课程设计
图像描述生成

基于编解码框架的方法

大纲

- 任务介绍
- 数据集
- 评测指标
- 主流模型
- 实战
- 作业

任务介绍

任务介绍

- 自动为图片生成流畅关联的自然语言描述

图像：

图像描述模型

文本描述：

穿着紫色衣服的小孩和穿着红色衣服的小孩正坐在野餐垫上吃零食

数据集

数据集

- Pascal Sentence
  - 25%的描述没有动词，15%的描述包含如sit、stand、wear和look这类的静态动词
- Flickr8k, Flickr30k
  - 21%的描述没有动词或包含静态动词
- Microsoft COCO
  - 使用最广泛的图像描述生成数据集
- AIC-ICC
  - 发布于2017 AI Challenger
  - 210,000幅图像组成的训练集，30,000幅图像组成的验证集，30,000幅图像组成的测试集A和30,000幅图像组成的测试集B
  - 每幅图像对应5个人工标注的***中文句子***描述

评测指标

BLEU (BiLingual Evaluation Understudy)

- n-gram层面的准确率

$$\mathrm{BLEU}_n(a,b) = \frac{\sum_{w_n \in a} \min\left(c_a(w_n), \max_{j=1,\ldots,|b|} c_{b_j}(w_n)\right)}{\sum_{w_n \in a} c_a(w_n)}$$

- a：候选句子（生成句子）
- b：参考句子集（真实句子集、标注句子集）
- w_n：n-gram
- $c_x(y_n)$：n-gram y_n在句子x中出现的次数

BLEU: A Method for Automatic Evaluation of Machine Translation. 2002

BLEU

- n-gram层面的准确率

$$\mathrm{BLEU}_n(a,b) = \frac{\sum_{w_n \in a} \min\left(c_a(w_n), \max_{j=1,\ldots,|b|} c_{b_j}(w_n)\right)}{\sum_{w_n \in a} c_a(w_n)}$$

- 例子
  - 参考句子1: The cat is playing on the desk.
  - 参考句子2: The cat is on the table.
  - 候选句子1: The cat is playing on the table.
  - 候选句子2: the the the the the the the.
  - 候选句子3: the

BLEU

- 简单计算召回率的问题
  - 例子
    - 参考句子1: I always do.
    - 参考句子2: I invariably do.
    - 参考句子3: I perpetually do.
    - 候选句子1: I always invariably perpetually do.
    - 候选句子2: I always do.
  - 候选句子1的召回率比候选句子2要大，但是并不合理

BLEU

- BLEU值

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

  - p_n就是$BLEU_n$

  - BP为短句惩罚项，根据候选句子的长度c，选择一个最相近的参考句子的长度r

$$BP = \begin{cases} 1 & \text{if c>r} \\ e^{(1-r/c)} & \text{else} \end{cases}$$

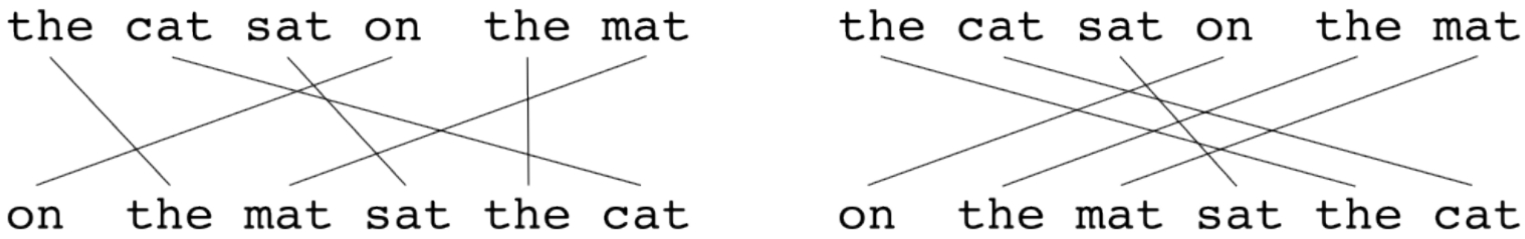
BLEU

- 一般计算BLEU-n，n取1到4

- 优点：容易计算
- 缺点
  - 没有考虑n-gram的顺序
  - 平等对待所有的n-gram
  - 衡量的是句子之间的流畅性而非语义相似度

METEOR (Metric for Evaluation of Translation with Explicit ORdering)

- 考虑同义词的F值，鼓励连续词匹配
- 计算方法：
- 1. 在候选句子和参考句子之间做词到词的映射
  - 首先列出所有可能的匹配
    - 匹配原则：完全匹配、词根匹配、同义词匹配（WordNet）
  - 然后在所有可能的匹配中选择一个匹配成功词最多的，如果两个匹配成功的词一样多，就选择其中交叉最少的那个



METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. 2005

METEOR

- 2. 计算METEOR值，对于多个参考句子，取得分最高的作为最终结果

$$\text{METEOR} = \max_{j=1,\ldots,|b|} \left(\frac{10PR}{R+9P} \right) \left(1 - \frac{1}{2} \left(\frac{\#\text{chunks}}{\#\text{matched unigrams}} \right)^3 \right)$$

  - P为一元组的准确率，R为一元组的召回率
  - $\#chunks$指的是chunk的数量，chunk就是既在候选句子中相邻又在参考句子中相邻的被匹配的一元组聚集而成的单位
  - chunk例子
    - 候选句子: the president spoke to the audience
    - 参考句子: the president then spoke to the audience

METEOR

- 优点
  - 词到词的映射方式，考虑了词的语义和位置因素
  - 引入了chunk计数来进行任意长度的n-gram匹配，在句子结构上衡量了两个句子的相似程度
  - 使用chunk的数量来确定的n-gram匹配，无需指定n的具体值
  - 通过候选句子和参考句子的一对一匹配规避了多参考句子下召回率计算的问题，从而可以计算召回率
- 缺点
  - 词匹配在实际运行的时候偏慢

ROUGE (Recall Oriented Understudy of Gisting Evaluation)

- n-gram层面的召回率

$$\mathrm{ROUGE}_n(a,b) = \frac{\sum_{j=1}^{|b|} \sum_{w_n \in b_j} \min\left(c_a(w_n), c_{b_j}(w_n)\right)}{\sum_{j=1}^{|b|} \sum_{w_n \in b_j} c_{b_j}(w_n)}$$

- 例子
  - 参考句子: The cat is on the table.
  - 候选句子: The cat is playing on the table.

ROUGE: A Package for Automatic Evaluation of Summaries. 2004

ROUGE-L (F值)

- 计算方法

$$ROUGE-L=\frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}$$

$$R_{lcs}=\frac{LCS(X,Y)}{m}$$

$$P_{lcs}=\frac{LCS(X,Y)}{n}$$

  - 其中，LCS为最长公共子序列，$\beta = P_{lcs}/R_{lcs}$

- 多个参考句子时，单独计算所有的参考句子的ROUGE-L值，取最大的一个作为最终的结果
- 优点：使用的是最长公共子列，无需n-gram的完全匹配，且无需预先定义匹配的n-gram的长度；考虑了词序
- 缺点：仅仅参考了最长的公共子列，候选句子和参考句子中的其他相同的部分都被省略掉了

CIDEr

- BLEU+向量空间模型

$$CIDEr_n(a,b) = \frac{1}{|b|}\sum_{j=1}^{|b|}\frac{\mathbf{g^n}(a)\cdot\mathbf{g^n}(b_j)}{\left\|\mathbf{g^n}(a)\right\|\left\|\mathbf{g^n}(b_j)\right\|}$$

其中，$g^n(x)$为句子x的n-gram形式的TF-IDF表示，对所有词预先执行stem操作，变成词根形式

- CIDEr最终也是1到4 gram的结果的平均值

CIDEr: Consensus-based Image Description Evaluation. 2014

CIDEr

- 优点
  - CIDEr引入了TF-IDF为n-gram进行加权，这样就避免评价候选句子时因为一些常见却不够有信息量的n-gram打上高分
    - 参考句子: The book **is on the top** of the desk.
    - 候选句子: The cat **is on the top** of the shelf.

- 缺点
  - CIDEr取词根的操作会让一些动词的原型和名词匹配成功
  - 高置信度的词重复出现的长句的CIDEr得分也很高

CIDEr: Consensus-based Image Description Evaluation. 2014

CIDEr-D

- 对于动词原形和名词匹配成功的问题，CIDEr-D不再取词根
- 对于包括高置信度的词的长句，CIDEr-D增加了惩罚生成句子和参考句子的长度差别的权重，并且通过对n-gram计数的截断操作不再计算生成句子中出现次数超过参考句子的n-gram

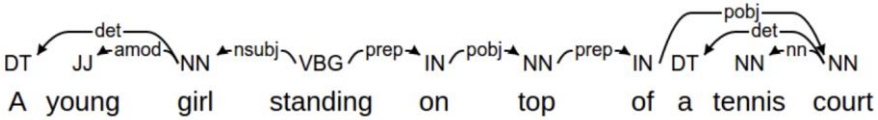
$$\text{CIDEr-D}_n(c_i, S_i) = \frac{10}{m} \sum_j e^{\frac{-(l(c_i)-l(s_{ij}))^2}{2\sigma^2}} \times \frac{\min(\mathbf{g^n}(c_i), \mathbf{g^n}(s_{ij})) \cdot \mathbf{g^n}(s_{ij})}{\|\mathbf{g^n}(c_i)\| \|\mathbf{g^n}(s_{ij})\|}$$

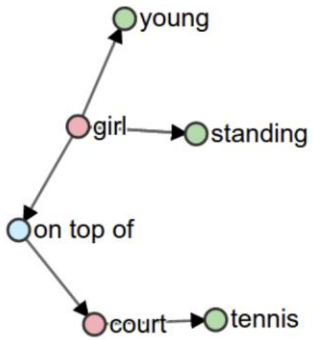
- CIDEr-D也是计算1到4 gram的结果的平均值

CIDEr: Consensus-based Image Description Evaluation. 2014

SPICE

- 以名词为中心的度量
- 候选句子和参考句子集的场景图相似度





- 场景图的物体、属性和关系都会分别转化为一元、二元、三元组组成的集合
  - {(girl), (court), (girl, young), (girl, standing), (court, tennis), (girl, on-top-of, court)}

SPICE: Semantic Propositional Image Caption Evaluation. 2016

SPICE

- 计算步骤
  - 利用Stanford Scene Graph Parser将候选句子和参考句子集转化为场景图
  - 比较候选句子和参考句子集中元组的precision、recall，最终计算出F1 score
- 优点
  - 在语义而非n-gram层级度量
  - 每个句子映射到场景图后可以从中提取出模型关于某些关系或者属性的识别能力
- 缺点
  - 缺少n-gram来度量句子的流畅性
  - 度量的准确性受到场景图解析器的制约

