
[Show and Tell: A Neural Image Caption Generator.]

提出使用soft-attention和hard-attention机制



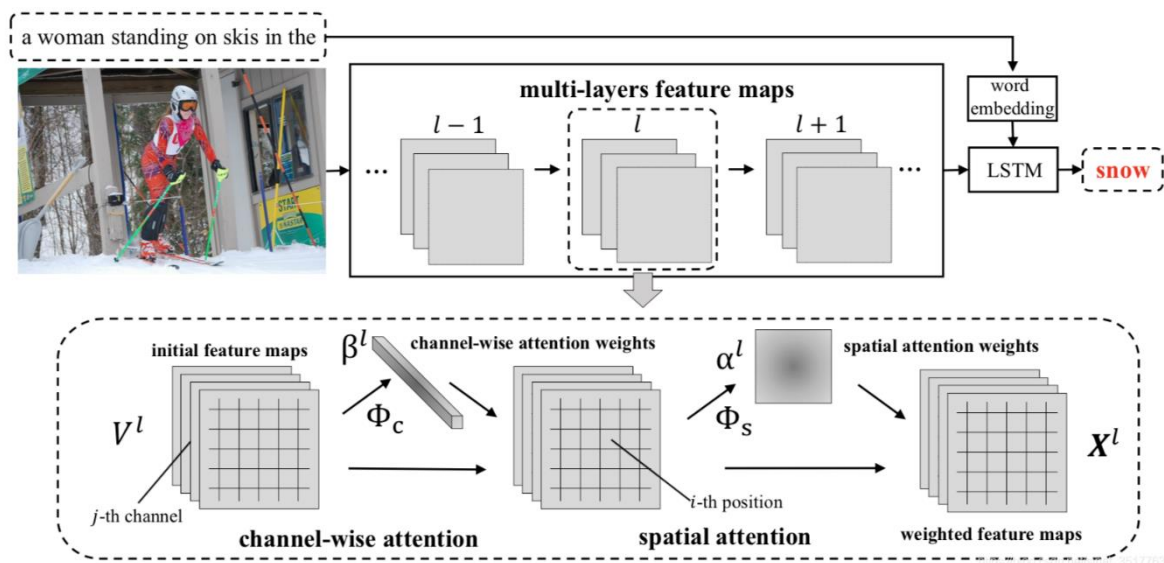
[Show, Attend and Tell: Neural Image Caption Generation with Visual Attention]

提出使用Sentinel哨兵机制机制

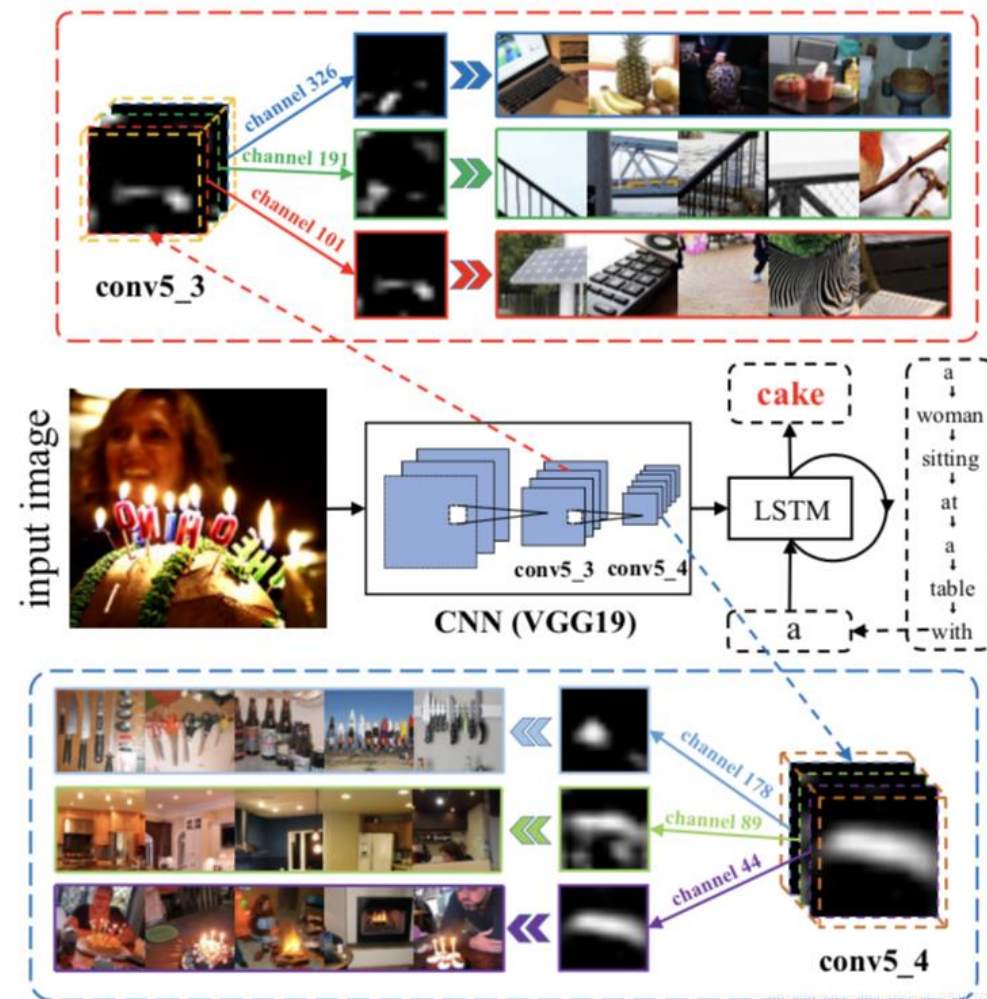


[Show, Attend and Tell: Neural Image Caption Generation with Visual Attention]

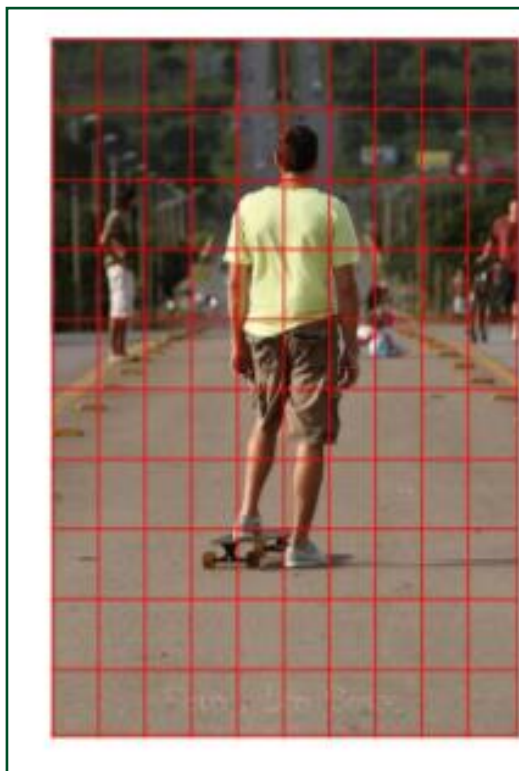
提出增加通道attention机制



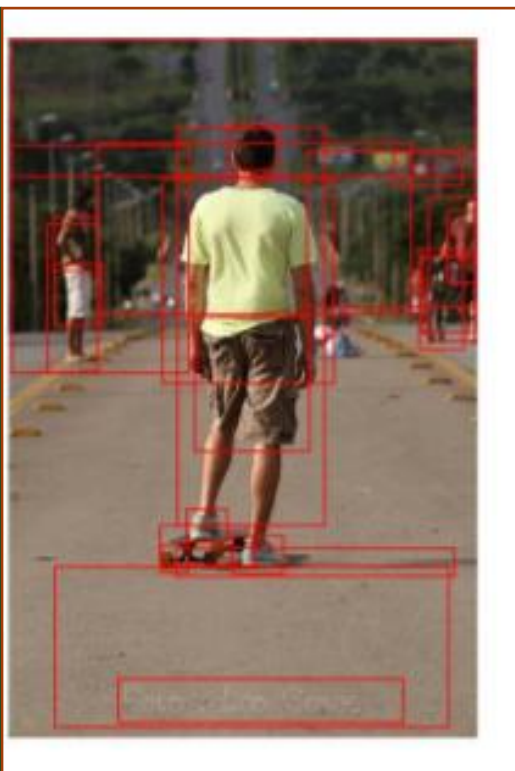
[SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning]



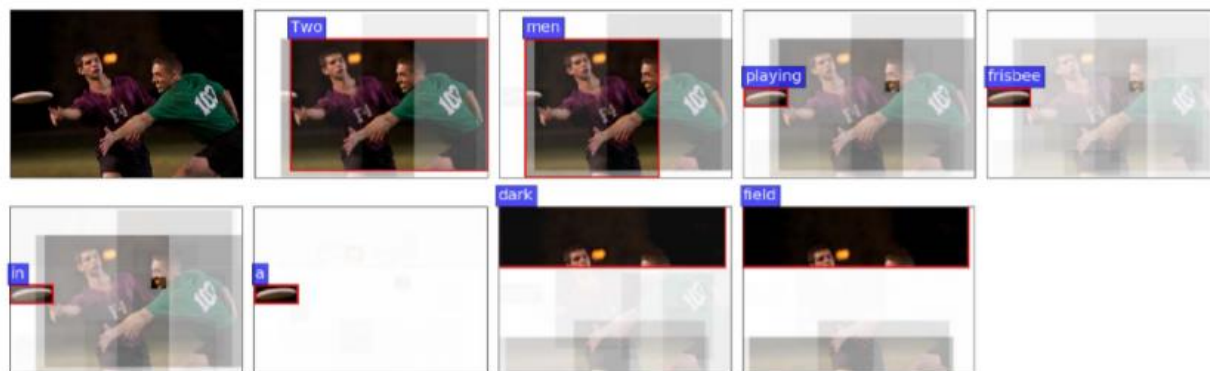
Top-Down



Bottom-Up

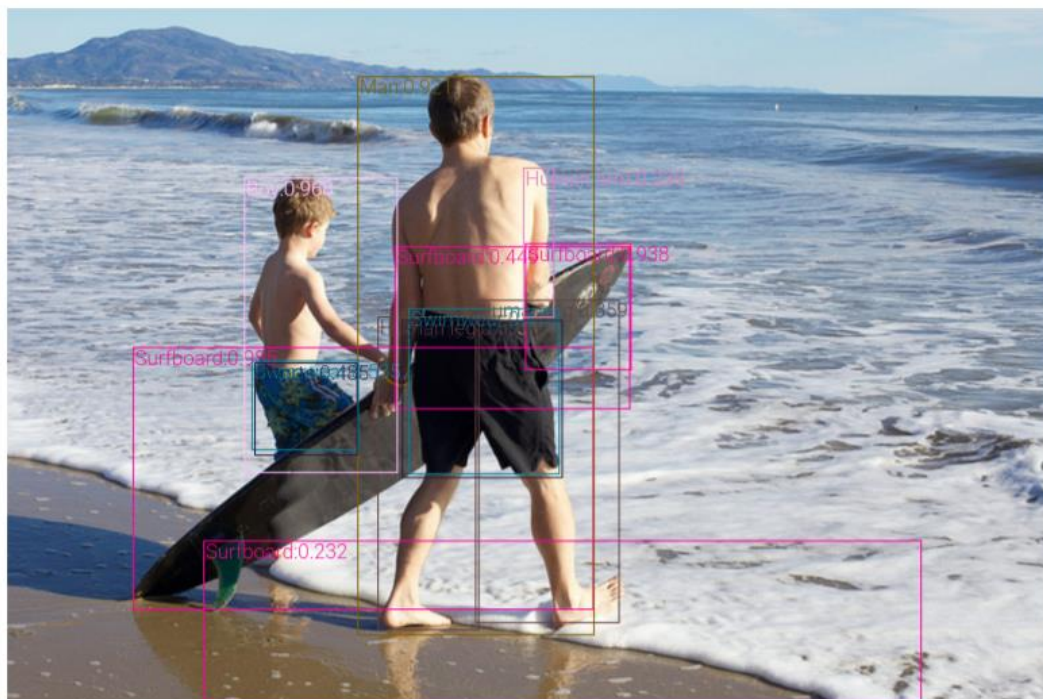


结合**Top-Down**和**Bottom-Up**机制，利用**目标检测Fast-Rcnn**获得自底向上的**注意力特征**

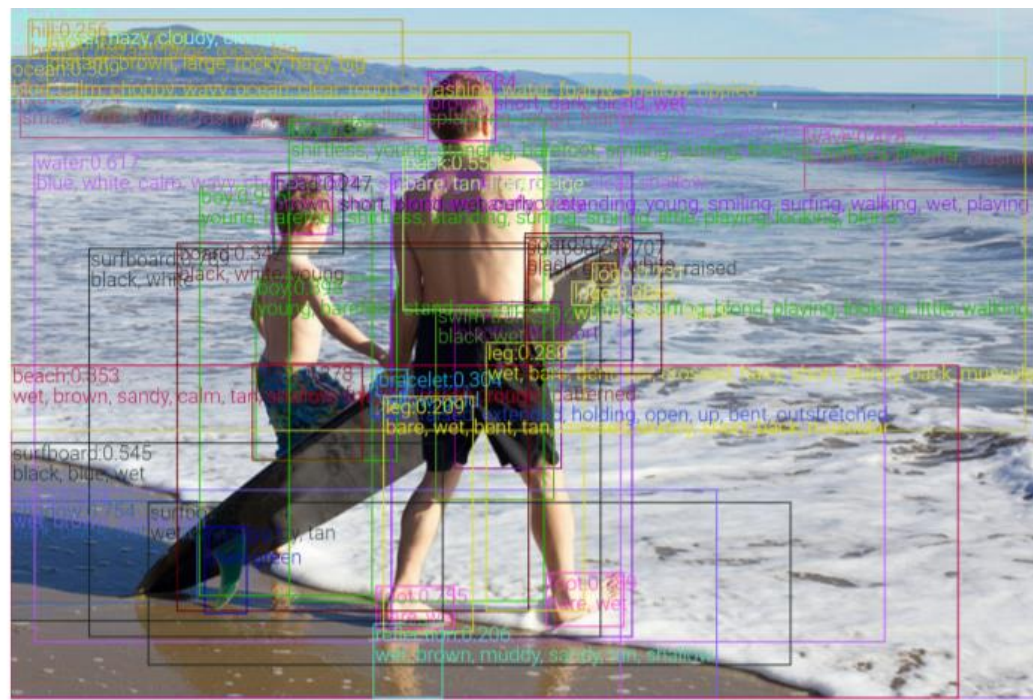


[Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering]