主流模型

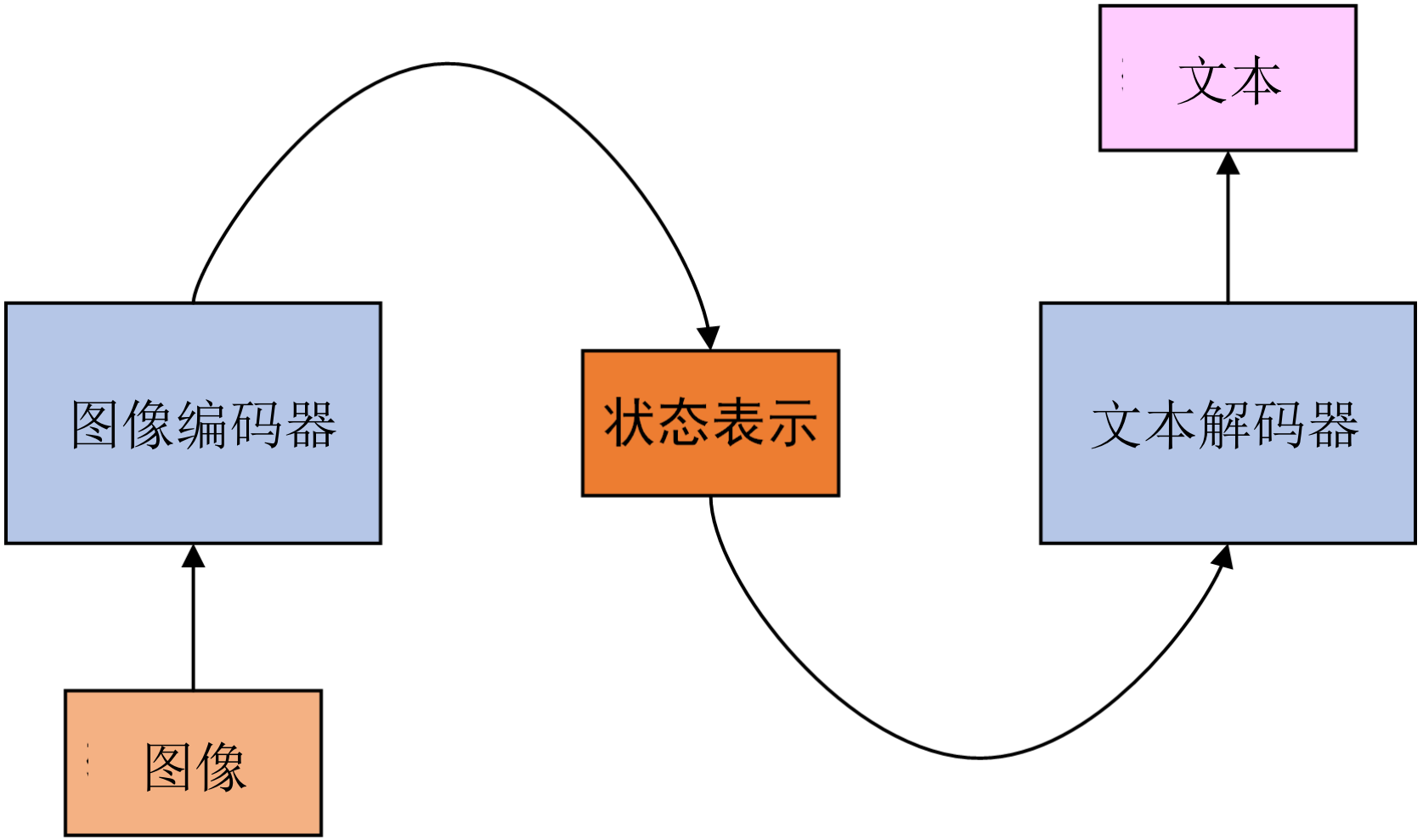
传统的基于管道式结构的方法

- 1. 使用计算机视觉技术来对场景进行分类，检测图像中存在的对象，预测它们的属性以及它们之间的关系，识别发生的动作，将它们映射为一些基本的自然语言描述单元，例如单词、短语或其他结构化描述。
- 2.通过自然语言生成技术（例如，模板，n-gram，语法规则等）将这些单词或者短语进行组合，生成自然语言描述句子。

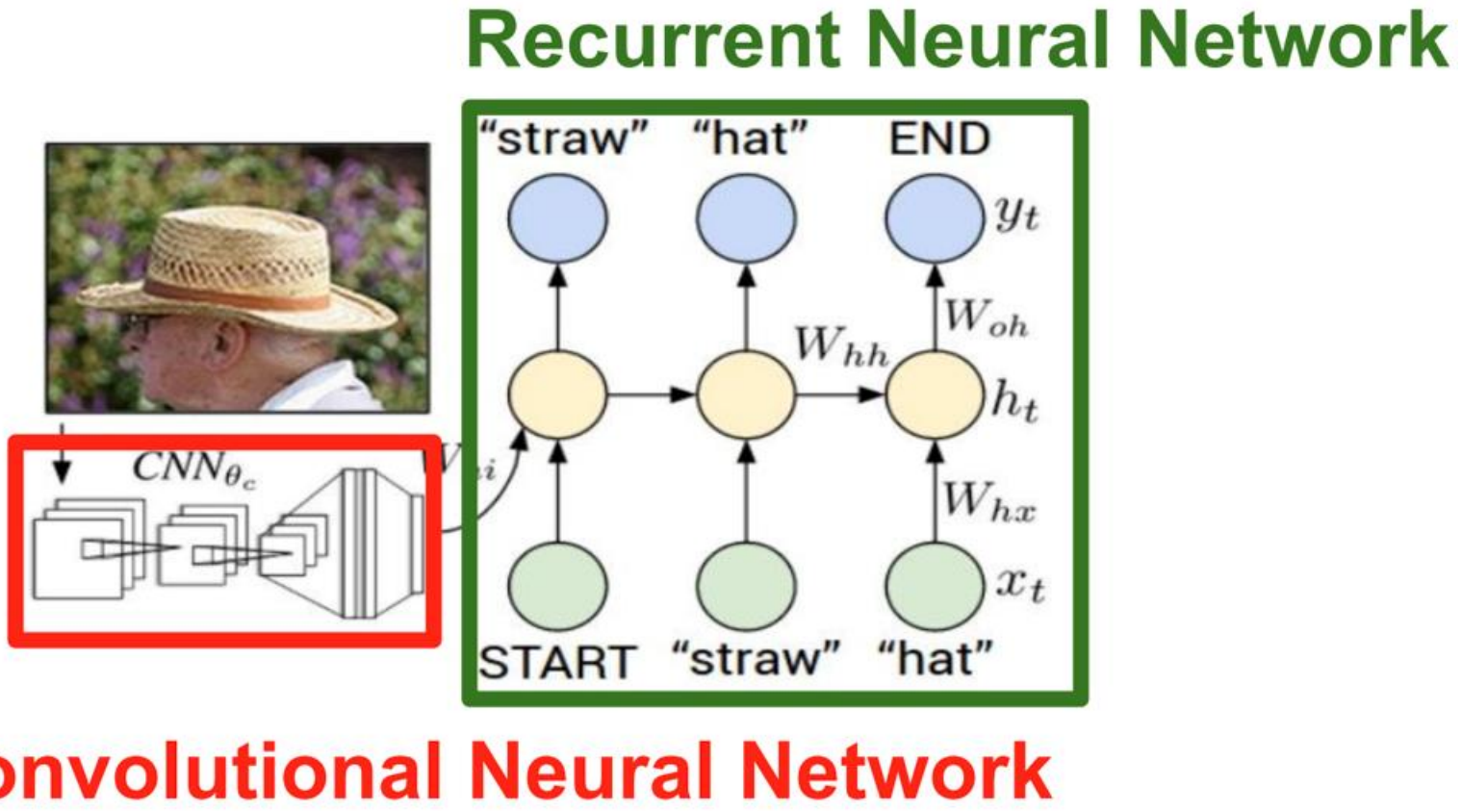
传统的基于管道式结构的方法：问题

- 分阶段的方式限制了两个模态数据间的信息交互
- 高度依赖于预先定义的场景、对象、属性和动作的封闭语义类集
- 分阶段的模型存在误差累积问题，前面任务的误差在后面阶段会放大
- 训练误差不能前向传递

基于编解码框架的方法



编解码框架：一个例子



Deep Visual-Semantic Alignments for Generating Image Descriptions. 2015

编解码框架：一个例子

test image

编解码框架：一个例子

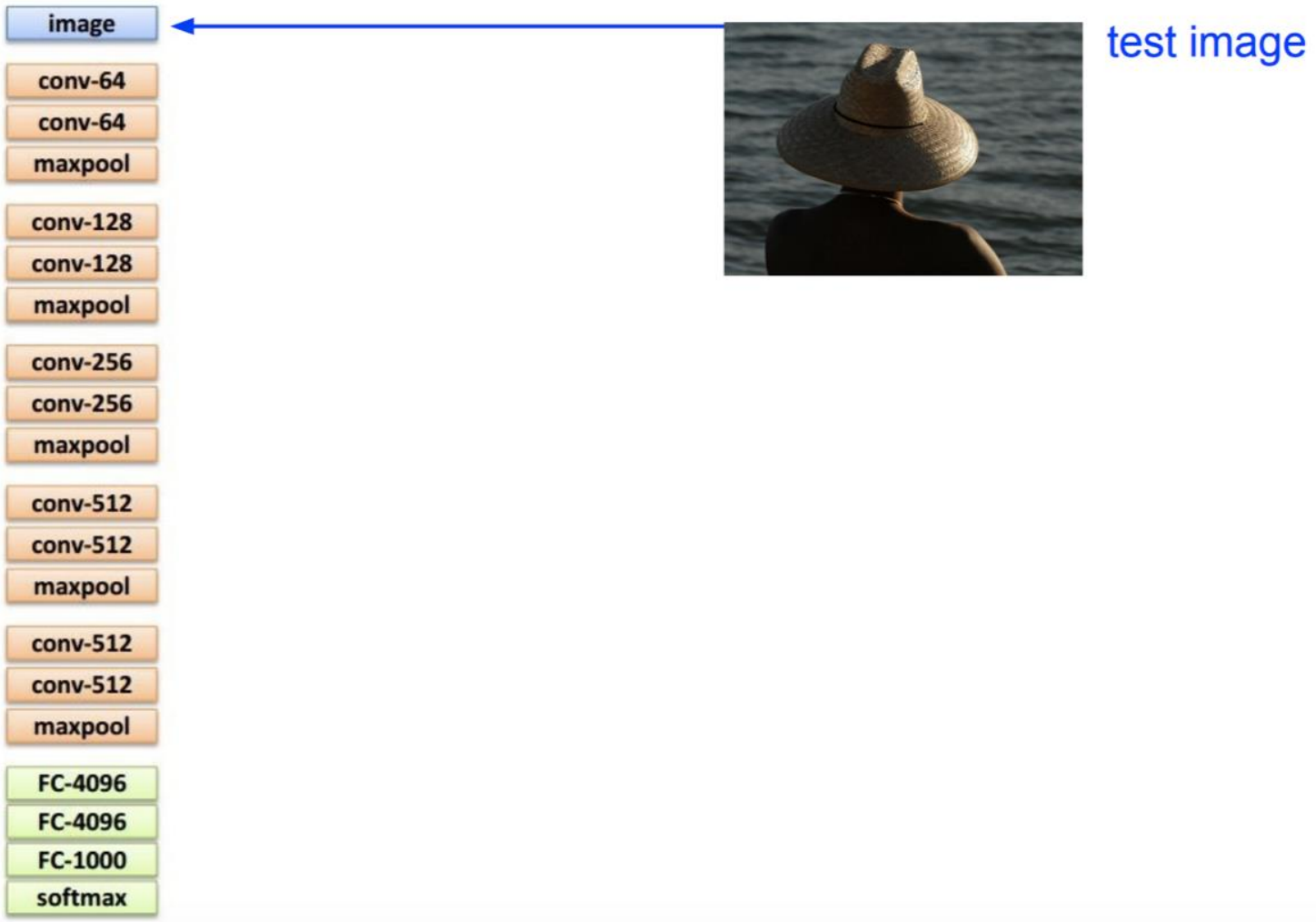


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
FC-1000
softmax
test image

编解码框架：一个例子

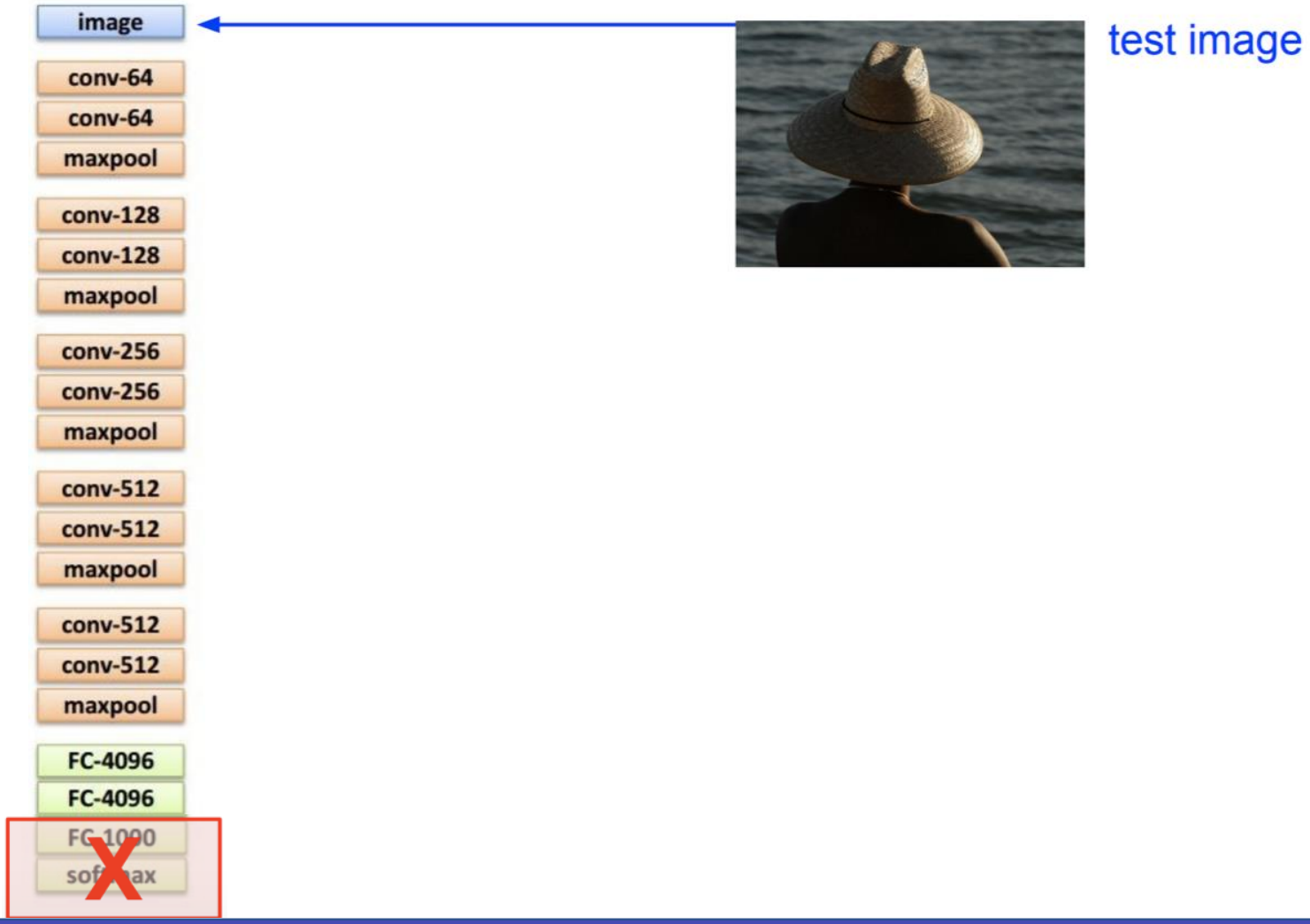


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
FC-1000
softmax
test image

编解码框架：一个例子

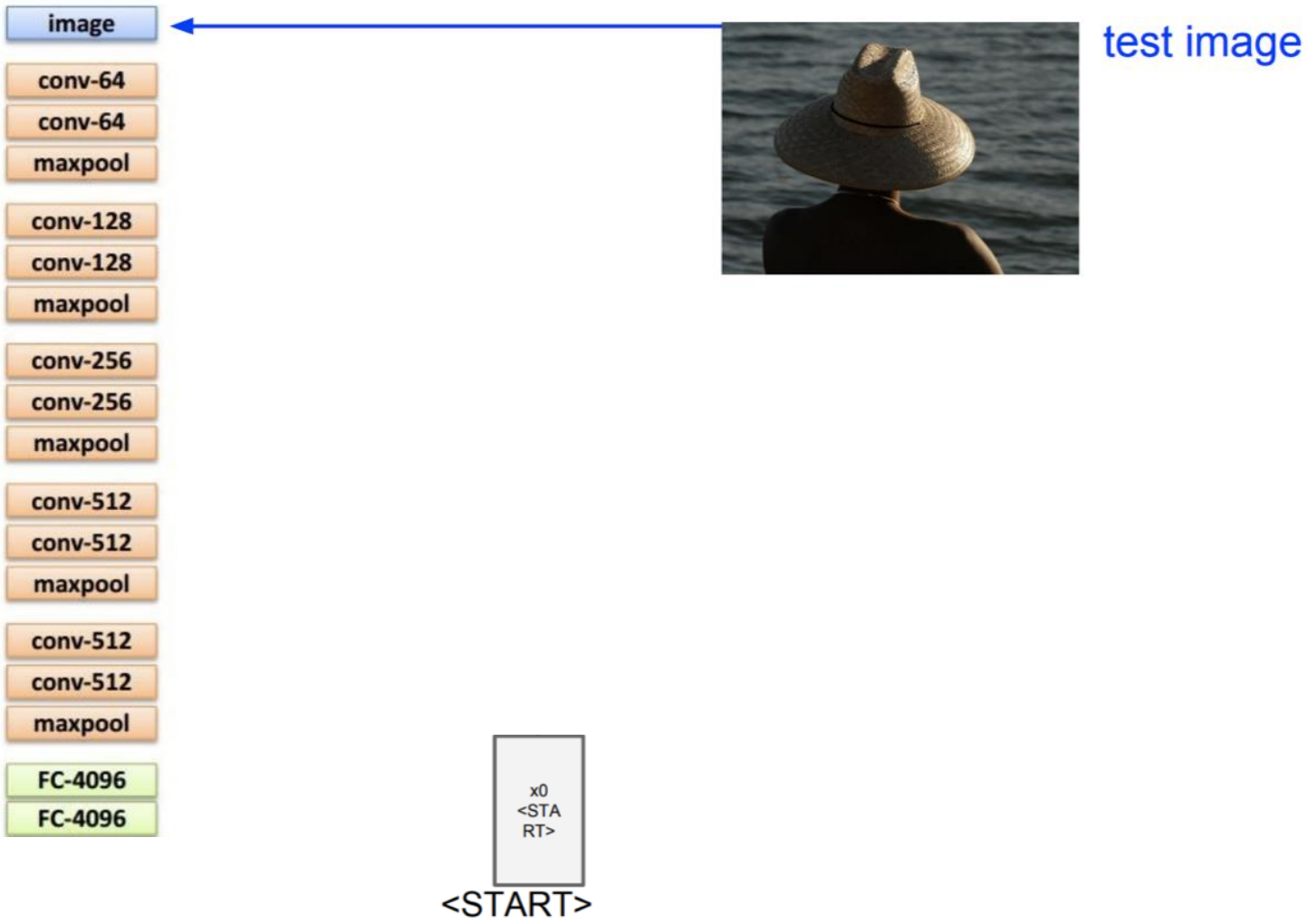
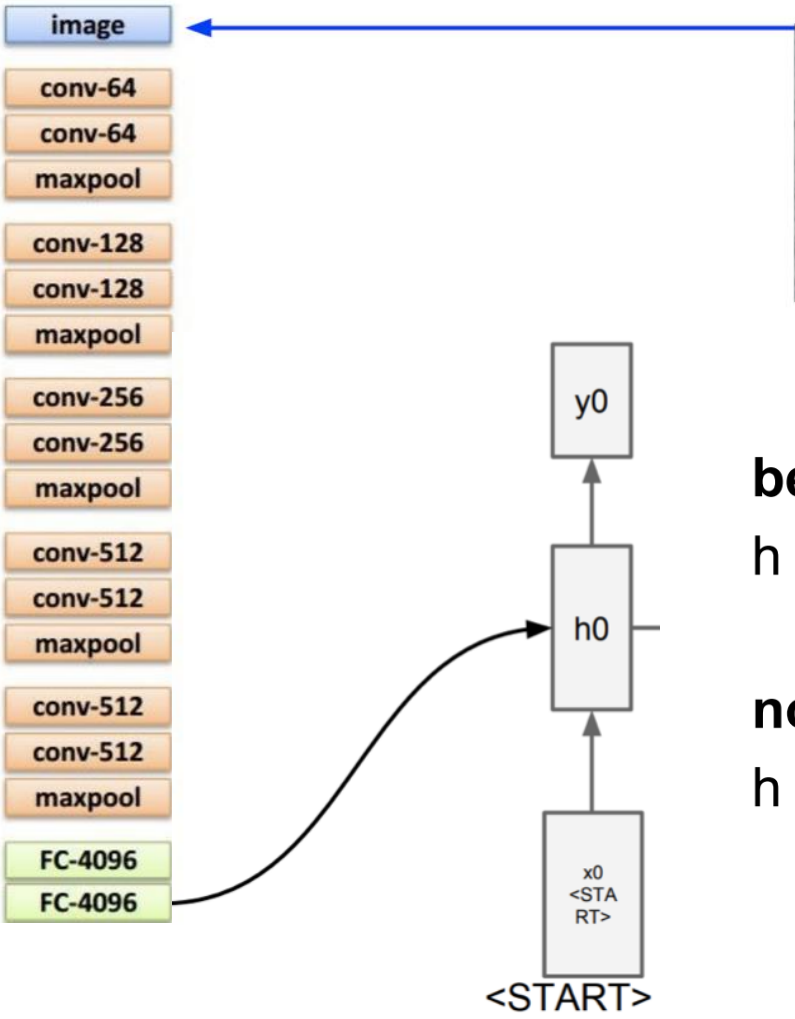


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
x0
<STA
RT>
<START>

编解码框架：一个例子



test image

before:

$h = \tanh(W_{xh} * x + W_{hh} * h)$

now:

$h = \tanh(W_{xh} * x + W_{hh} * h + W_{ih} * v)$

编解码框架：一个例子

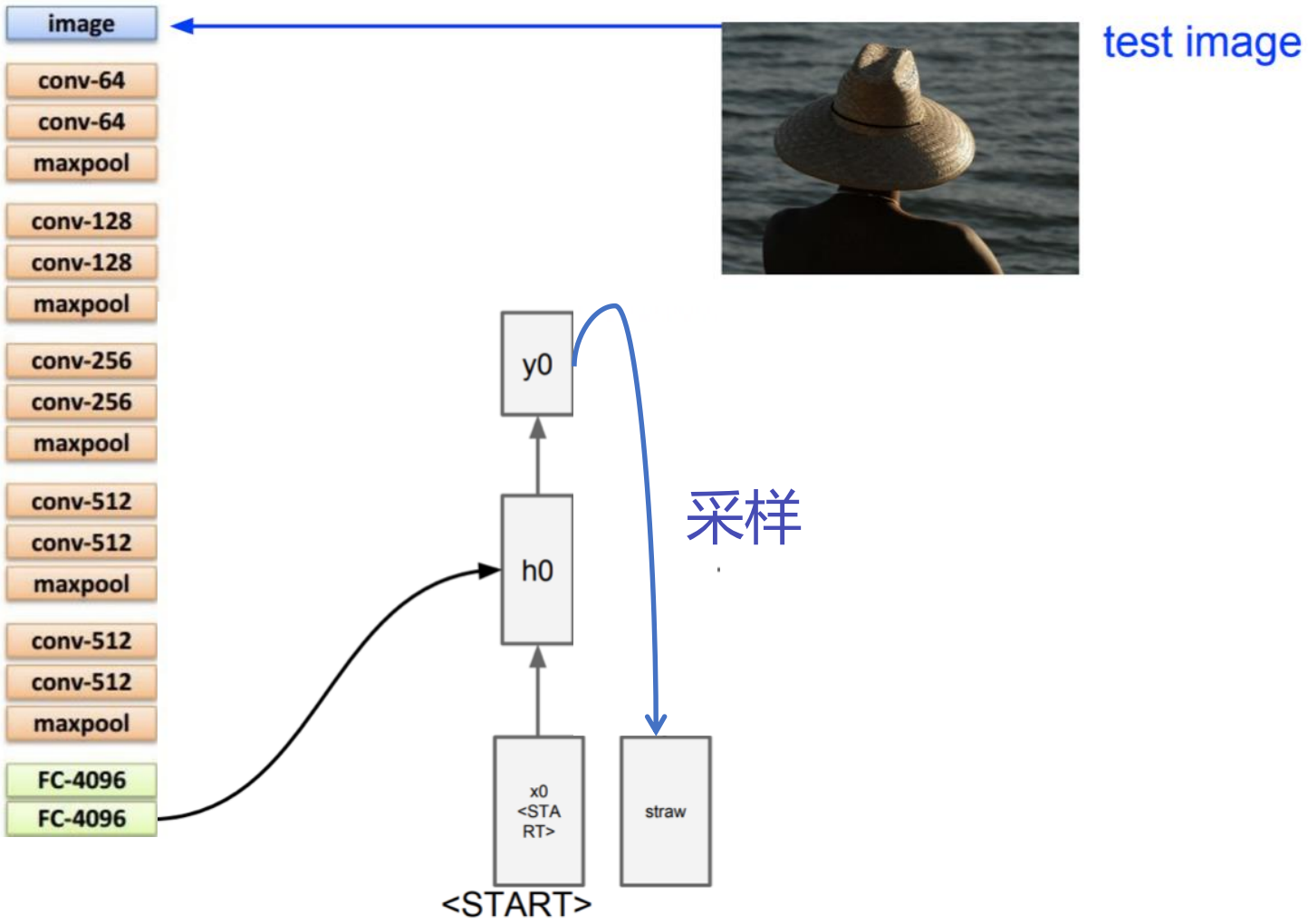


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y0
h0
采样
x0
<STA
RT>
straw
<START>