之前的目标检测任务



本文的目标检测任务



[VinVL: Revisiting Visual Representations in Vision-Language Models]



Visual Encoding

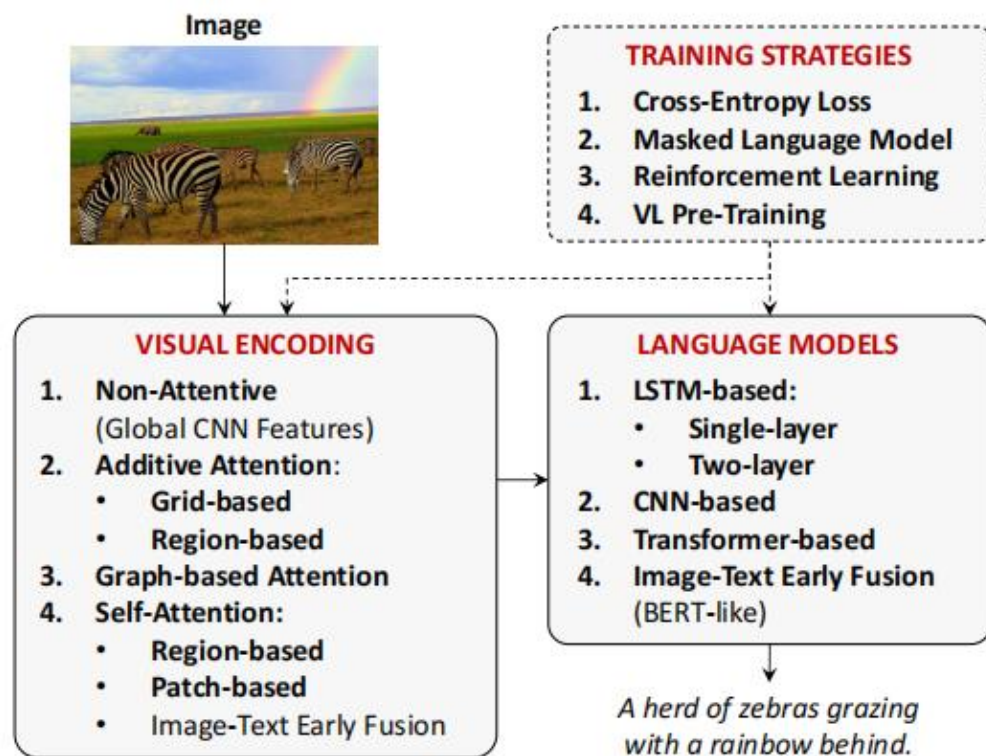
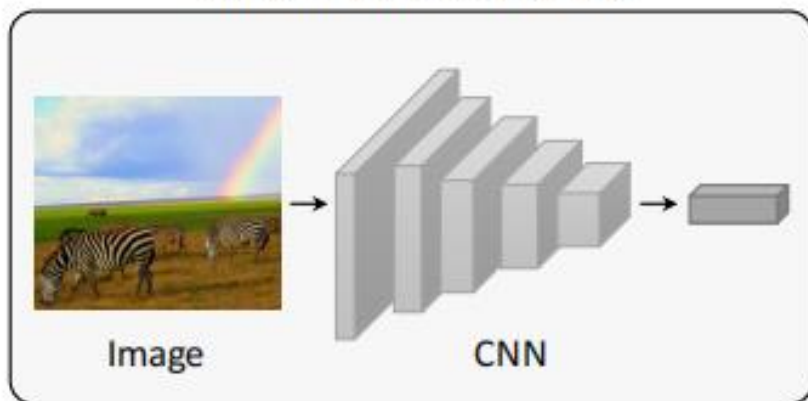


Fig. 1: Overview of the image captioning task and a simple taxonomy of the most relevant approaches.

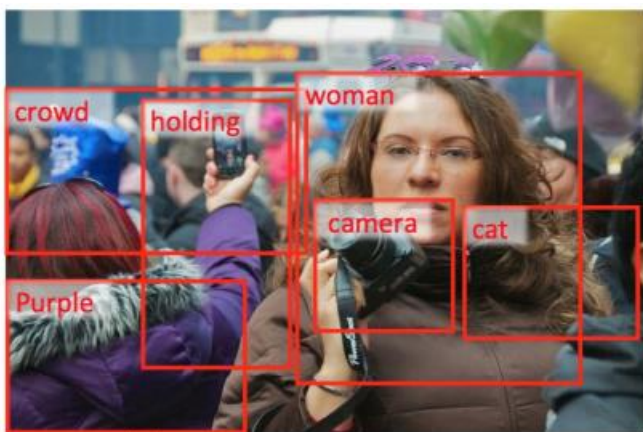
目前的视觉编码分为4大类:

- 1) Non-Attentive
- 2) Additive Attention
- 3) Graph-based Attention
- 4) Self-Attention

Global CNN Features



(a)



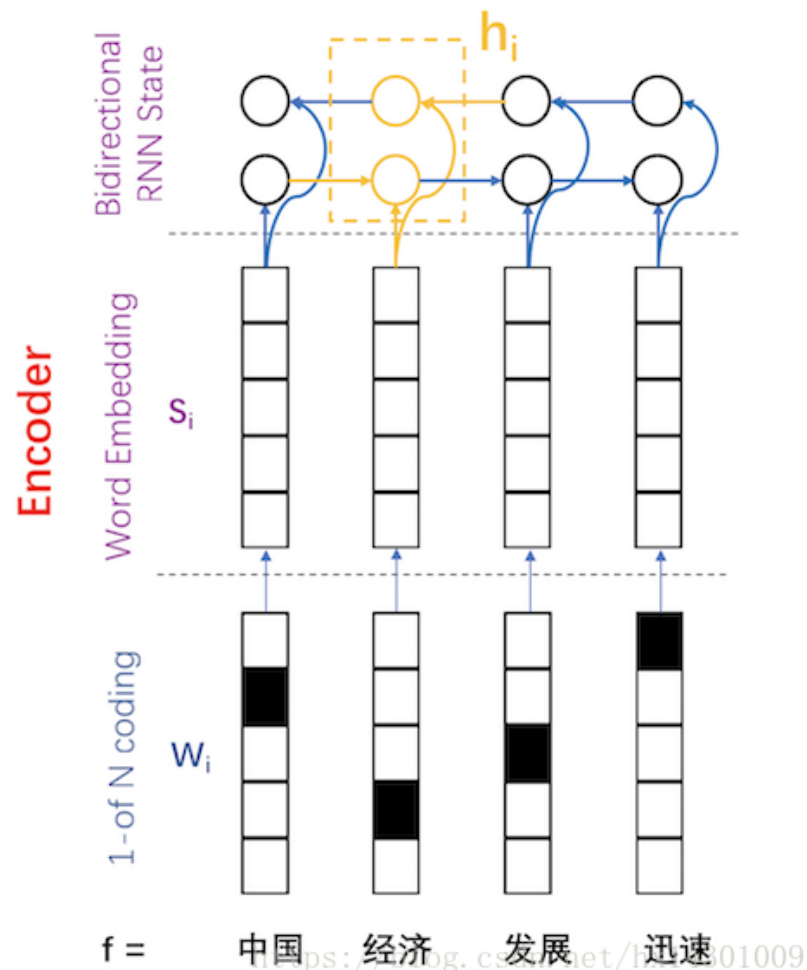
Global CNN Features就是将整幅图片直接输入到CNN网络中。

最简单的方法为：在CNN 的最后一层的激活层提取更高级和固定大小的表示方法，然后将其用作语言模型的条件元素。

优点在于它们表示的简单性和紧凑性，包括从整个输入中提取和浓缩信息并考虑图像的所有上下文的能力。

但是缺点也显而易见，这种方式导致信息过度压缩，所有显着对象和区域都融合在一个向量中，使得字幕模型难以产生特定且细节的描述。

H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt et al., "From captions to visual concepts and back," in CVPR, 2015.



定义：这种方法的直观印象可以看做加权平均法。注意机制最初是设计用来建模两个元素序列（即循环编码器和解码器的隐藏状态之间的关系），但它可以用于将一组视觉表示与语言模型的隐藏状态联系起来。

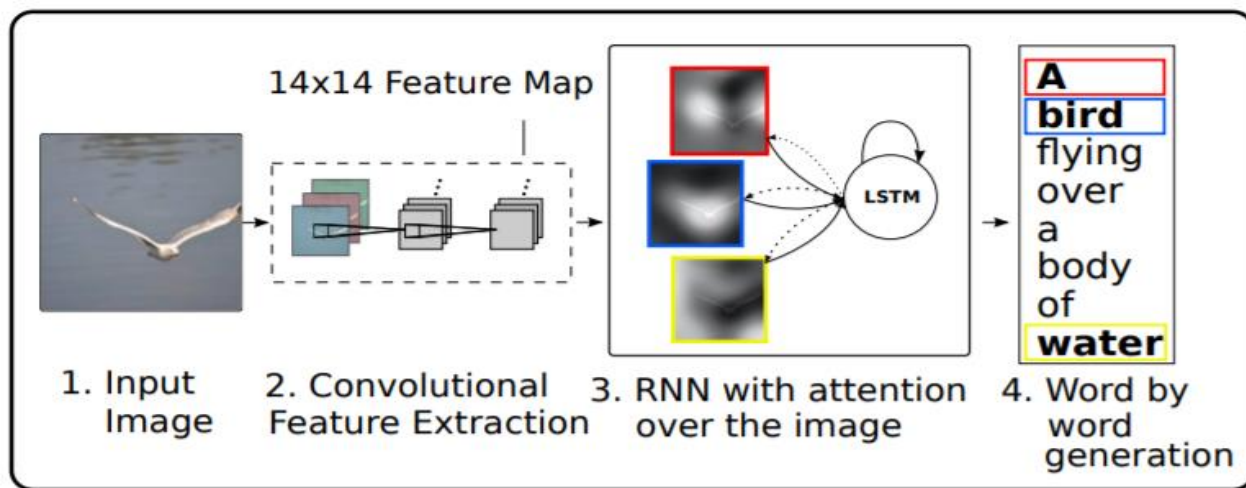
形式上，给定两组向量：输入向量 $\{x_1, \dots, x_n\}$ 和隐藏序列向量 $\{h_1, \dots, h_i\}$ 则 h_i 和 x_j 之间的附加注意力得分计算如下：

$$f_{\text{att}}(h_i, x_j) = W_3^T \tanh(W_1 h_i + W_2 x_j)$$

其中 W_1 和 W_2 是权值矩阵， W_3 是执行线性组合的权值向量

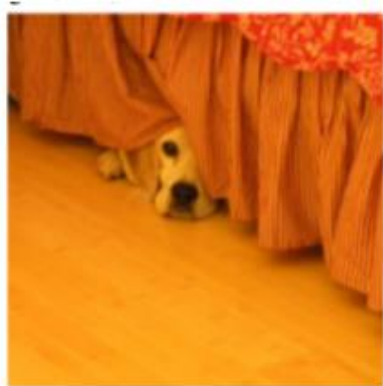
D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in ICLR, 2014

- 1、输入为图片数据
- 2、经过CNN进行卷积，提取图片特征信息最终形成图片的特征图信息
- 3、attention对提取的特征图进行加强与抑制，作为后续进入LSTM模型的输入数据
- 4、LSTM模型最终输出文本信息

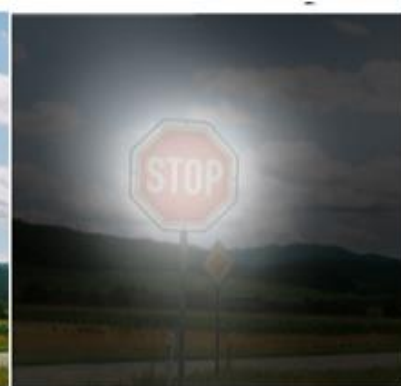


K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in ICML, 2015.