编解码框架：一个例子

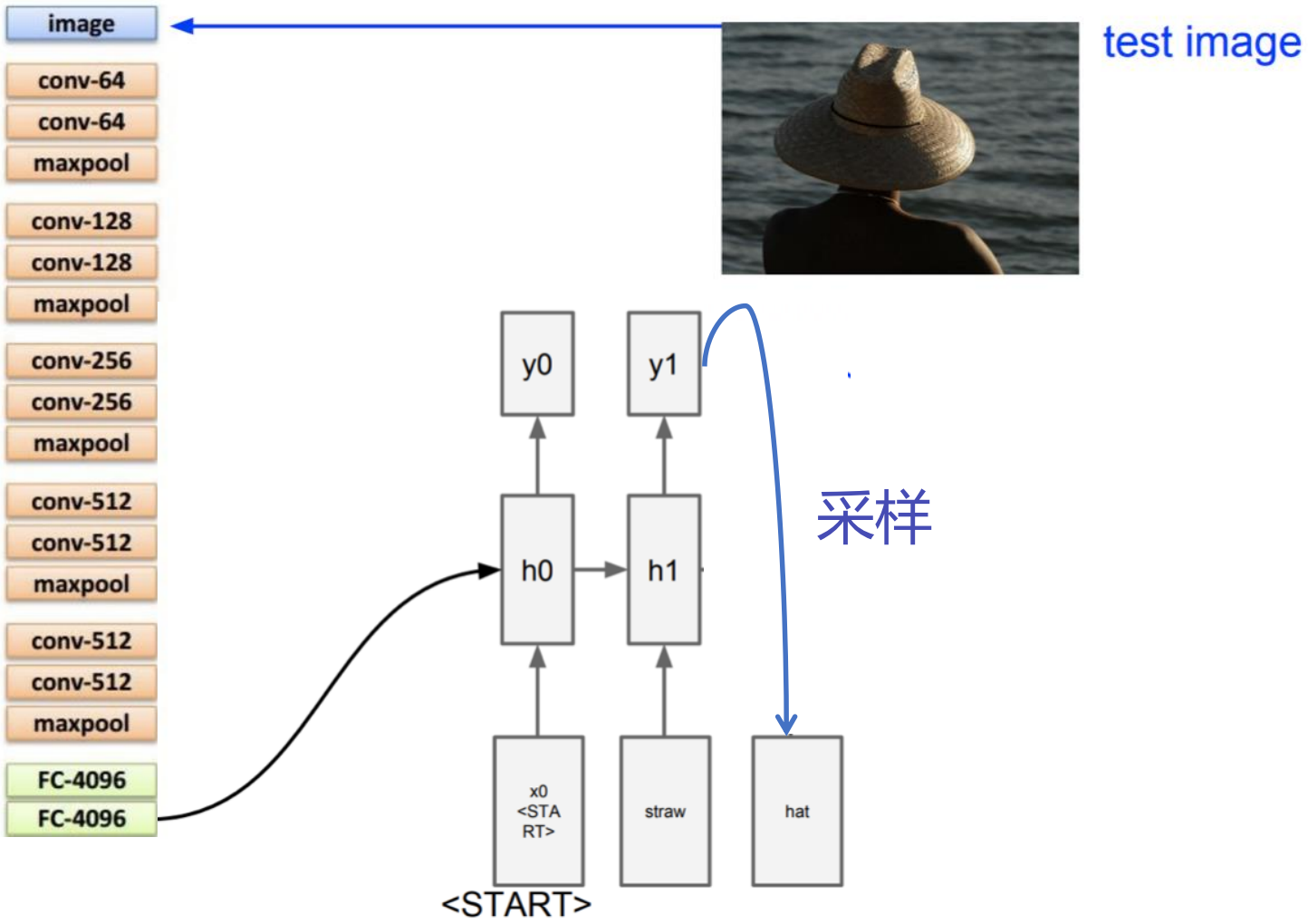


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y0
y1
采样
h0
h1
x0
<STA
RT>
straw
hat
<START>

编解码框架：一个例子

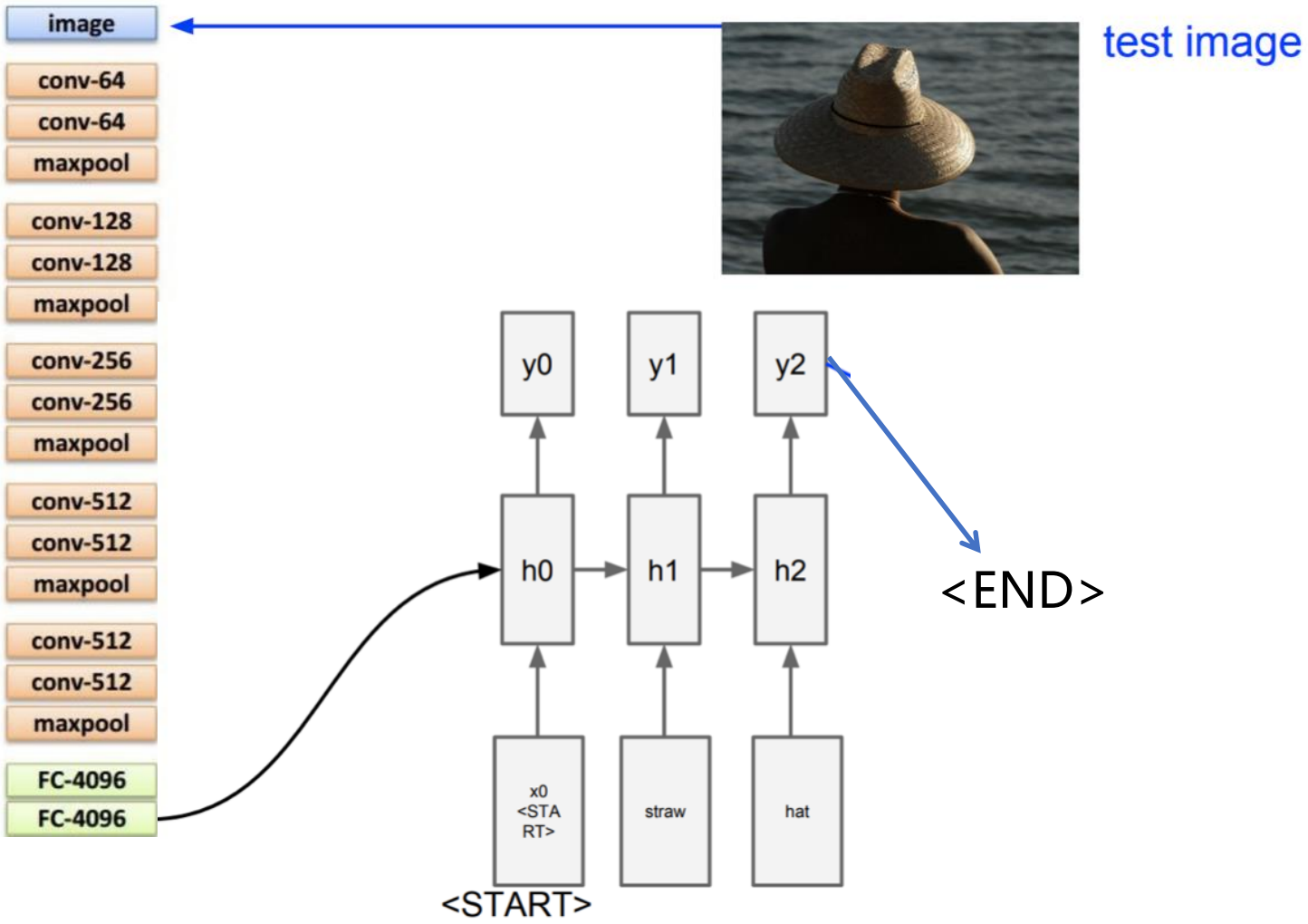


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y0
y1
y2
h0
h1
h2
<END>
x0
<START>
straw
hat
<START>

编码器+解码器

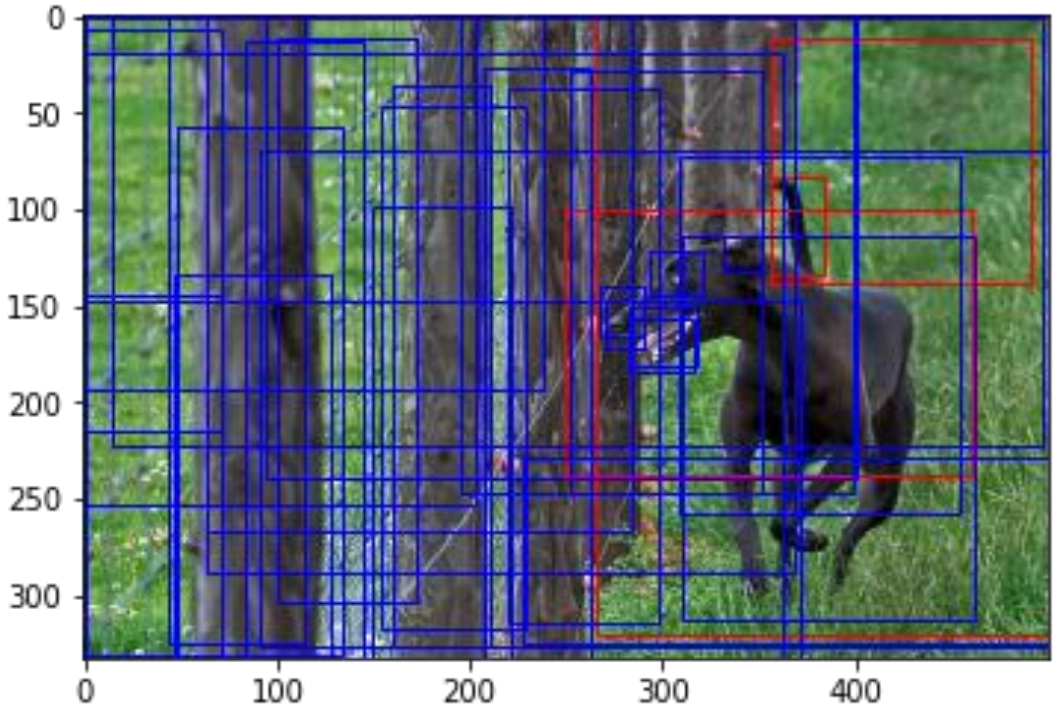
- 整体表示+RNN
- 局部表示+注意力
- 局部表示、自注意力+注意力
- 局部表示、图网络+注意力
- 局部表示、Transformer编码器+Transformer解码器
- 视觉Transformer+Transformer解码器