- 1、输入为图片数据
- 2、经过CNN进行卷积，提取图片特征信息最终形成图片的特征图信息
- 3、attention对提取的特征图进行加强与抑制，作为后续进入LSTM模型的输入数据
- 4、LSTM模型最终输出文本信息



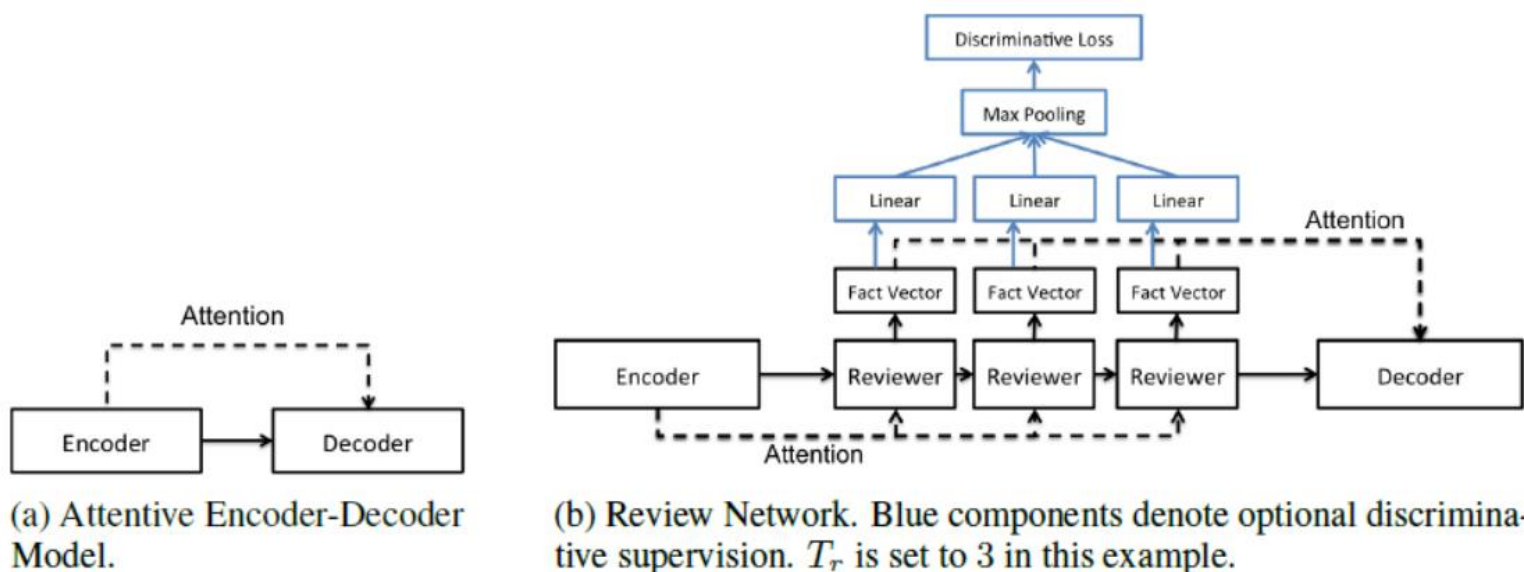
A dog is standing on a hardwood floor.



A stop sign is on a road with a mountain in the background.

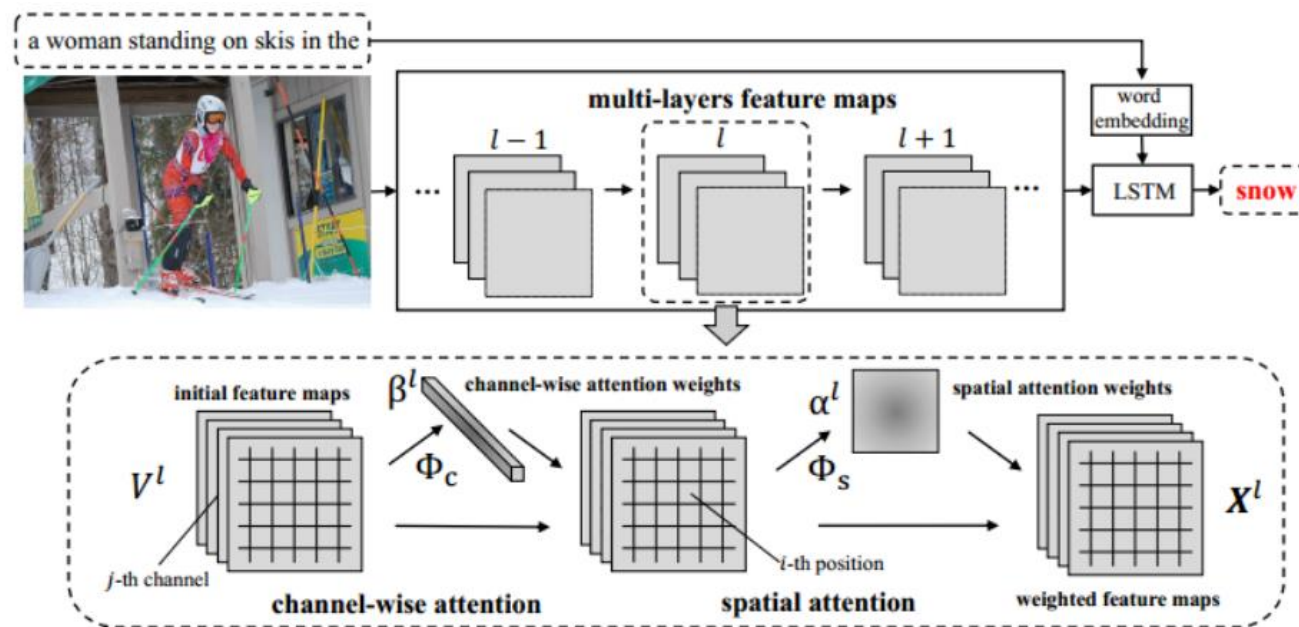
K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in ICML, 2015.

主要思想：注意力机制每次只关注到局部，没有考虑全局因素对预测的影响。该文将feature map作为图片的全局信息，然后通过LSTM单元获得一个比feature map更能表示图片全局信息的一个更紧凑、更抽象的thought vector。

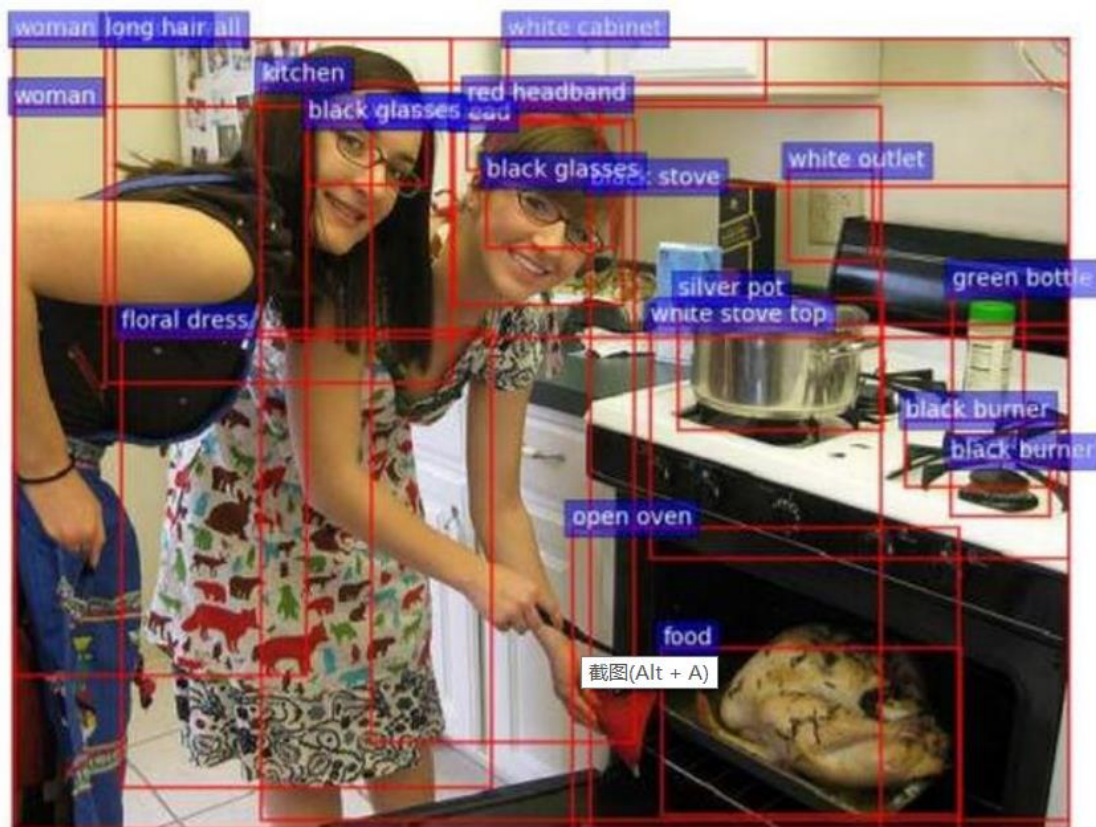


Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov, "Review Networks for Caption Generation," in NeurIPS, 2016.

主要思想：基于CNN的原理，其所提取的特征图具有spatial、channel-wise (semantic)、multi-layer的属性。但现有的一些image caption方法主要考虑spatial的属性。因而在此基础上论文提出，充分利用CNN以上三个属性来进行image caption的任务，提出利用三者结合的属性进行attention机制的应用。



L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning," in CVPR, 2017.



注意力一般分为两种：

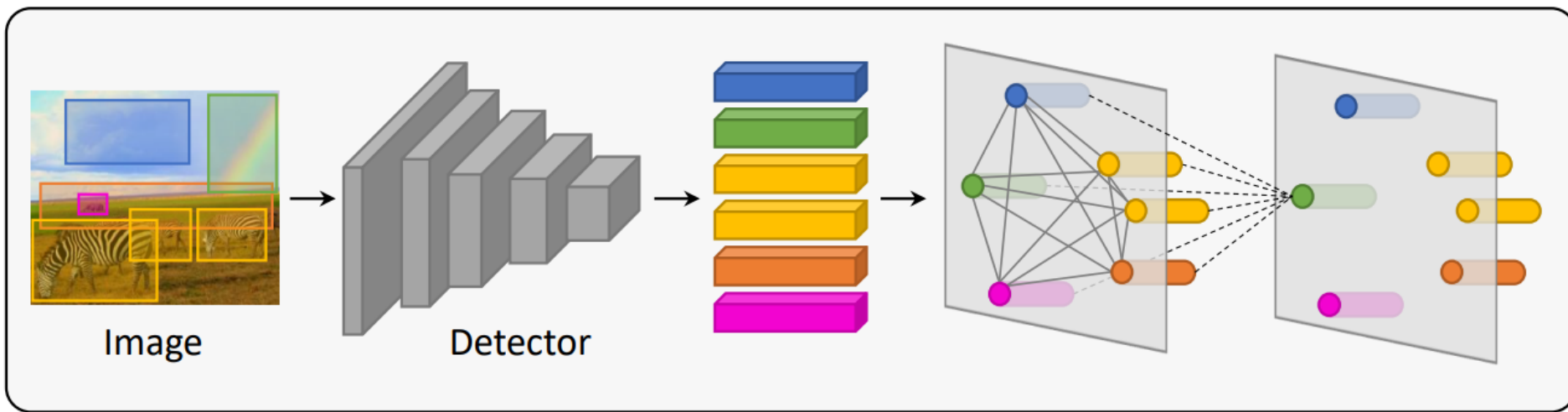
一种是自上而下 (top-down) 的有意识的注意力，称为聚焦式注意力。聚焦式注意力是指有预定目的、依赖任务的、主动有意识地聚焦于某一对象的注意力；

另一种是自下而上(bottom-top)的无意识的注意力，称为基于显著性 (Saliency-Based) 的注意力，也可以说是基于刺激。

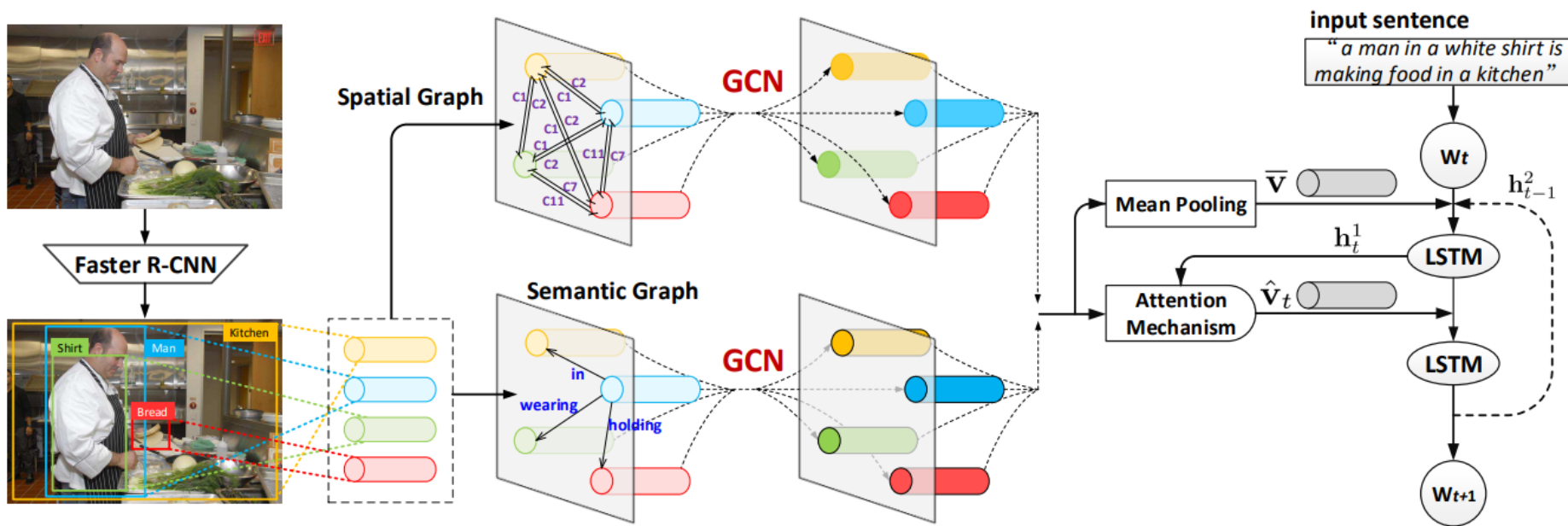
L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning," in CVPR, 2017.