图像编码器

- 整体表示编码器
  - CNN的倒数第二层（CNN整体表示）
  - 视觉Transformer的<CLS>对应的表示（ViT整体表示）

- 局部表示编码器
  - CNN的最后一个卷积层（网格表示）
  - 目标检测区域表示（区域表示）
  - 视觉Transformer提取的块表示（块表示）

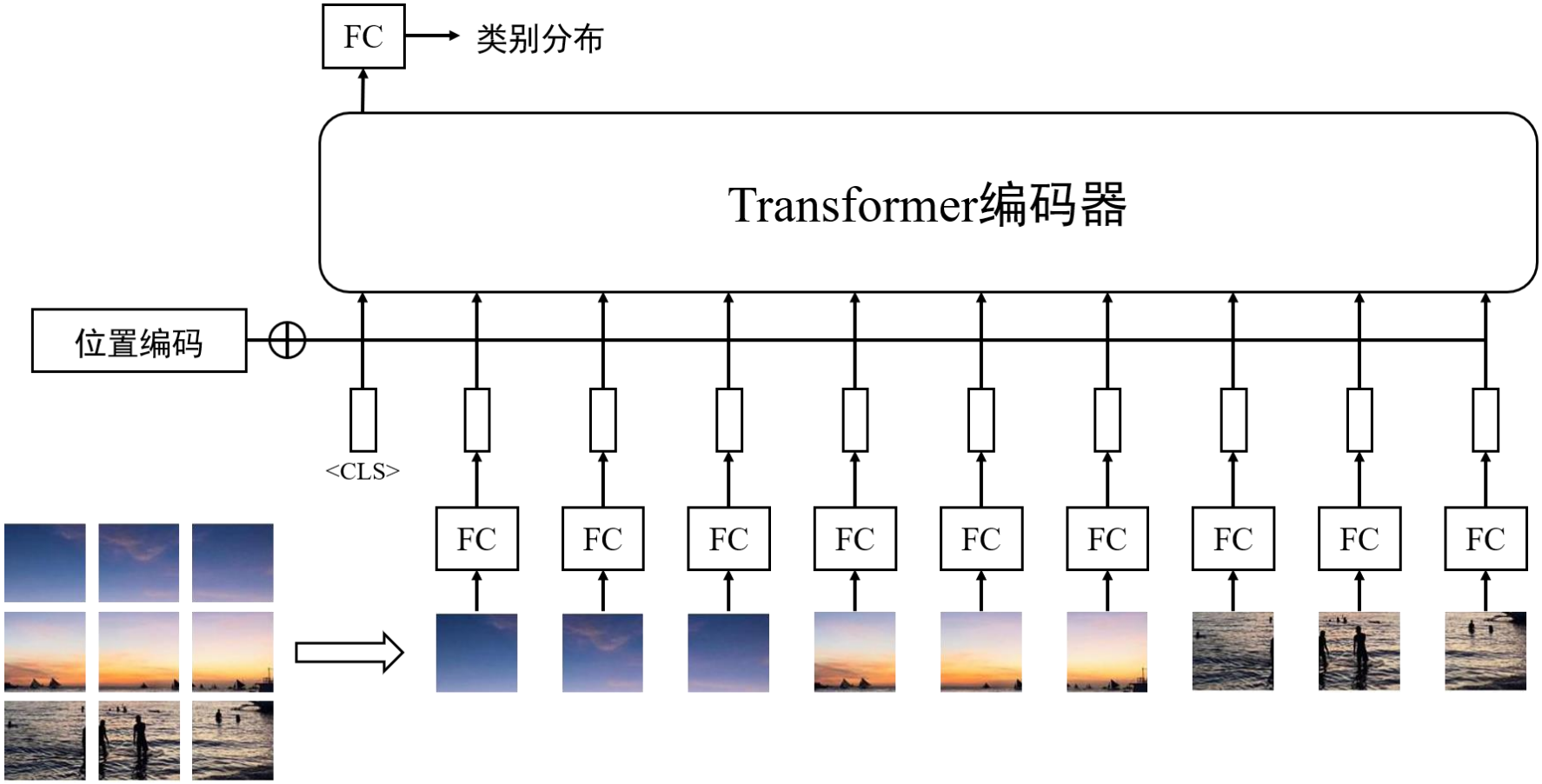
基于目标检测模型的区域表示



- 固定表示：
  36x(2048+4)
- 自适应表示：
  Kx(2048+4)
  (36<=K<=100)

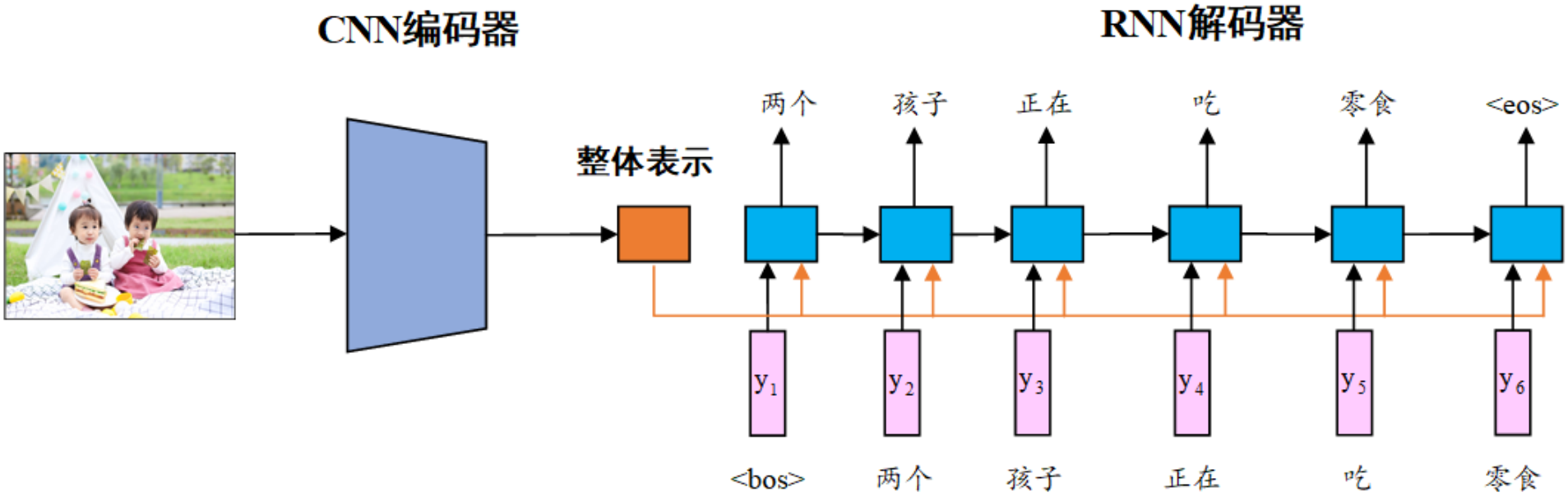
- 基于caffe的开源实现（原始实现）
  - https://github.com/peteanderson80/bottom-up-attention
- 基于pytorch的开源实现
  - https://github.com/facebookresearch/detectron2

视觉Transformer：总体框架



An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, Dosovitskiy et al. 2020

CNN整体表示+RNN



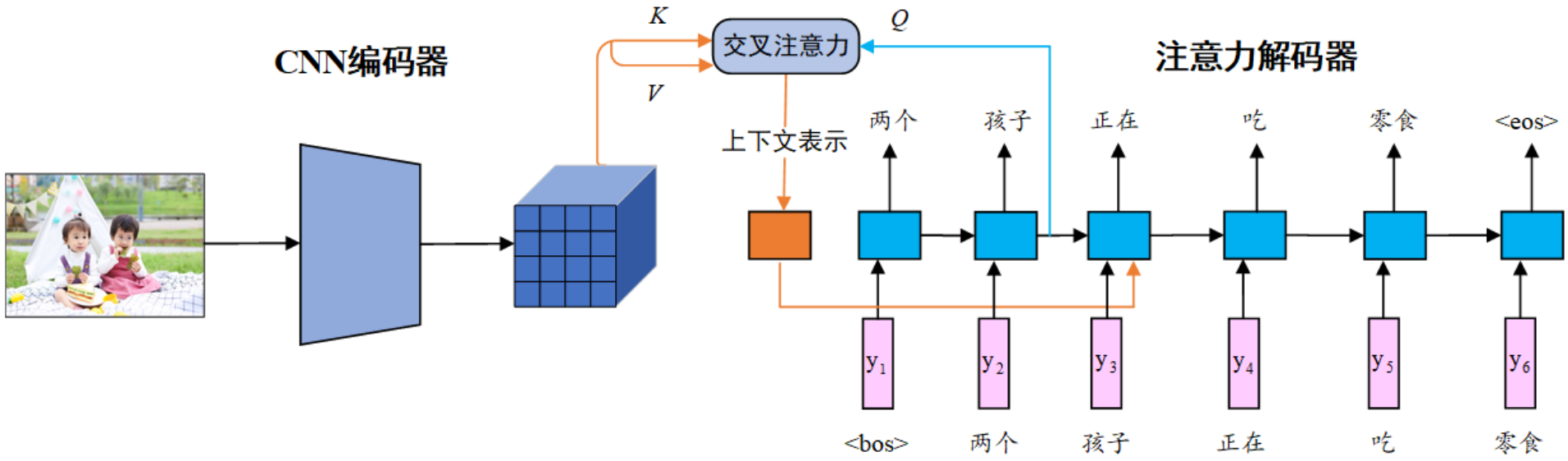
Explain Images with Multimodal Recurrent Neural Networks.
Deep Visual-Semantic Alignments for Generating Image Descriptions.
Show and Tell: A Neural Image Caption Generator.
Long-term Recurrent Convolutional Networks for Visual Recognition and Description.
Learning a Recurrent Visual Representation for Image Caption Generation.

文本生成策略

- 直接采样
  - 每次循环均采样概率最大的单词
  - 这种贪心的选取词的策略在计算复杂度低，但是无法获得全局最优解，因为每次都选取概率最大的词并不能保证整个句子出现的概率最大。
- 束搜索（Beam Search）采样
  - 假设窗口大小为 K，在 t 时刻有 K 个候选句子，在 t+1 时刻每个候选句子将采样概率值最大的前 K 个单词生成 K 个新的候选句子。如此，将生成 个新的候选句子，模型从中选择概率最大的前 K 个句子作为 t+1 时刻的候选句子。
  - 尽管束搜索策略也无法保证获得全局最优解，但是K越大，搜索空间就越大，也就越有可能获得更高概率的句子。
  - 当K=1时，束搜索策略等价于直接采样策略。

网格表示+注意力

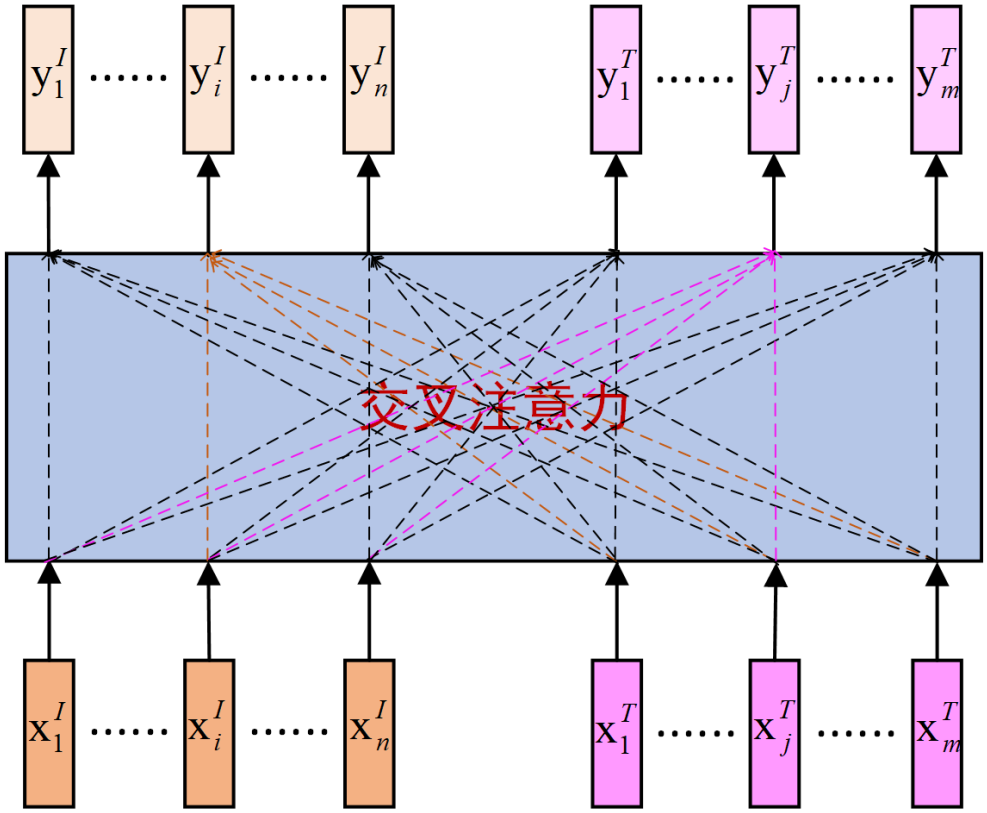
- 生成每个词时依赖不同的图像上下文向量
- 上下文向量是“注意”不同图像网格区域的结果