图编码提供了一种利用检测目标之间的关系机制，允许相邻节点间交换信息，从而完成局部交换。它能够无缝集成外部语义信息。而且手动构建的图结构能够限制视觉特征的交互。

- 一些图结构：
 - Spatial and semantic graphs
 - Scene graph
 - Hierarchical trees



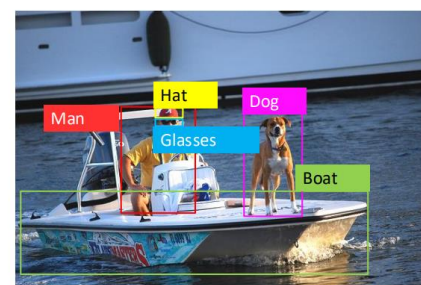
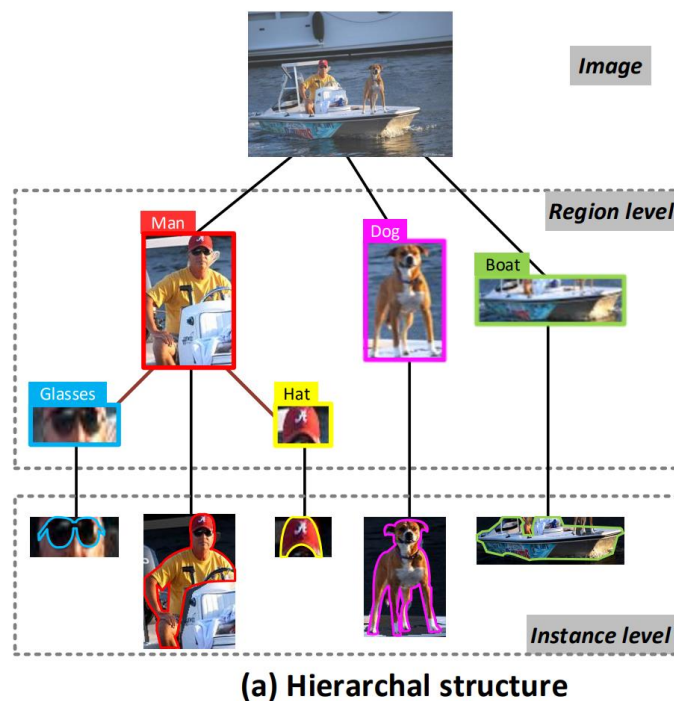
使用图卷积网络GCN来整合目标之间的语义和空间关系。语义关系图预测目标对之间的动作或交互，空间关系图则描述目标对边界框之间的几何度量（即交并比IoU、相对距离和角度等）。



Yao et al. "Exploring Visual Relationship for Image Captioning," in ECCV, 2018.

Guo et al. "Aligning linguistic words and visual semantic units for image captioning," in ACM Multimedia, 2019

Yao等人用树把图像表示为层次结构。其中根节点代表图像整体，中间节点表示图像区域和它们包含的子区域，叶节点则表示区域内的分割对象。之后将图像树送入一个TreeLSTM来获得图像编码。



(b) Regions

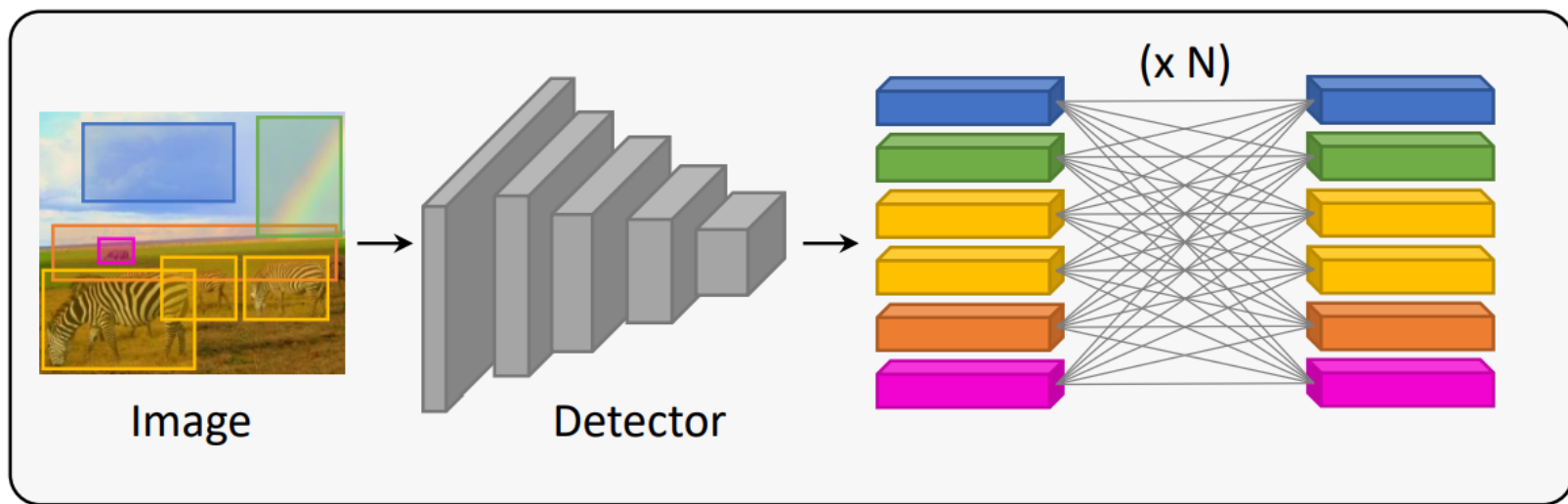


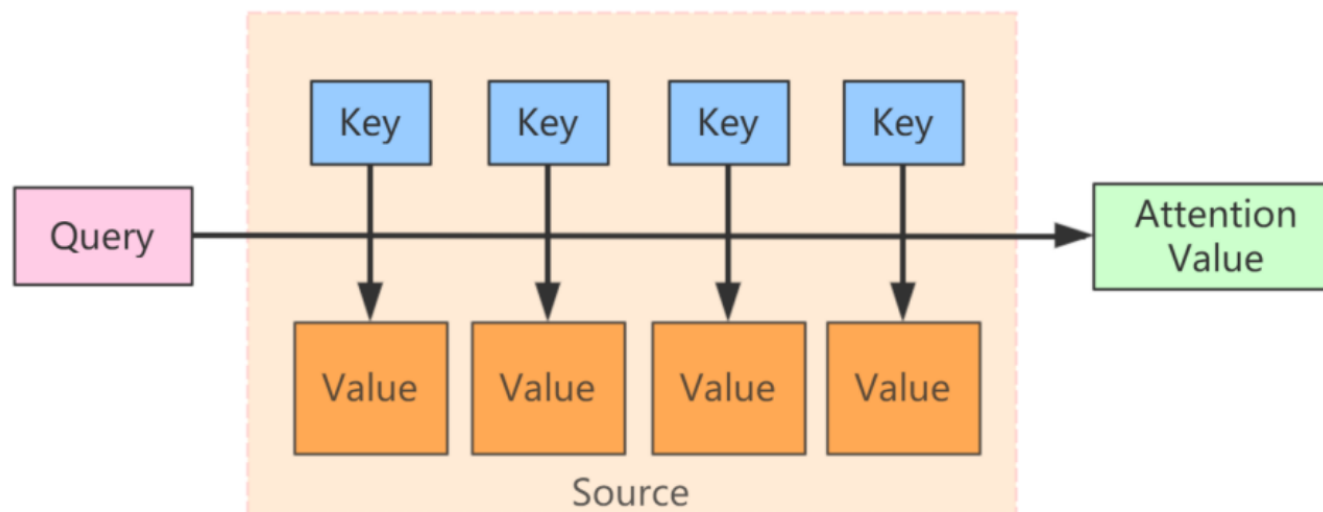
(c) Instances

T. Yao, Y. Pan, Y. Li, and T. Mei, "Hierarchy Parsing for Image Captioning," in ICCV, 2019.

Self-attention自注意力机制可以看作是连接了所有元素来构建的完整的图表示。

自注意力机制最早在2017年由Vaswani提出，由此提出的Transformer在NLP领域占据主导地位，近年来也广泛应用到了计算机视觉中。它是一种特殊的注意力机制，一个集合内的每个元素都与其它所有元素相连，并通过残差的连接计算出该集合内元素的一个新的表示。



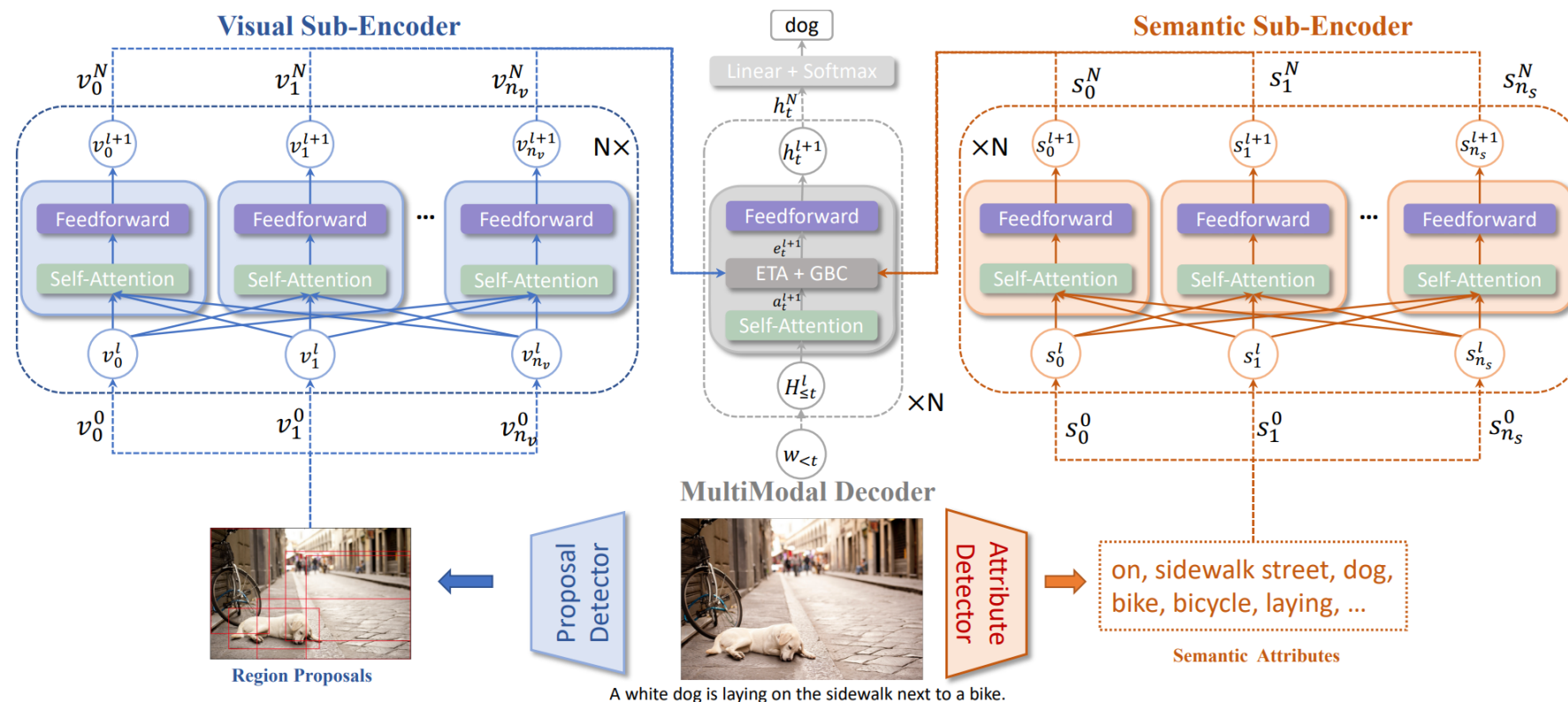


$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} \right) \mathbf{V}$$

Self-attention中, Q,K,V三个向量由同一组元素线性映射得到

Yang等人使用一个自注意力模型来编码区域特征之间的关系。

Li等人提出一个Transformer结构的模型，包含一个视觉encoder和一个语义encoder。



Li et al. "Learning to Collocate Neural Modules for Image Captioning," in ICCV, 2019.



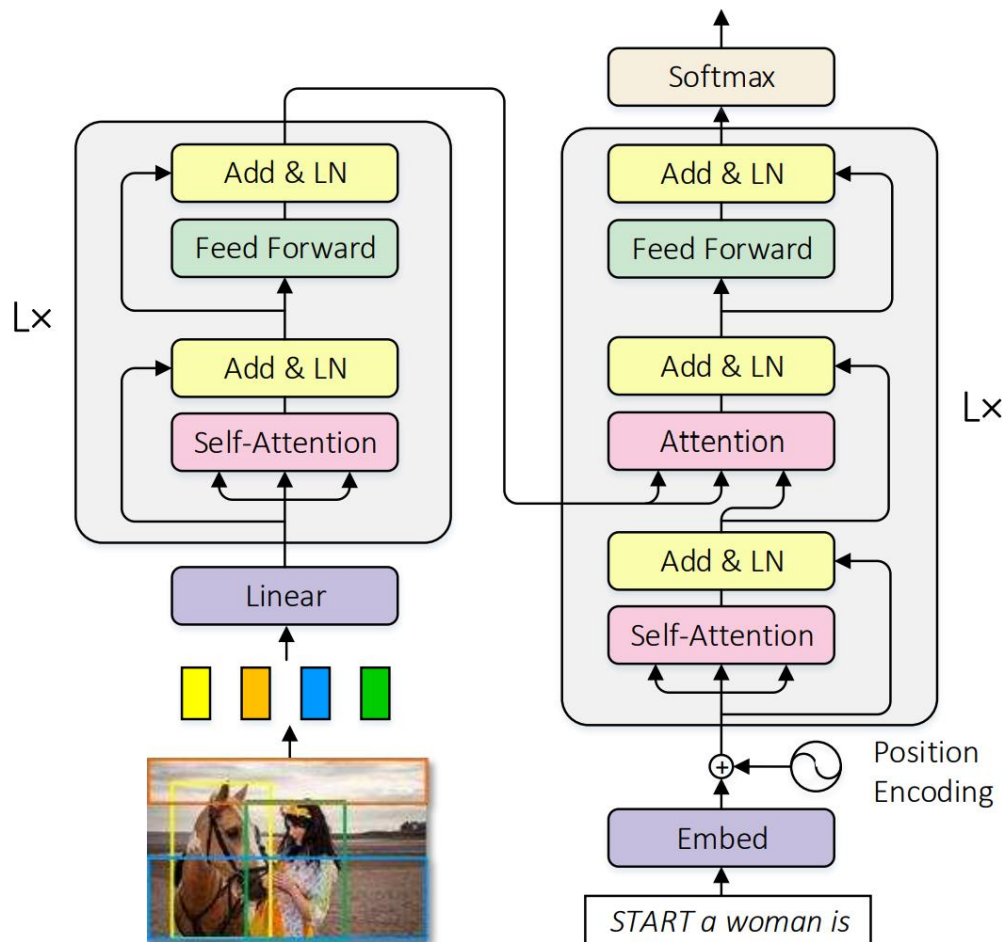
一些工作提出了专门为Image Captioning设计的自注意力变体。

改进self-attention算子：

- Geometry-Aware Self-Attention (几何感知的自注意力计算)
- Attention on Attention
- X-Linear Attention X-线性注意力 [1]
- Memory-augmented Attention 记忆增强注意力 [2]

其它self-attention方法：

- 向特征序列添加一个全局向量 [3]
- 混合使用区域和网格特征编码 [4]
- 直接应用于网格特征 [5]



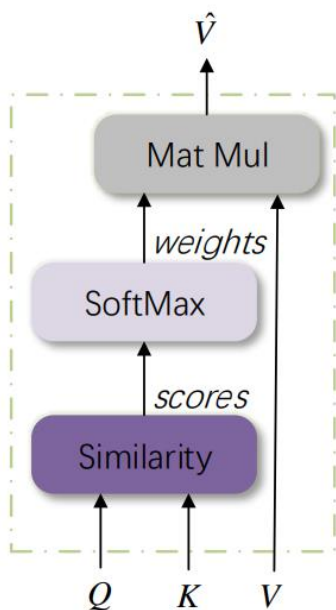
Geometry-Aware Self-Attention (几何感知的自注意力计算)

$$Z = \text{Attention}(Q, K, V) = \text{Softmax}(E) V$$

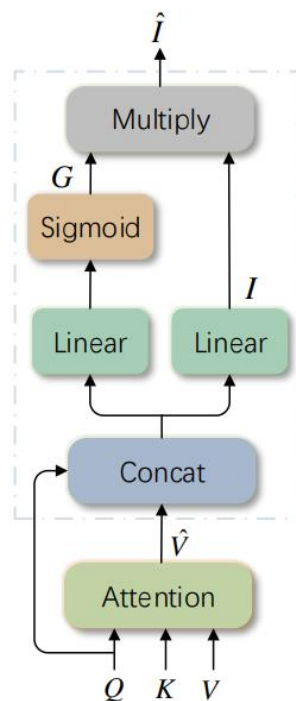
$$E = QK^T$$

$$E = QK^T + \phi(Q', K', G)$$

Herdade et al. "Normalized and Geometry-Aware Self-Attention Network for Image Captioning," in CVPR, 2020



(a) Attention



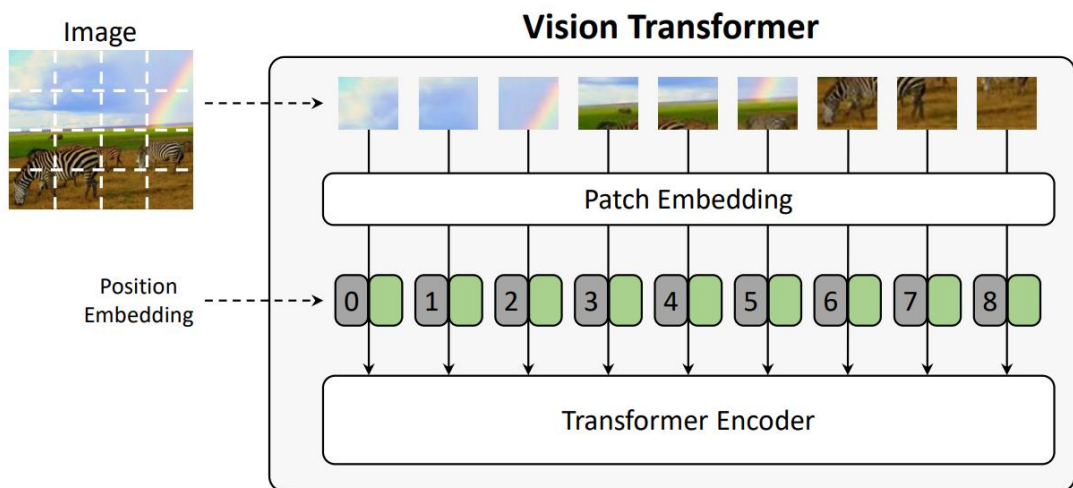
(b) Attention on Attention

Attention on Attention

经典的注意力存在问题：无论Q是否和K/V相关，attention都会为每个查询输出加权平均值。

AoA 测量attention结果和查询之间的相关性，可以避免无关向量的误导信息。

L. Huang et al. "Attention on Attention for Image Captioning," in ICCV, 2019.



Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ICLR, 2021

除了向上述方法那样将object region作为attention的输入序列，Transformer结构还可以应用在图像块patches上，从而减少或直接舍弃卷积运算。

Liu等人设计了第一个convolution-free的image captioning架构。它使用了预训练的ViT作为encoder，然后使用标准的Transformer decoder去生成字幕。[6]



- [1] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-Linear Attention Networks for Image Captioning," in CVPR, 2020.
- [2] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed Memory Transformer for Image Captioning," in CVPR, 2020
- [3] J. Ji, Y. Luo, X. Sun, F. Chen, G. Luo, Y. Wu, Y. Gao, and R. Ji, "Improving Image Captioning by Leveraging Intra- and Interlayer Global Representation in Transformer Network," in AAAI, 2021.
- [4] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, and R. Ji, "Dual-Level Collaborative Transformer for Image Captioning," in AAAI, 2021.
- [5] X. Zhang, X. Sun, Y. Luo, J. Ji, Y. Zhou, Y. Wu, F. Huang, and R. Ji, "RSTNet: Captioning with Adaptive Attention on Visual and Non-Visual Words," in CVPR, 2021.
- [6] W. Liu, S. Chen, L. Guo, X. Zhu, and J. Liu, "CPTR: Full Transformer Network for Image Captioning," arXiv preprint arXiv:2101.10804, 2021.



语言模型 (Language Model)

语言模型主要是用来预测一个给定单词序列在句子中出现的概率，是许多NLP任务的一个关键组成部分，赋予了及其在一个随机过程中理解和解决语言问题的能力

对于一个给定 n 个单词的序列 $(y_1, y_2, \dots, y_n)$, image caption的语言模型会为这个序列分配一个概率

$$P(y_1, y_2, \dots, y_n | \mathbf{X}) = \prod_{i=1}^n P(y_i | y_1, y_2, \dots, y_{i-1}, \mathbf{X})$$

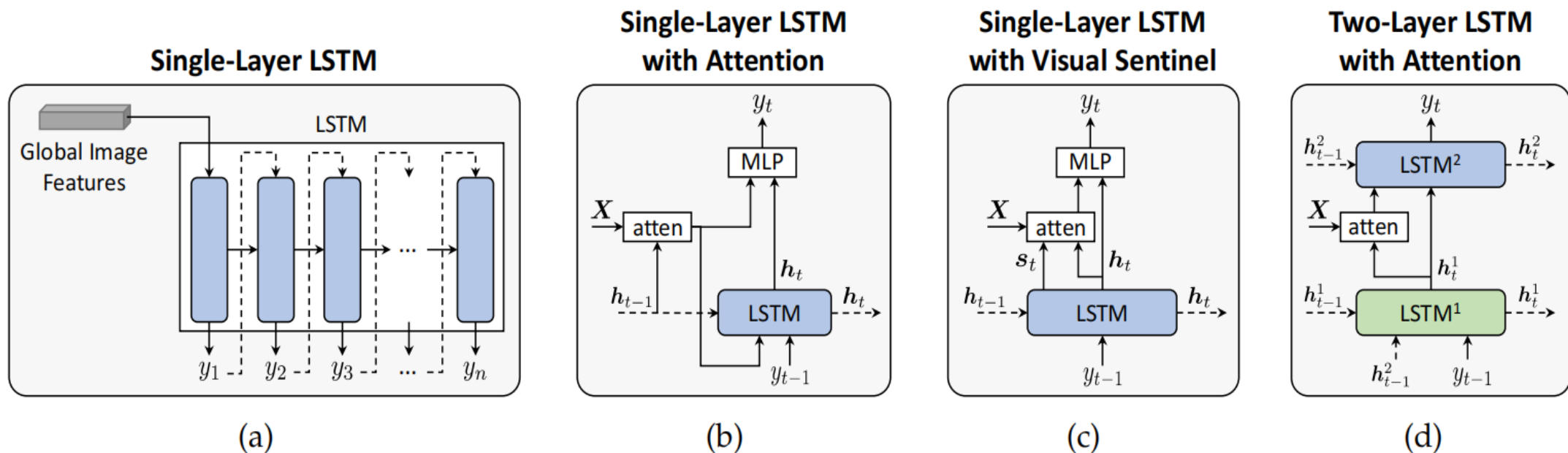
- 在预测每一个单词时都是以已知单词为条件，在已知单词的基础上预测下一个单词。
- 同时在image caption中也会通过输出一个序列结束的token来终止image caption的生成。

LANGUAGE MODELS

1. LSTM-based:
 - Single-layer
 - Two-layer
2. CNN-based
3. Transformer-based
4. Image-Text Early Fusion (BERT-like)

单层的LSTM:将视觉编码作为LSTM的初始隐藏状态, 然后生成image caption。在每一个时间步中通过将隐藏状态投影到与词汇相同大小的向量上, 然后将softmax作为激活函数来预测单词, 每一步的输出就是单词表中所有单词的概率。

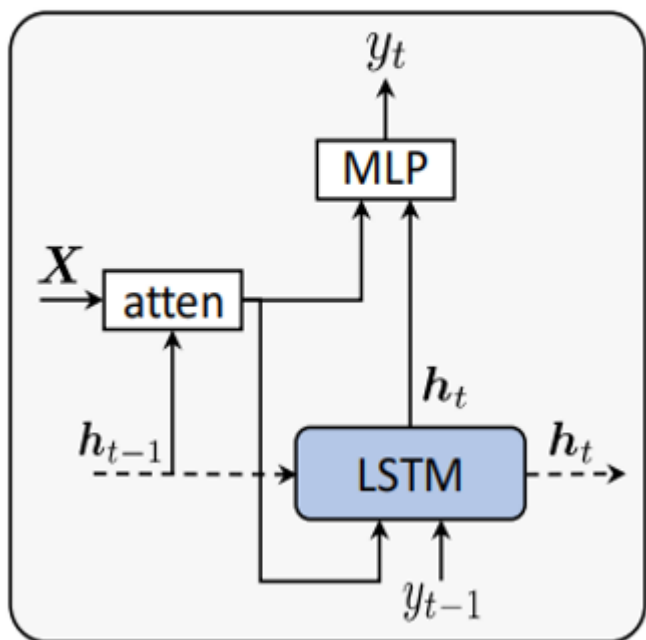
- 在训练过程中, 输入是ground-truth的句子, 也就是word embedding。
- 在推理过程中, 输入是之前已经产生的单词。



O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in CVPR, 2015

单层带attention的LSTM：使用动态和时变的图像表示来代替静态的全局向量。

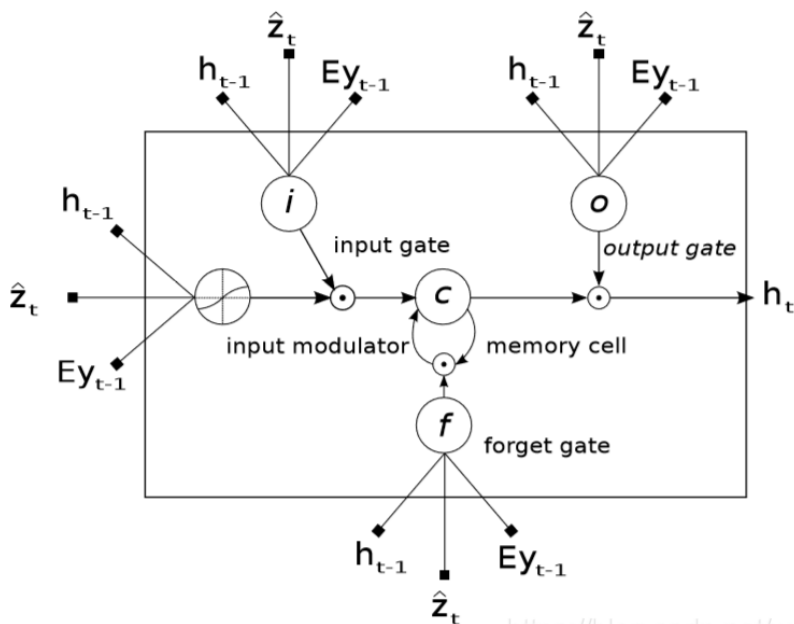
Single-Layer LSTM with Attention



在提取到图像的位置特征之后，利用 attention 机制使 decoder 具有在特征中选择的能力，在早期的隐藏状态会指导视觉特征的 attention，计算出一个上下文向量，然后将其放入 MLP 中来预测输出的单词。