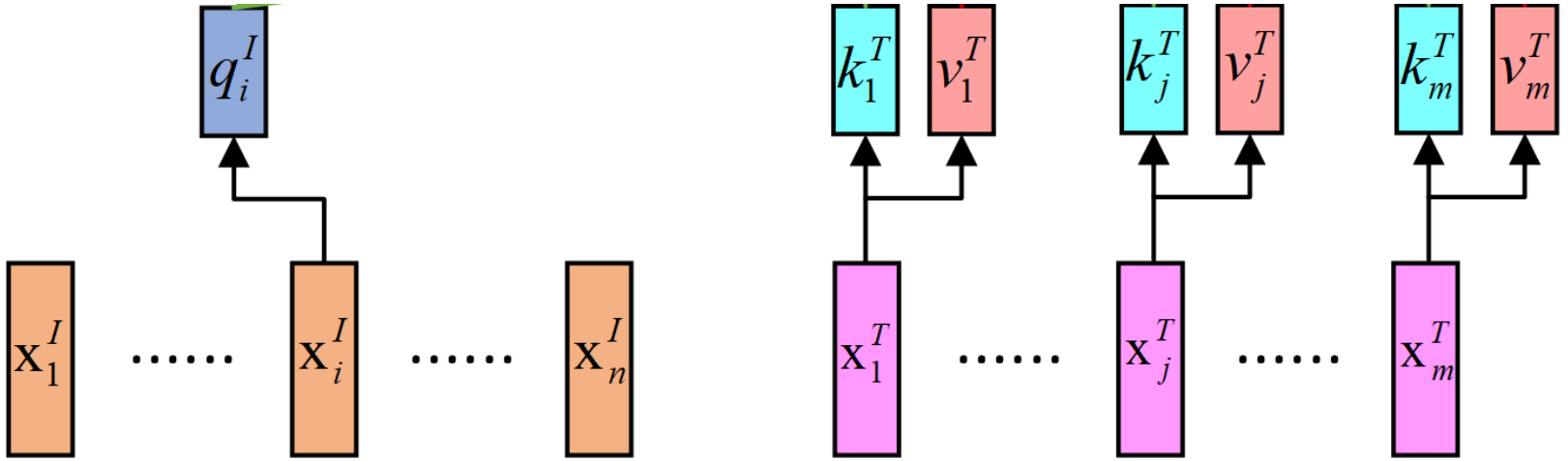
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. 2015

交叉注意力

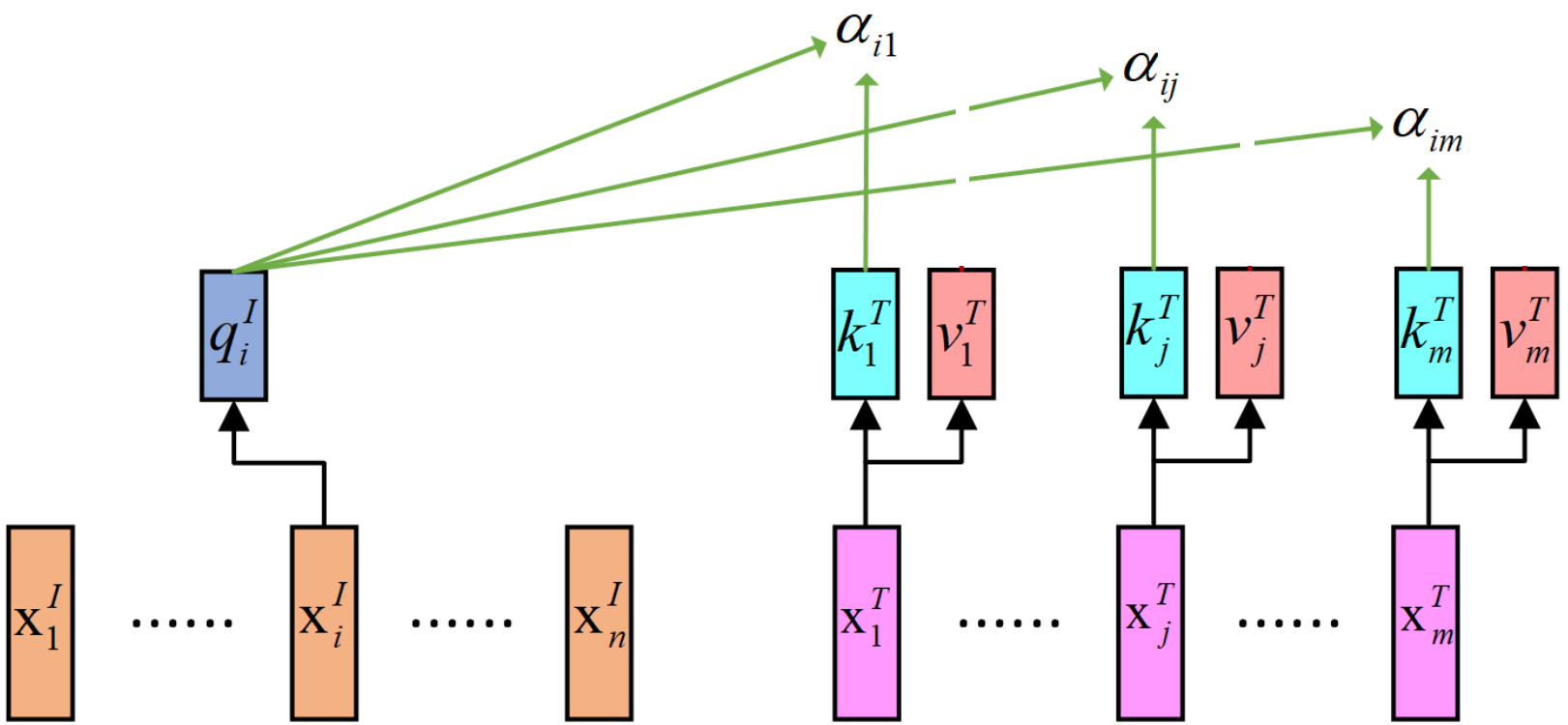
- Q、K、V不再来源于同一组输入，而是Q来源于一组输入，K和V来源于另一组输入



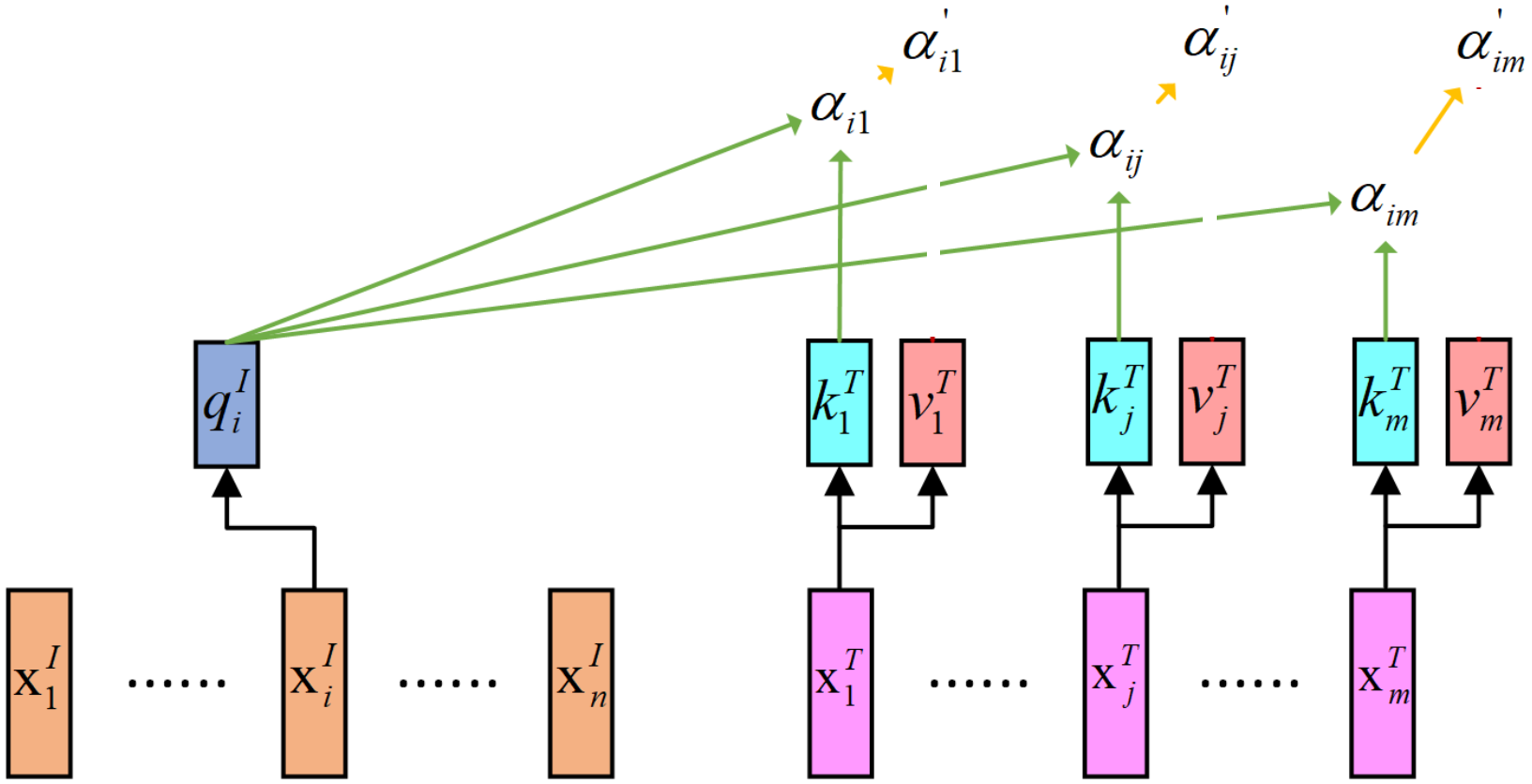
计算步骤1：Q、K、V



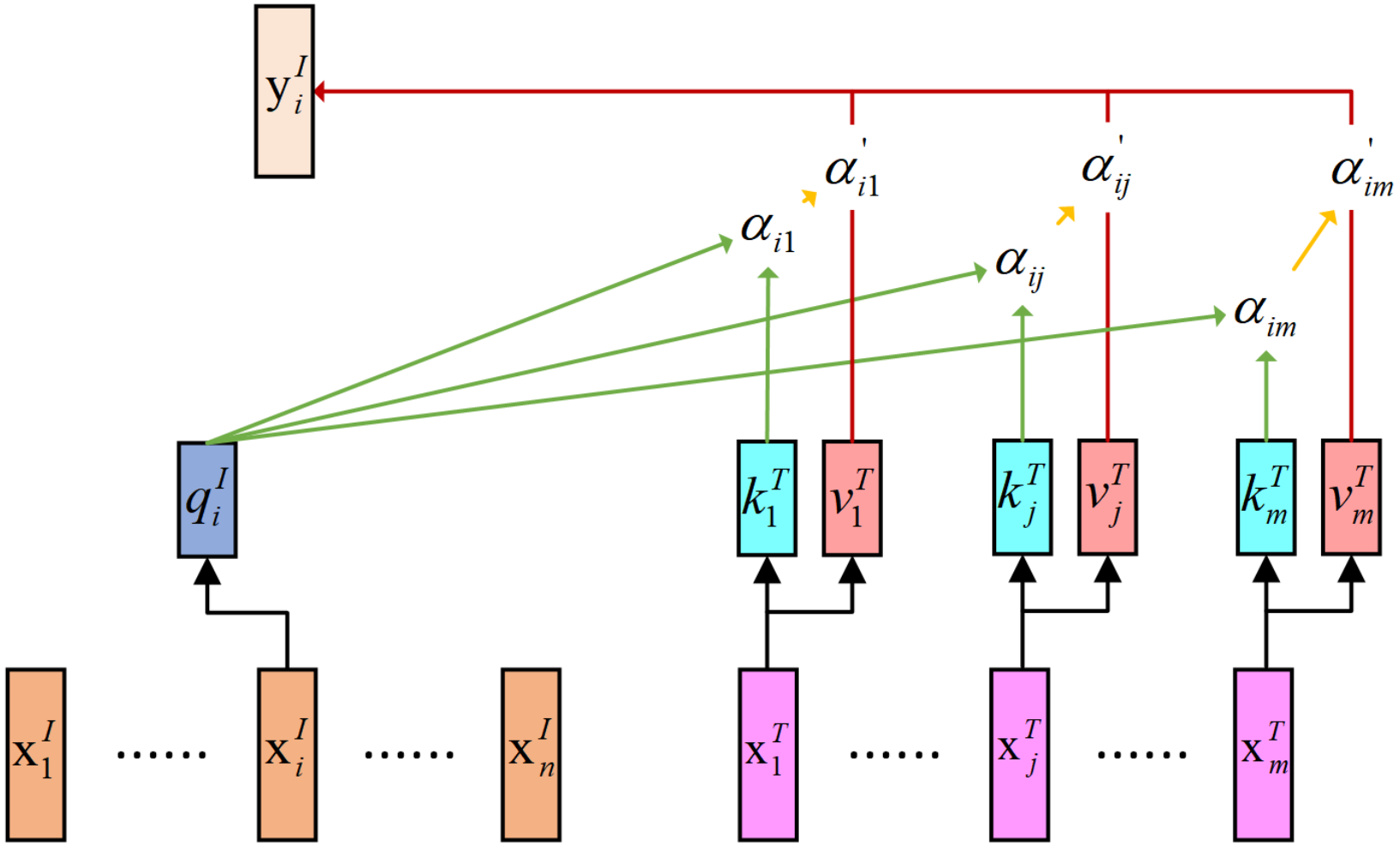
计算步骤2：注意力得分


α_{i1}
α_{ij}
α_{im}
q_i^I
k_1^T
v_1^T
k_j^T
v_j^T
k_m^T
v_m^T
x_1^I
x_i^I
x_n^I
x_1^T
x_j^T
x_m^T

计算步骤3：归一化注意力得分

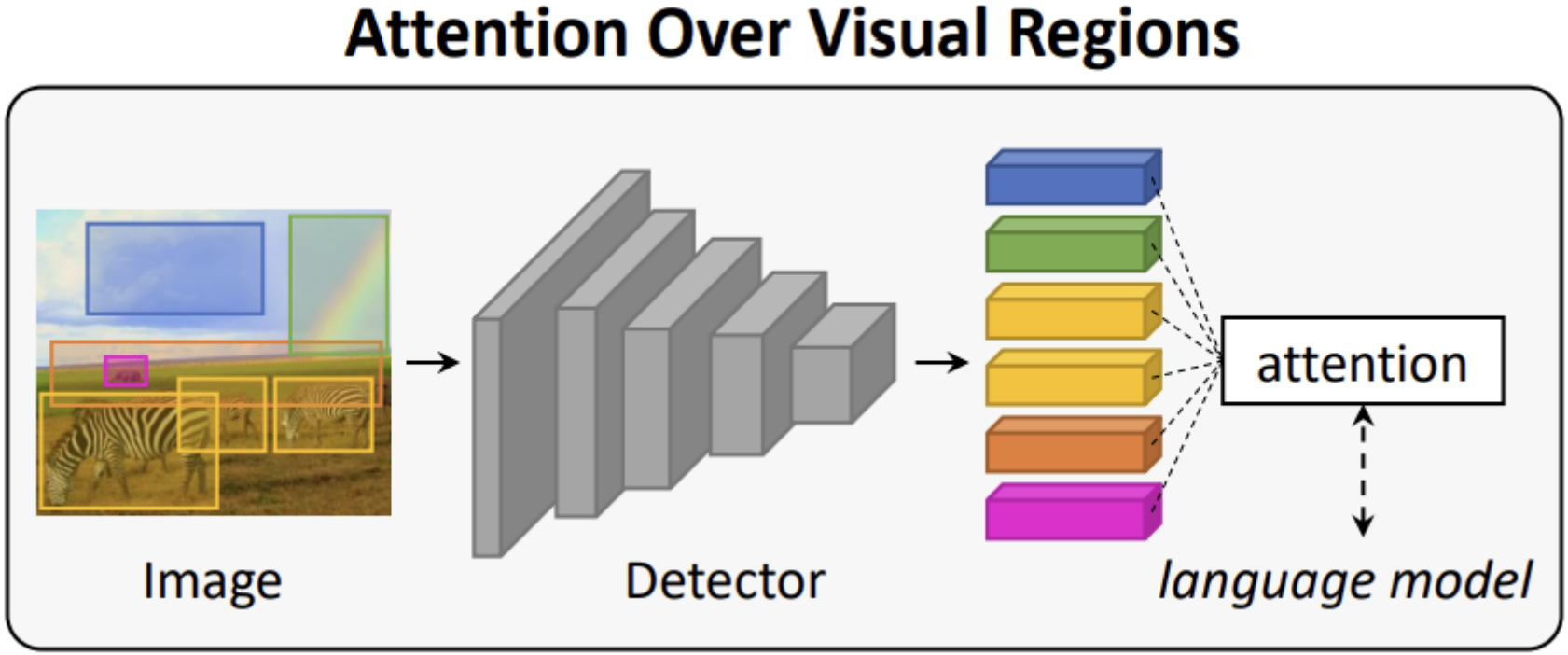


计算步骤4：加权求和


y_i^I
α'_{i1}
α'_{ij}
α'_{im}
α_{i1}
α_{ij}
α_{im}
q_i^I
k_1^T
v_1^T
k_j^T
v_j^T
k_m^T
v_m^T
x_1^I
x_i^I
x_n^I
x_1^T
x_j^T
x_m^T

区域表示+注意力

- 生成每个词时依赖不同的图像上下文向量
- 上下文向量是“注意”不同图像目标区域的结果



Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. 2018

网格表示、Transformer编码器+Transformer解码器

Input: Image **I**
Output: Sequence **y** = $y_1, y_2, ..., y_T$

Decoder: $y_t = T_D(\mathbf{y}_{0:t-1}, \mathbf{c})$
where $T_D(.)$ is the transformer decoder

Encoder: $\mathbf{c} = T_W(\mathbf{z})$
where **z** is spatial CNN features
$T_W(.)$ is the transformer encoder

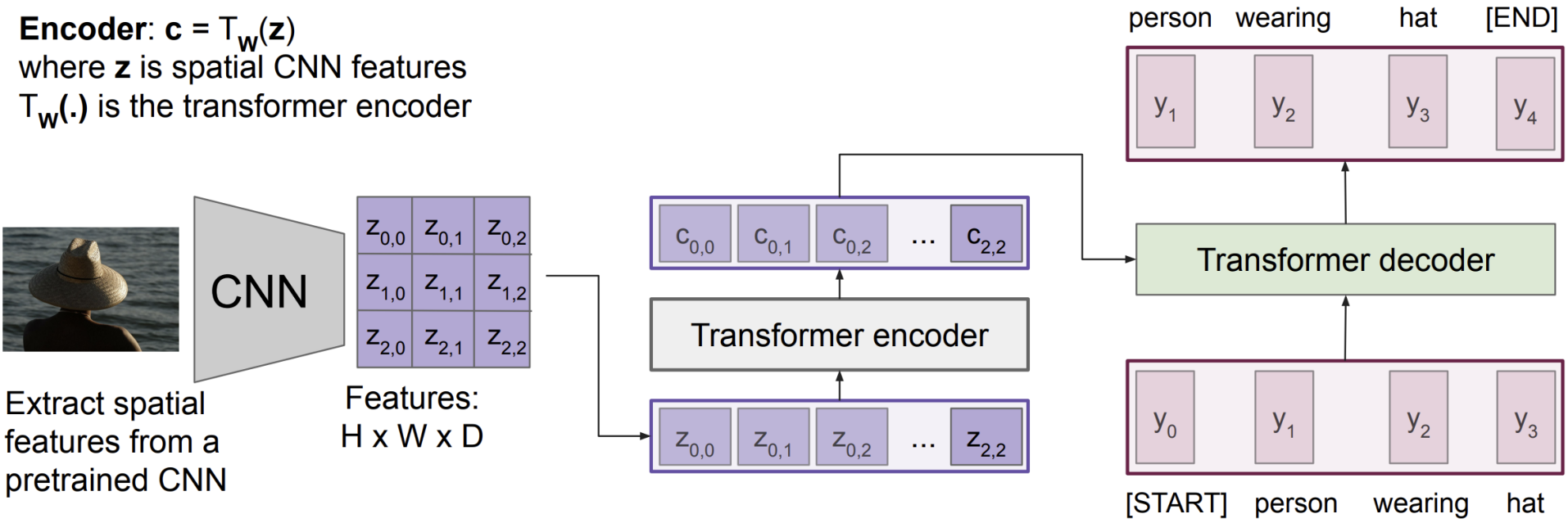
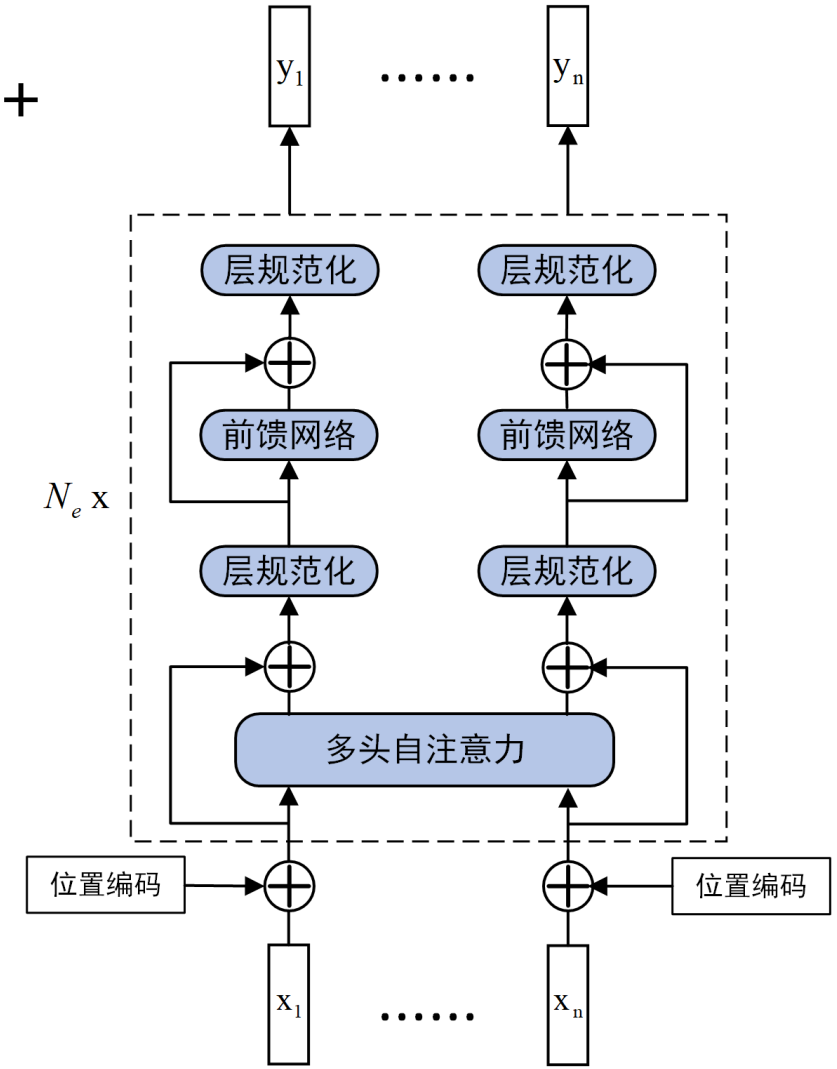