K. Xu, J. Ba, R. Kiros "Show, attend and tell: Neural image caption generation with visual attention," in ICML, 2015.

单层带attention的LSTM：使用动态和时变的图像表示来代替静态的全局向量。



经过attention的图像信息，隐藏状态和word embedding一起送入LSTM

在生成第 t 个单词的时候，传入LSTM的向量为 z_t ，前一阶段的隐藏状态为 h_{t-1} ，图片特征为 $a = \{a_1, a_2, \dots, a_L\}, a_i \in R^D$ (位置特征的集合)

$$z_t = \sum_{i=1}^L \alpha_{t,i} a_i$$

Attention 的核心就是去学习这个位置集合的权重

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^L \exp(e_{tk})}$$

$$e_{ti} = f_{\text{att}}(a_i, h_{t-1})$$

f_{att} : attention model

K. Xu, J. Ba, R. Kiros "Show, attend and tell: Neural image caption generation with visual attention," in ICML, 2015.

其他方法：

- Visual sentinel:

J. Lu et al. "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in CVPR, 2017. 使用visual sentinel来捕获空间图像特征。在每个时间步长，visual sentinel从之前的隐藏状态计算和生成的单词。

- Hidden state reconstruction:

X. Chen et al. "Regularizing RNNs for Caption Generation by Reconstructing The Past with The Present," in CVPR, 2018. 利用第二个LSTM来重构基于当前状态的先前隐藏状态来规范语言模型的动态转换。

H. Ge et al. "Exploring Overall Contextual Information for Image Captioning in Human-Like Cognitive Style," in ICCV, 2019. 通过使用带有辅助模块的双向 LSTM 来更好地捕获上下文信息。

- Multi-stage generation:

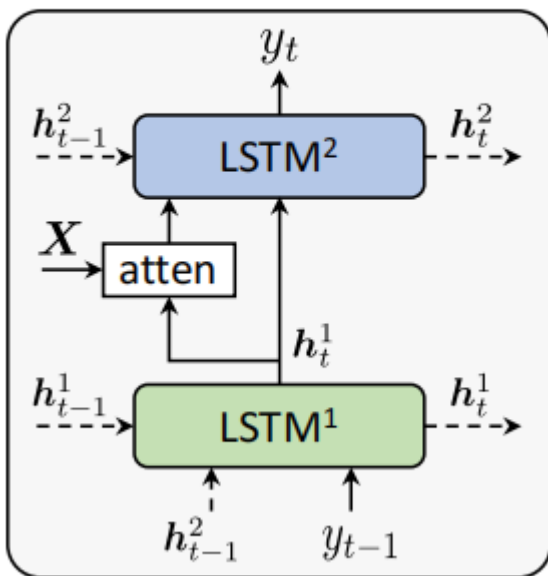
Y. Wang et al. "Skeleton Key: Image Captioning by Skeleton-Attribute Decomposition," in CVPR, 2017. 生成过程分解为两个阶段，两个阶段都是单层的LSTM。

J. Gu et al. "Stack-Captioning: Coarse-to-Fine Learning for Image Captioning," in AAAI, 2018.

一个粗到细的多阶段框架，使用 LSTM 解码器序列，每个解码器都操作前一个的输出，以产生越来越精细的标题。

双层LSTM:将LSTM扩展到多层结构, 来增强捕获高阶语义信息的能力。最早的两层LSTM作为image caption的语言模型是将第一层的隐藏状态作为第二层的输入。

Two-Layer LSTM with Attention



Two-layers and additive attention:

将第一个LSTM层作为一个自上而下的视觉注意模型, 它可以获取之前生成的单词, 早先的隐藏状态和平均池化图像特征; 然后当前的隐藏状态被用于计算带注意力机制的图像区域上的概率分布。将得到的图像特征放入到第二个LSTM层中, 与第一层的隐藏状态相结合, 从而生成词汇的概率分布。

P. Anderson et al. "Bottom-up and top-down attention for image captioning and visual question answering," in CVPR, 2018.

Variants of two-layers LSTM:

- Neural Baby Talk:

J. Lu et al. "Neural Baby Talk," in CVPR, 2018.

- Reflective attention:

L. Ke et al. " Reflective Decoding Network for Image Captioning," in ICCV, 2019.

- Look back and predict forward:

Y. Qin et al. "Look Back and Predict Forward in Image Captioning," in CVPR, 2019.

- Adaptive attention time:

L. Huang et al. "Adaptively Aligned Image Captioning via Adaptive Attention Time," in NeurIPS, 2019.

- Recall mechanisms:

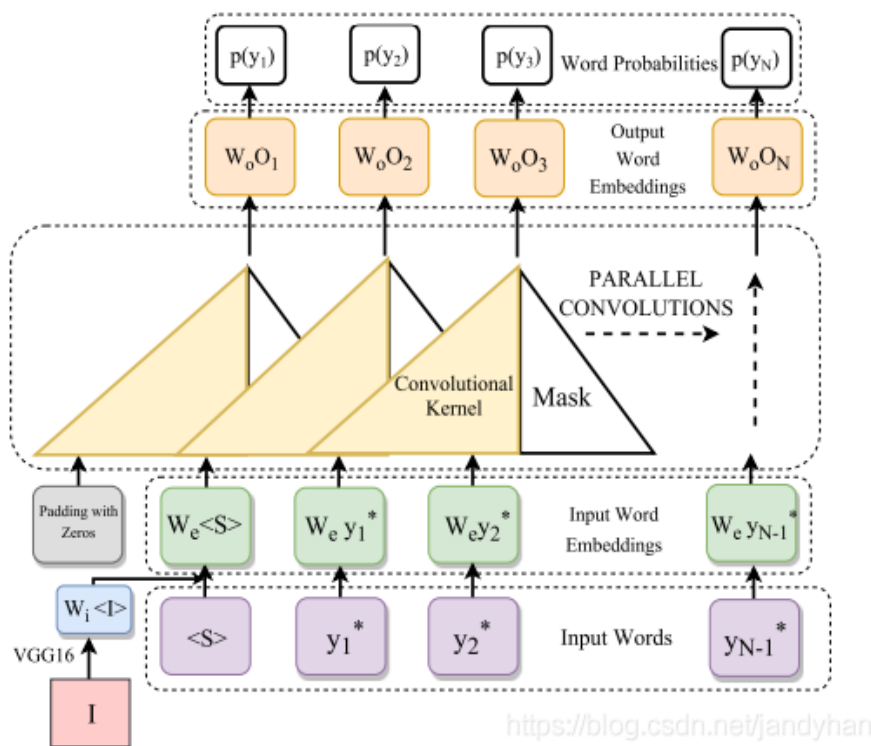
L. Wang et al. "Show, Recall, and Tell: Image Captioning with Recall Mechanism," in AAAI, 2020.

X. Yang et al. "Auto-Encoding Scene Graphs for Image Captioning," in CVPR, 2019.

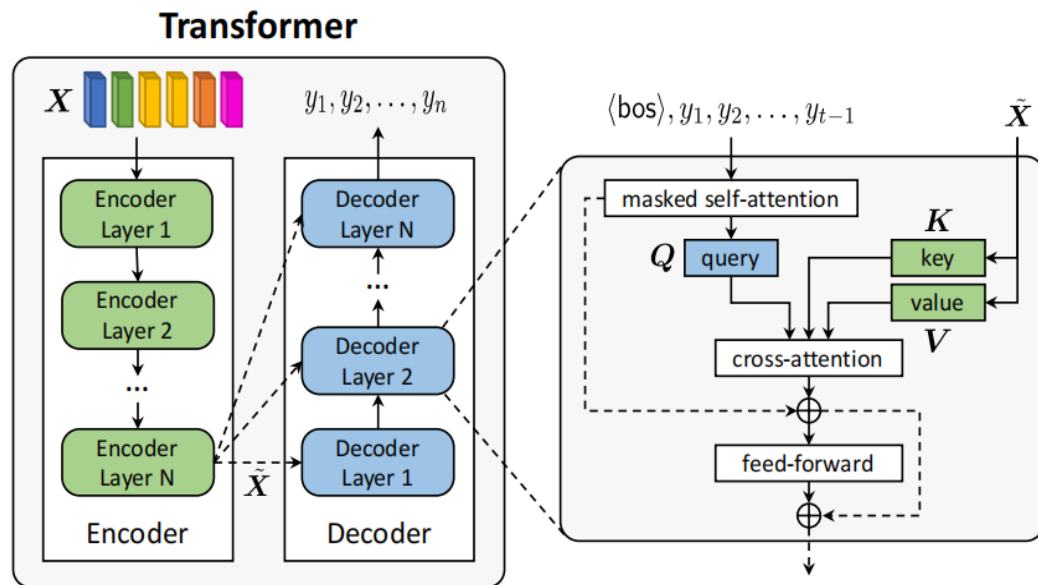
Z. Shi et al. "Improving Image Captioning with Better Use of Captions," in ACL, 2020.

T. Yao et al. "Hierarchy Parsing for Image Captioning," in ICCV, 2019

将全局图片特征向量和词语embedding相结合并输入置CNN中。在训练过程中会并行地对所有单词操作并且在推理中按顺序操作。卷积是被向右掩盖的，防止模型使用之后单词标记的信息。尽管并行训练具有优势，但是由于性能和transformer的出现，在语言模型中进行卷积的操作并未受到广泛欢迎。



J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in CVPR, 2018.



标准的transformer解码器执行隐藏注意力操作，并将其应用到单词上，然后执行交叉自注意操作，其中单词作为query，最后一个编码器层的输出作为keys和value，然后接入前馈神经网络。在训练过程中mask机制用于已知的单词，其目的是为了约束这个单向生成过程。

Gating mechanisms

G. Li, L. Zhu, P. Liu, and Y. Yang, “Entangled Transformer for Image Captioning,” in ICCV, 2019.

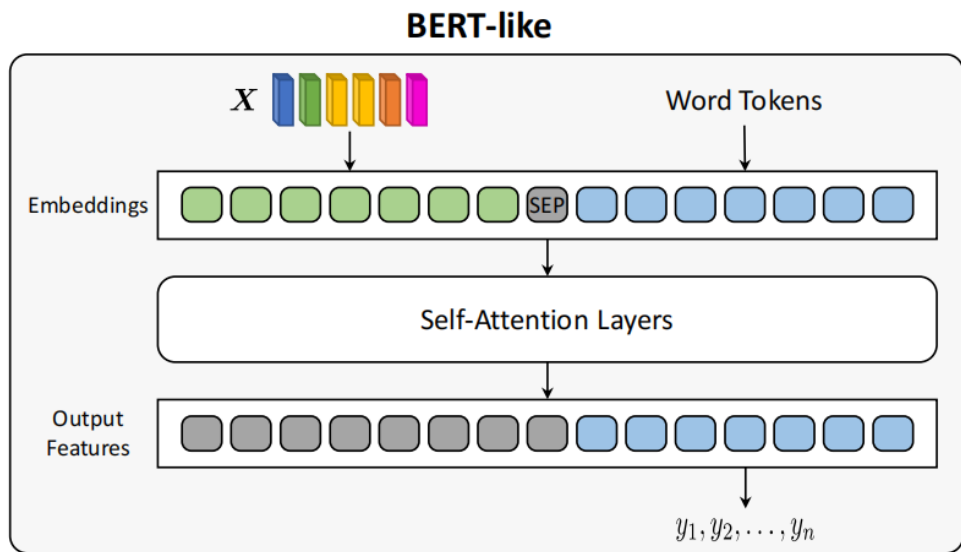
一种交叉注意操作符的门控机制，该机制通过结合 和调制来自外部标记器的图像区域表示和语义属性来控制视觉和语义信息的流动。

J. Ji, Y. Luo, X. Sun, F. Chen, G. Luo, Y. Wu, Y. Gao, and R. Ji, “Improving Image Captioning by Leveraging Intra- and Interlayer Global Representation in Transformer Network,” in AAAI, 2021.

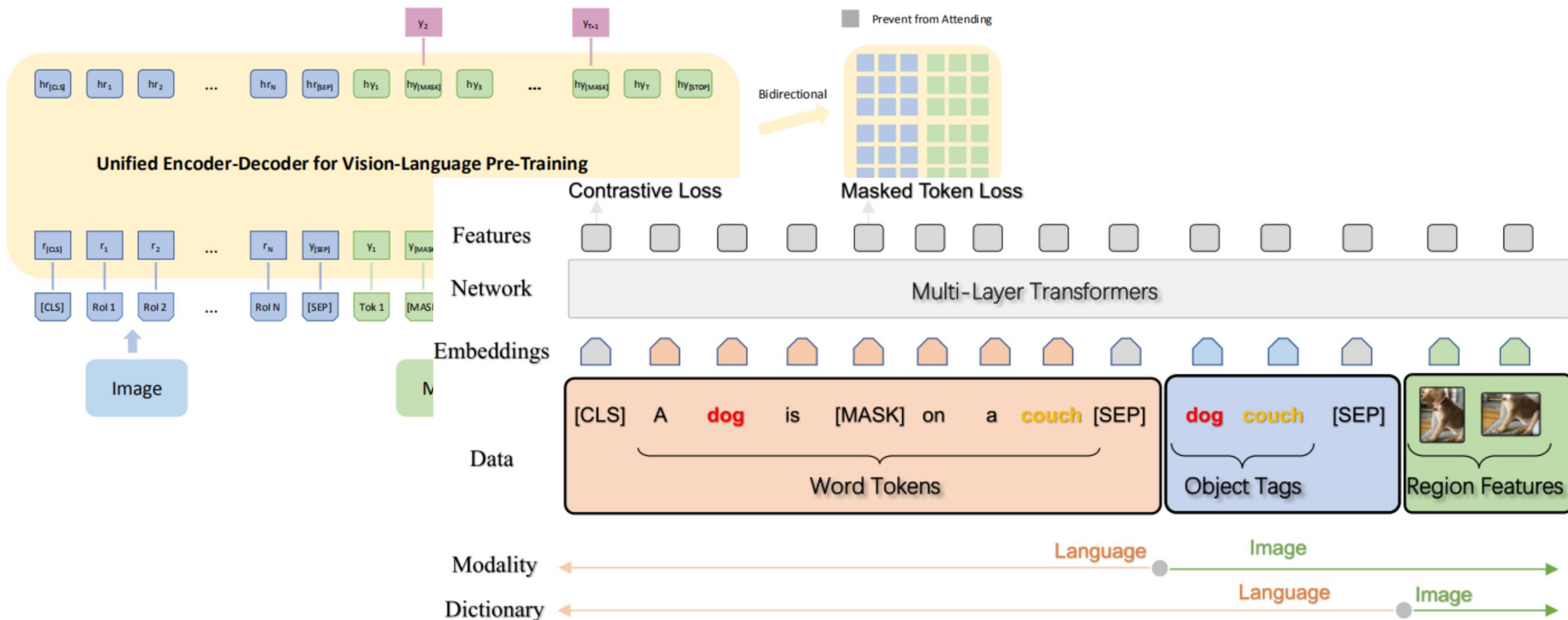
集成了一个上下文门 控机制来调节全局图像表示对每个生成的单词的影响，通过多头注意进行建模

M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, “Meshed-Memory Transformer for Image Captioning,” in CVPR, 2020.

考虑所有的编码层，而不是只对最后一个编码层执行交叉注意。设计了网格解码器，它包含一个独立调制所有编码层的贡献，以及一个门来对文本查询引导的这些贡献进行权重。



文本和视觉特征在早期就融合到一起作为image caption的语言模型。这种模型的优点是**可以从大量的文本语料库中预训练学习的参数来初始化作为处理文本的层的参数**



L. Zhou et al. "Unified Vision-Language Pre-Training for Image Captioning and VQA," in AAAI, 2020

X. Li et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks," in ECCV, 2020.



数据集和评价指标

VizWiz

TextCaps

COCO

TextCaps



*Woman on
a horse jumping
over a pole jump.*

*A glass bowl contains
peeled tangerines and
cut strawberries.*



A person is holding a small container of cream upside down.



*The billboard displays
'Welcome to Yakima The
Palm Springs of Washington'.*



Conceptual Captions

Fashion Captioning

CUB-200

Fashion Captioning

CUB-200



Cars are on the streets.

Small stand of trees, just visible in the distance in the previous photo.



*A decorative leather
padlock on a compact bag
with croc embossed leather.*



This bird is blue with white on its chest and has a very short beak.



(a)

(b)

	Domain	Nb. Images	Nb. Caps (per Image)	Vocab Size	Nb. Words (per Cap.)
MS COCO	Generic	132K	5	27K (10K)	10.5
Flickr30K	Generic	31K	5	18K (7K)	12.4
Flickr8K	Generic	8K	5	8K (3K)	10.9
CC3M	Generic	3.3M	1	48K (25K)	10.3
CC12M	Generic	12.4M	1	523K (163K)	20.0
SBU Captions	Generic	1M	1	238K (46K)	12.1
VizWiz	Assistive	70K	5	20K (8K)	13.0
CUB-200	Birds	12K	10	6K (2K)	15.2
Oxford-102	Flowers	8K	10	5K (2K)	14.1
Fashion Cap.	Fashion	130K	1	17K (16K)	21.0
BreakingNews	News	115K	1	85K (10K)	28.1
GoodNews	News	466K	1	192K (54K)	18.2
TextCaps	OCR	28K	5/6	44K (13K)	12.4
Loc. Narratives	Generic	849K	1/5	16K (7K)	41.8

表1 主要数据集概述

- **Standard captioning datasets:** 早期的Image Captioning架构通常在 Flickr30K 和 Flickr8K 数据集上进行训练和测试，这些数据集由从 Flickr 网站收集的图片组成，包含日常活动、事件和场景，每个都配有5个caption。目前，最常用的Image Captioning数据集是 Microsoft COCO，它由复杂场景的图像组成，其中包含自然环境中的人、动物和常见日常物体。
- **Pre-training datasets:** 多样性较少和质量较低的Image Captioning数据集，或者用于其他任务（例如视觉问答）数据集可以用来预训练。SBU Captions，最初用于将Image Captioning作为检索任务处理。Conceptual Captions数据集，该数据集包含大约 330 万张 (CC3M) 和 1200 万张 (CC12M) 图像，并搭配一个从网络上自动收集的弱关联描述。
- **Domain-specific datasets:** VizWiz Captions 旨在支持Image Captioning对辅助技术的研究，该数据集中的图像是由视障人士用手机拍摄的。CUB-200 和 Oxford-102分别包含鸟和花的图像，并配有10 个caption。Fashion Captioning包含可能共享相同caption的不同姿势和颜色的服装的图像。BreakingNews 和 GoodNews 强制使用更丰富的词汇，它们的图像取自新闻文章。“Localized Narratives”，其caption是通过记录人们自由叙述他们在图像中看到的内容来收集的(包含局部的描述)。



Flickr8k: <https://academictorrents.com/details/9dea07ba660a722ae1008c4c8afdd303b6f6e53b>

Flickr30k: <https://pan.baidu.com/s/1r0RVUwctJsI0iNuVXHQ6kA> code: hrf3

COCO: <https://pan.baidu.com/s/1QT-s0iwVYlClMThVySzuTQ> code: 2pyr

SBU Captions: <http://www.cs.virginia.edu/~vicente/sbucaptions/>

CC3M: <https://ai.google.com/research/ConceptualCaptions/download>

CC12M: <https://github.com/google-research-datasets/conceptual-12m>

VizWiz: <https://vizwiz.org/tasks-and-datasets/image-captioning/>

CUB 200: <https://pan.baidu.com/s/1OL3s7XmzoaYmbBocYJjYyQ> code: 7jeb

Oxford-102: <https://s3.amazonaws.com/fast-ai-imageclas/oxford-102-flowers.tgz>

Fashion Cap.: https://github.com/xuewyang/Fashion_Captioning

GoodNews: <https://github.com/furkanbiten/GoodNews>

Text Caps: <https://textvqa.org/textcaps/>

Localized Narratives: <https://google.github.io/localized-narratives/>

			Inputs		
			Pred	Refs	Image
Standard	BLEU	Translation	✓	✓	
	METEOR	Translation	✓	✓	
	ROUGE	Summarization	✓	✓	
	CIDEr	Captioning	✓	✓	
	SPICE	Captioning	✓	(✓)	(✓)
Diversity	Div	Captioning	✓		
	Vocab	Captioning	✓		
	%Novel	Captioning	✓		
Embedding-based	WMD	Doc. Dissimilarity	✓	✓	
	Alignment	Captioning	✓	✓	
	Coverage	Captioning	✓	(✓)	(✓)
Learning-based	TIGer	Captioning	✓	✓	✓
	BERT-S	Text Similarity	✓	✓	
	CLIP-S	Captioning	✓	(✓)	✓

表2 图像caption指标的分类和主要特征



$$BLEU = BP \cdot \exp\left(\sum_{n=1}^N w_n \log P_n\right)$$

$$BP = \begin{cases} 1 & c > r \\ e^{1-r/c} & c \leq r \end{cases}$$

$$P_n = \frac{\sum_{i \in n\text{-gram}} \min(h_i(C), \max_{j \in m} h_i(S_j))}{\sum_{i \in n\text{-gram}} h_i(C)}$$

i : C中的n-gram

$h_i(C)$: $n\text{-gram } i$ 在C中出现的次数

$h_i(S_j)$: $n\text{-gram } i$ 在第j个参考caption出现的次数

优点：易于计算且速度快,应用范围广泛，可以很轻松将模型与相同任务的基准作对比。

缺点：没有考虑语义、句子结构，不能很好地处理形态丰富的语句。BLEU 指标偏向于较短的翻译结果（BP效果不好）

- BLEU 的核心思想是比较候选caption和参考caption里的n-gram的重合程度，重合程度越高就认为译文质量越高。
- 1-gram用于衡量单词翻译的准确性，高阶n-gram用于衡量句子翻译的流畅性。实践中，通常是取N=1~4，然后对进行加权平均。公式中的P_n指的是n-gram的精确率。
- W_n指n-gram的权重，一般设置为均匀权重，即对任意n都有W_n=1/N。
- BP是过短惩罚因子，c是模型输出的候选句子的长度，r是最短参考译文的长度。如果译文的长度小于最短的参考译文，则 BP 小于 1。



$$R_{LCS} = \frac{LCS(C, S)}{len(S)}$$

$$P_{LCS} = \frac{LCS(C, S)}{len(C)}$$

$$F_{LCS} = \frac{(1 + \beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}}$$

- BLEU 的改进版，专注于召回率而非精度。
- ROUGE大致分为四种： - ROUGE-N （将BLEU的精确率优化为召回率） - ROUGE-L （将BLEU的n-gram优化为公共子序列） - ROUGE-W （将ROUGE-L的连续匹配给予更高的奖励） - ROUGE-S （允许n-gram出现跳词(skip)）
- ROUGE-L 中的 L 指最长公共子序列 (LCS)，ROUGE-L 计算的时候使用了机器译文C和参考译文S的最长公共子序列。

R_{LCS} : 召回率 P_{LCS} : 精确率 F_{LCS} : ROUGE-L



$$METEOR = (1 - pen) \times F_{means}$$

$$F_{means} = \frac{PR}{\alpha P + (1 - \alpha)R}$$

$$P = \frac{m}{c}$$

$$R = \frac{m}{r}$$

$$Pen = \frac{\#chunks}{m}$$

- α 为可控的参数
- m 为候选caption中能够被匹配的一元组(unigram)的数量
- c 为候选caption的长度
- r 为参考caption的长度
- Pen 为惩罚因子，惩罚的是候选翻译中的词序与参考翻译中的词序不同

- METEOR 是一种基于精度和召回率的机器翻译评估指标。使用 WordNet 计算特定的序列匹配，同义词，词根和词缀，释义之间的匹配关系，改善了BLEU的效果，使其跟人工判别共更强的相关性。

- #chunks指的是chunk的数量，chunk就是既在候选caption中相邻又在参考caption中相邻的被匹配的一元组聚集而成的单位，例：

Candidate: the president spoke to the audience.

Reference: the president then spoke to the audience.

在上述例子中，reference中的6个unigram都可以被匹配，但是其在reference中匹配的对象，却只有"the president" 和 "spoke to the audience"这两个字符串是相邻的，而这两个字符串就是两个chunk。

$$CIDEr_n(c_i, s_i) = \frac{1}{M} * \sum_{j=1}^M \frac{g^n(c_i) * g^n(s_{ij})}{\|g^n(c_i)\| * \|g^n(s_{ij})\|}$$

$$g_k(c_i) = TF(k) * IDF(k)$$

k: 句子中第k个n-gram

c_i : 第i个图片对应的候选caption

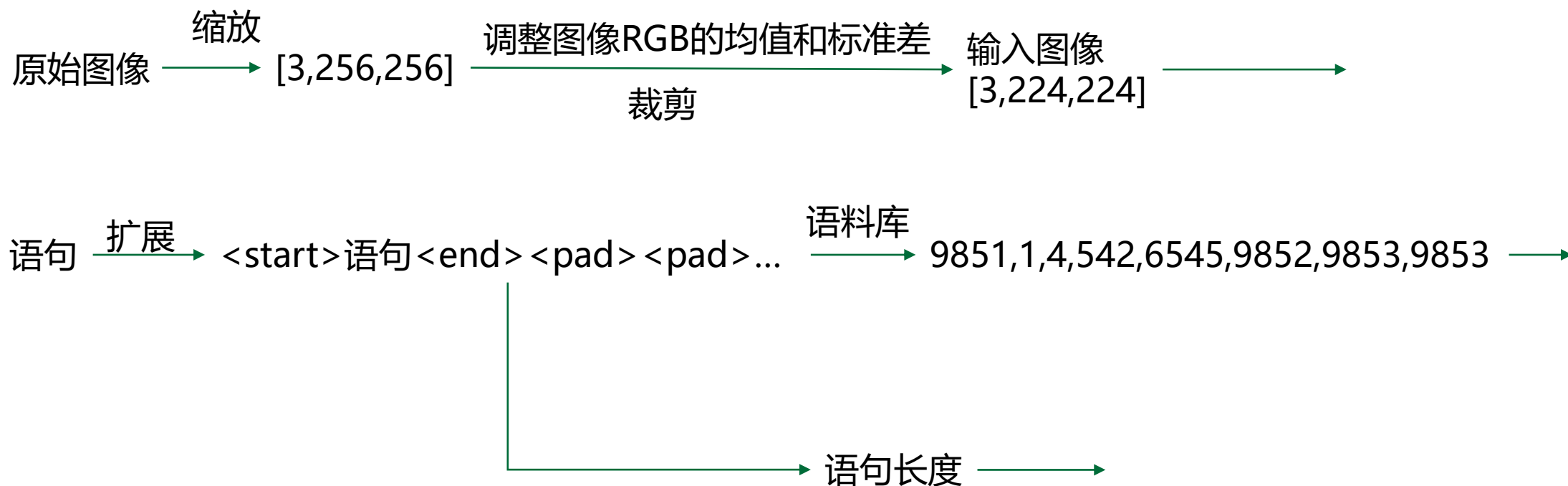
s_{ij} : 第i个图片对应的参考caption(多个)中的第j个caption