Image Captioning: Transforming Objects into Words. 2019
Normalized and Geometry-Aware Self-Attention Network for Image Captioning. 2020
Dual-Level Collaborative Transformer for Image Captioning. 2021

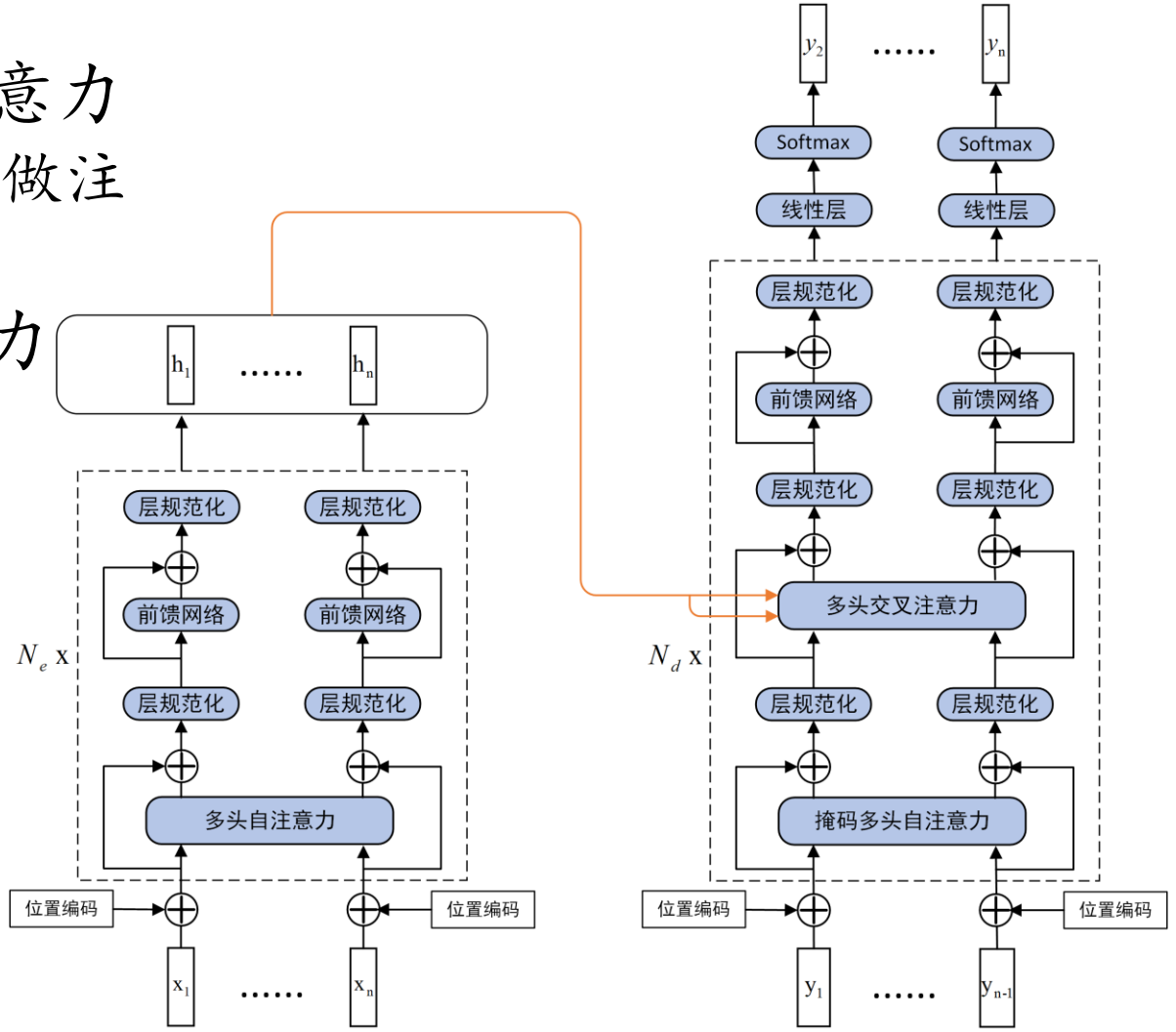
Transformer编码器

- 堆叠多个多头自注意力层+全连接层
  - $\mathrm{Max}(0, xW_1 + b_1)W_2 + b_2$
- 位置编码
- 残差连接
- Layer Normalization



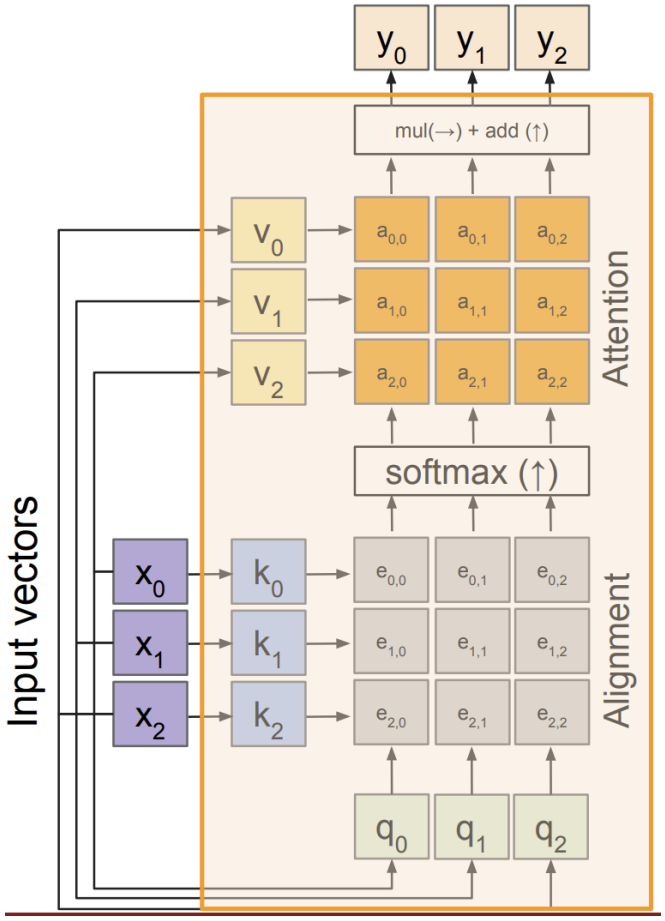
Transformer编码器+解码器

- 掩码多头自注意力
  - 只与之前单词做注意力计算
- 多头交叉注意力

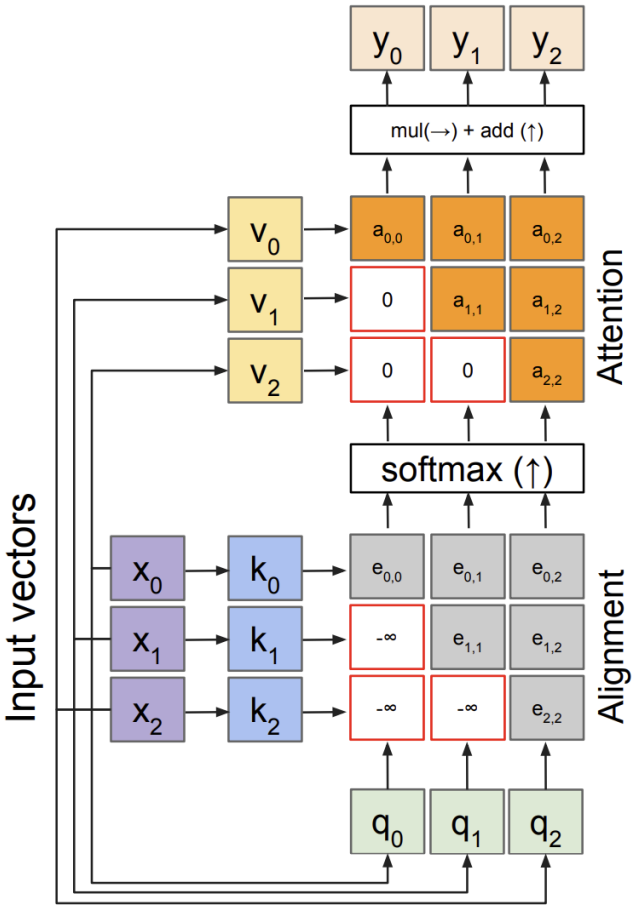


掩码自注意力

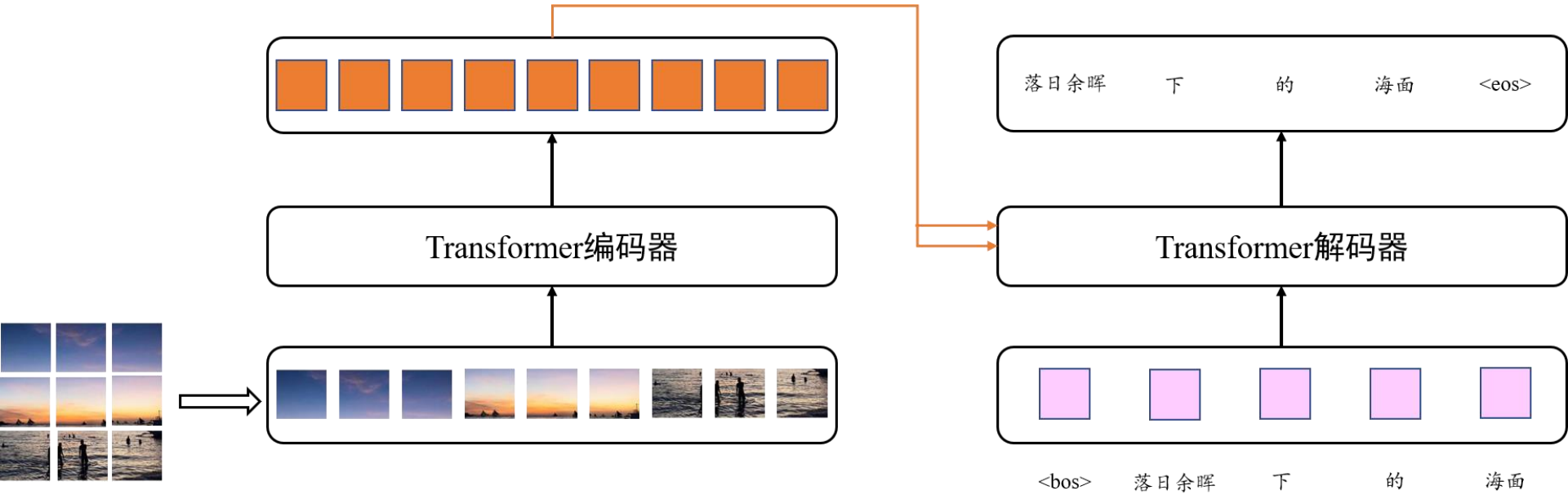
自注意力



掩码自注意力



视觉Transformer+Transformer解码器



CPTR: FULL TRANSFORMER NETWORK FOR IMAGE CAPTIONING. 2021

作业

服饰图像描述：必选任务

- 数据集：仅用到image和textual descriptions，https://github.com/yumingj/DeepFashion-MultiModal
  - 请注意文本描述不是单一句子
  - 训练/测试集切分会在下周发到课程QQ群
- 至少新实现下列模型结构中的两种：
  - CNN/ ViT整体表示+GRU
  - 网格/区域表示、自注意力+注意力
  - 网格/区域表示、Transformer编码器+Transformer解码器
  - 网格/区域表示、图网络+ Transformer解码器
  - 视觉Transformer+Transformer解码器
- 至少新实现以下评测标准中的两个：
  - METEOR、ROUGE-L、CIDEr-D、SPICE

服饰图像描述：可选任务

- 默认实现的交叉熵损失和评测指标不一致，请实现基于强化学习的损失函数，直接优化评测指标
- 微调多模态预训练模型或多模态大语言模型，并测试性能
- 利用训练的服饰图像描述模型和多模态大语言模型，为真实背景的服饰图像数据集增加服饰描述和背景描述，构建全新的服饰图像描述数据集
  - 在新数据集上重新训练服饰图像描述模型

谢谢