M: 每个图片对应的参考caption数目

- CIDEr将每个句子都看作“文档”，将其表示成TF-IDF向量的形式，计算参考 caption 与候选的 caption 的余弦相似度，来衡量图像标注的一致性的。
- 优点：是一种加权的评价指标，更关注caption是否说到了重点，而常见的词权重则没有那么高。这个评价指标和人类的评价相关性更高一些。



Demo 展示



Flickr30K Dataset + Andrej Karpathy's training, validation, and test splits

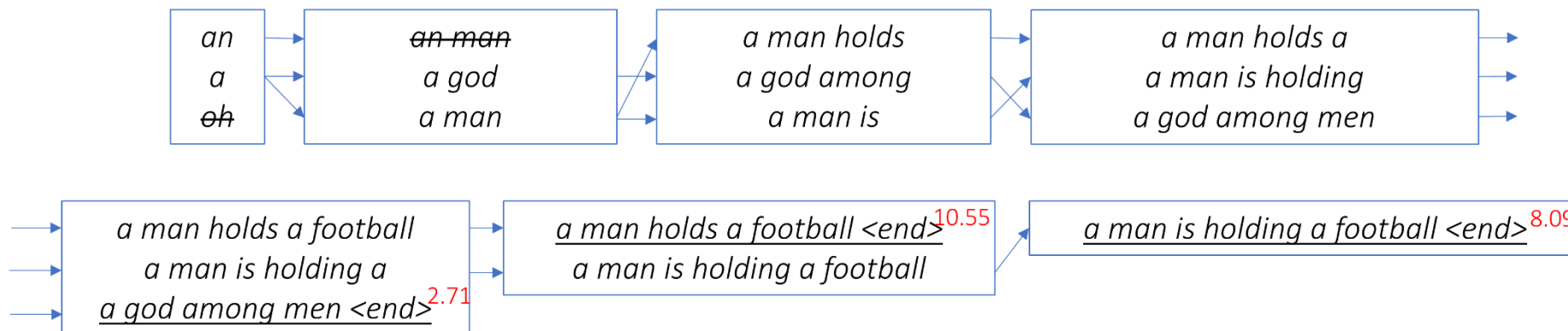


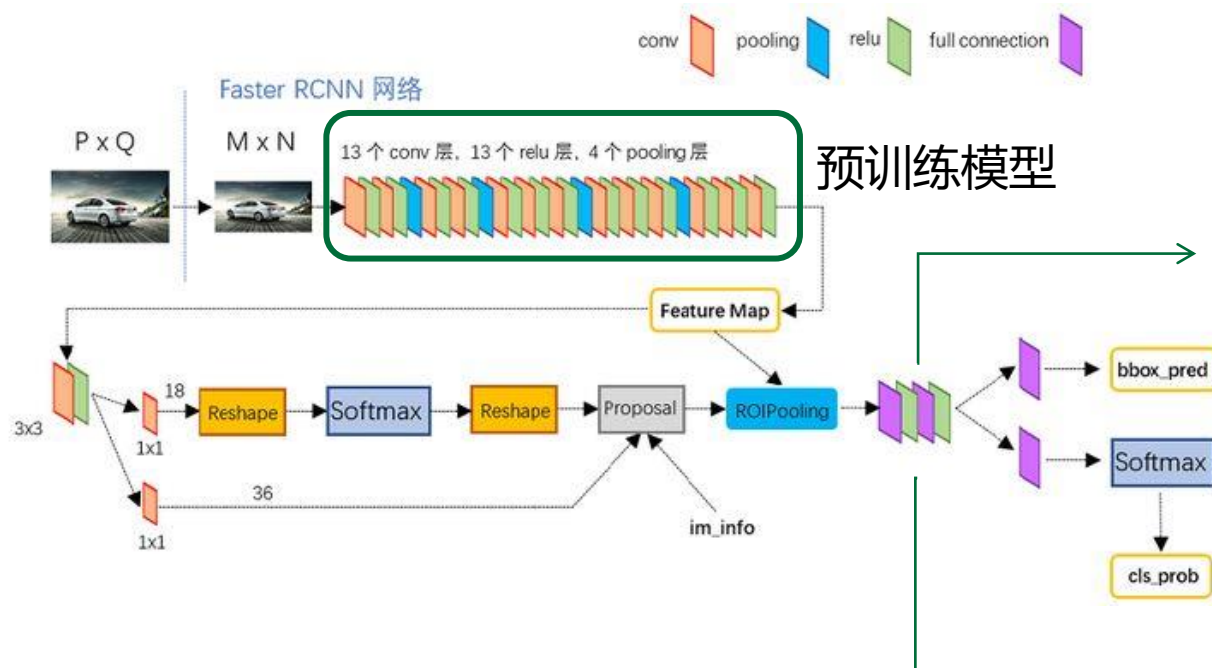
Beam Search with $k = 3$

Choose top 3 sequences at each decode step.

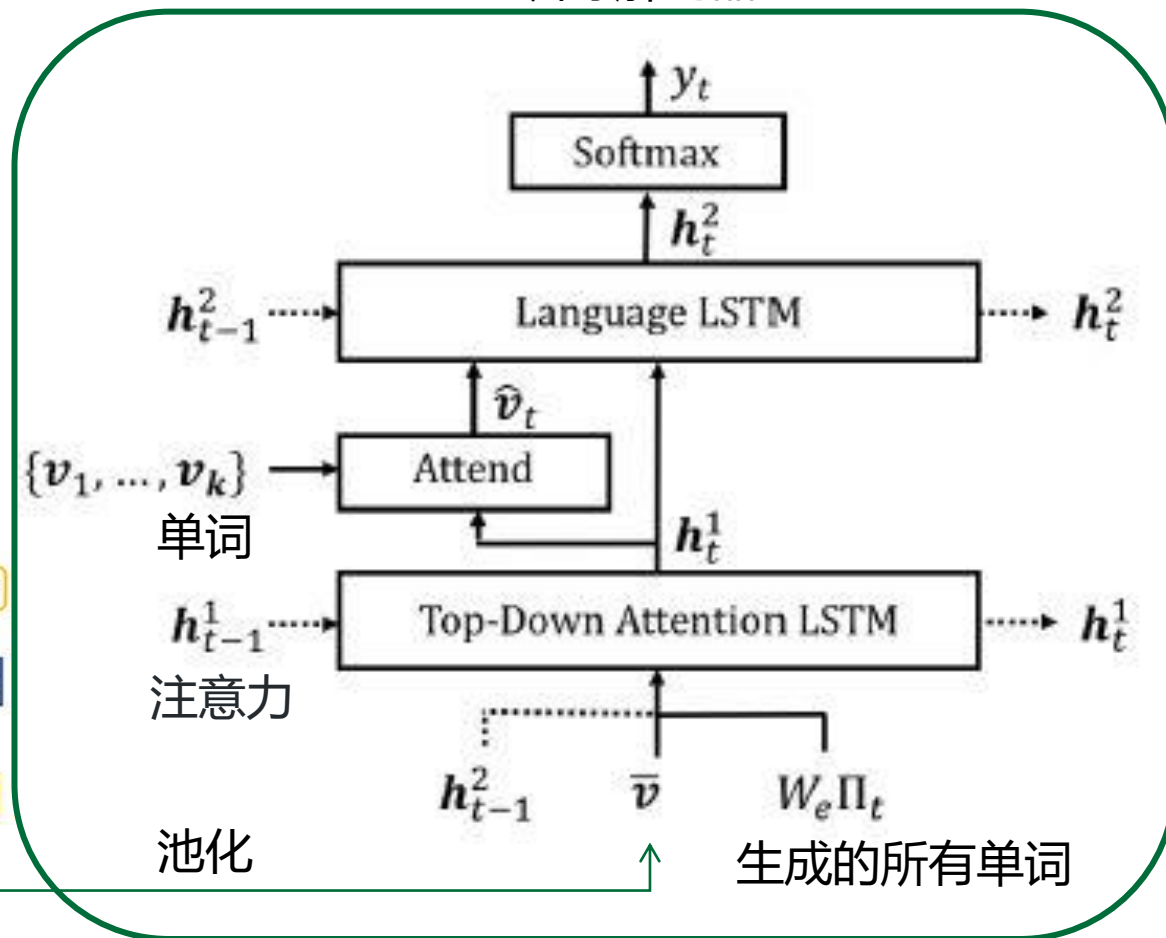
Some sequences fail early.

Choose the sequence with the highest score after all 3 chains complete.

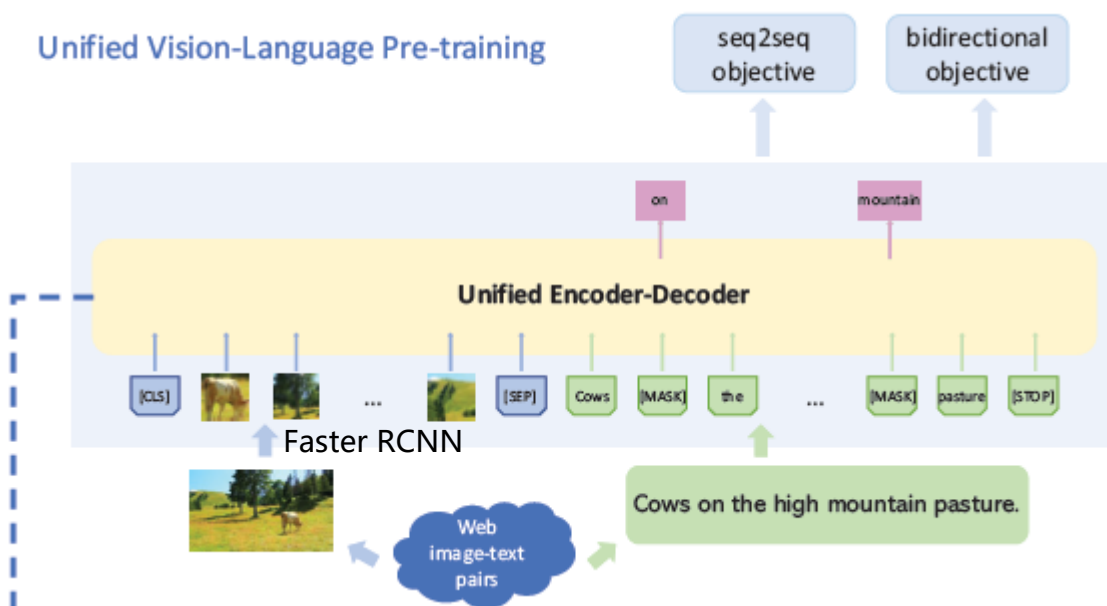




训练解码器

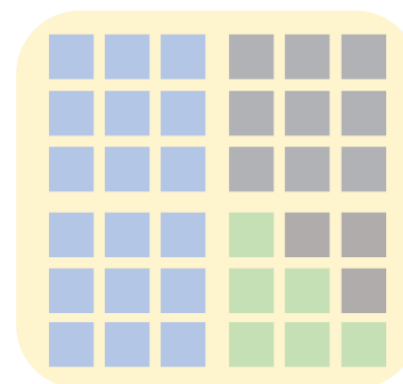


训练



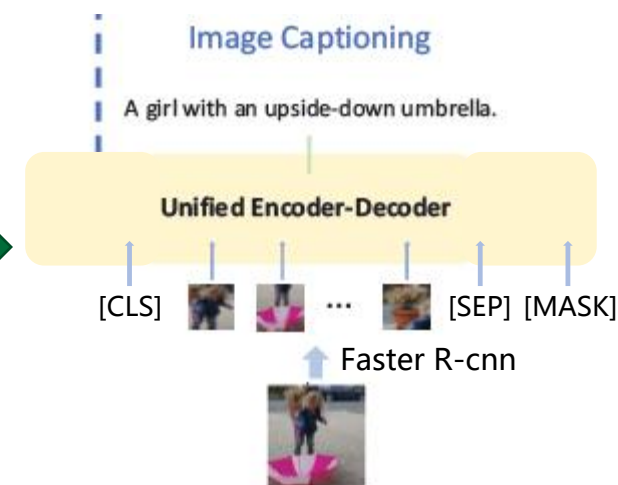
Seq-to-Seq

微调



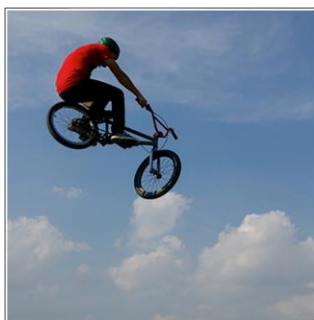
Self-attention mask

推理



Zhou L, Palangi H, Zhang L, et al. Unified vision-language pre-training for image captioning and vqa[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 13041-13049.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Up-Down	68.8%	50.5%	36.4%	26.1%
VLP	71.6%	54.5%	40.7%	30.1%



BUTD: A man in a red shirt is doing a trick on his bike .

VLP: A man in a red shirt is doing a trick on his bike .



BUTD: A group of dogs racing on a track .

VLP: Greyhounds race around a track .

BUTD: A man and a woman pose for a picture .

VLP: A man and a woman are smiling .



BUTD: A group of people are sitting around a table .

VLP: A group of people are sitting at a table in a restaurant .





北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

谢谢观看

答辩人：蔡永奇、陈盈盈、郭镇远

马昊哲、史凯阅、张津铭

时 间：2021-10-11