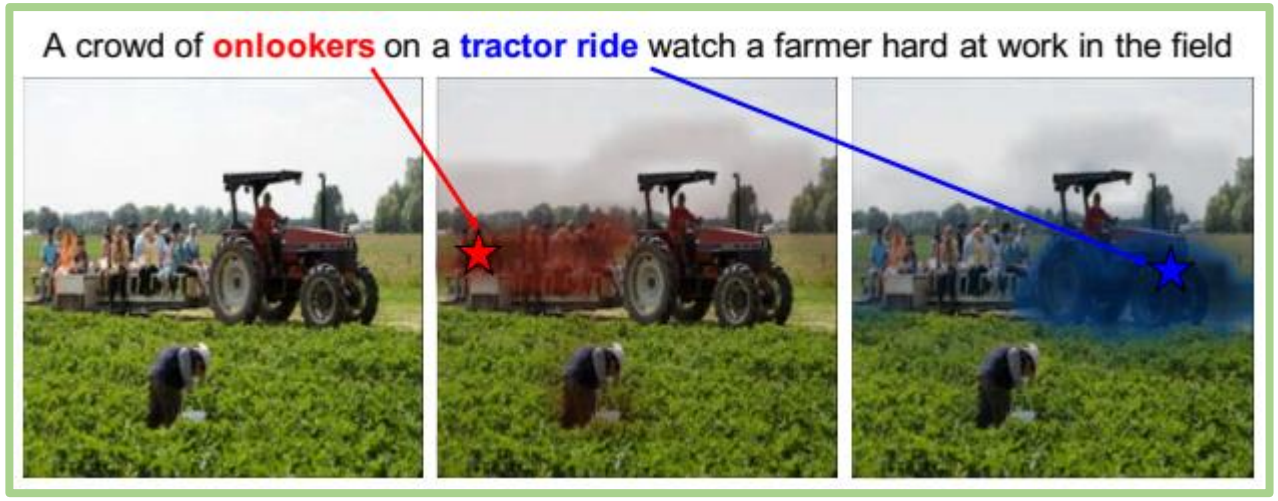
Visual Grounding

Luke Ye

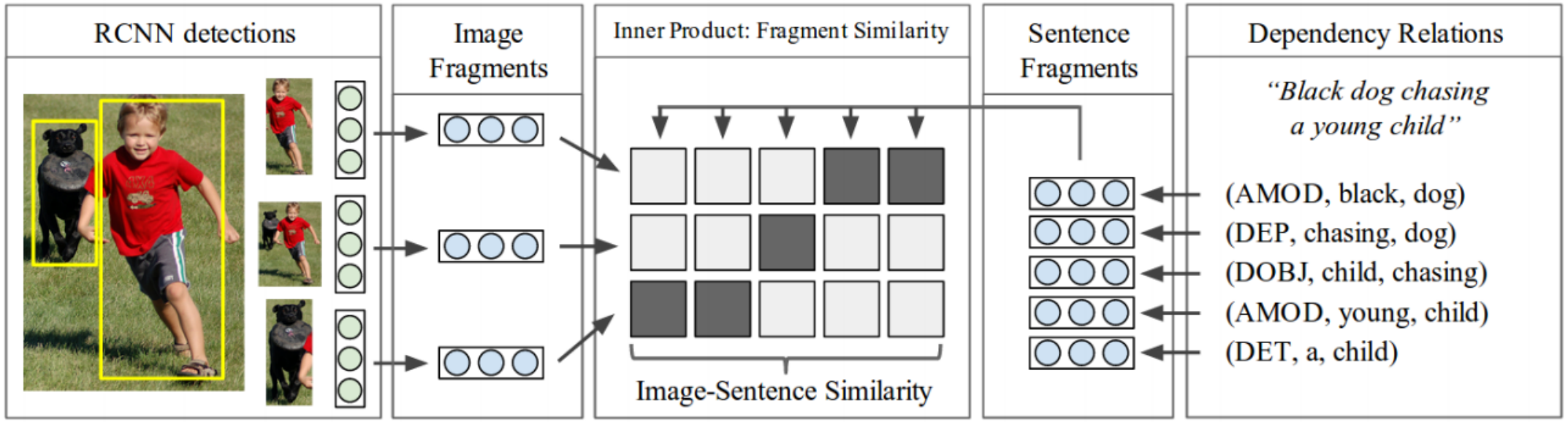
任务介绍

建立文本概念到视觉概念的对齐关系

Single-Object	Multi-Object
带有文本限定的单目标对齐	多个目标跨模态的对齐



起源



2014 NIPS, Deep Fragment Embeddings for Bidirectional Image Sentence Mapping

起源

Fragment Alignment Objective

$$\mathcal{C}_0(\theta) = \sum_i \sum_j max(0, 1 - y_{ij} v_i^T s_j).$$

$$\mathcal{C}_F(\theta) = \min_{y_{ij}} \mathcal{C}_0(\theta)$$

$$\text{s.t.} \quad \sum_{i \in p_j} \frac{y_{ij} + 1}{2} \geq 1 \quad \forall j$$

$$y_{ij} = sign(v_i^T s_j)$$

$$y_{ij} = -1 \quad \forall i, j \quad \text{s.t.} \quad m_v(i) \neq m_s(j) \text{ and } y_{ij} \in \{-1, 1\}$$

Global Ranking Objective

$$S_{kl} = \frac{1}{|g_k|(|g_l| + n)} \sum_{i \in g_k} \sum_{j \in g_l} max(0, v_i^T s_j).$$

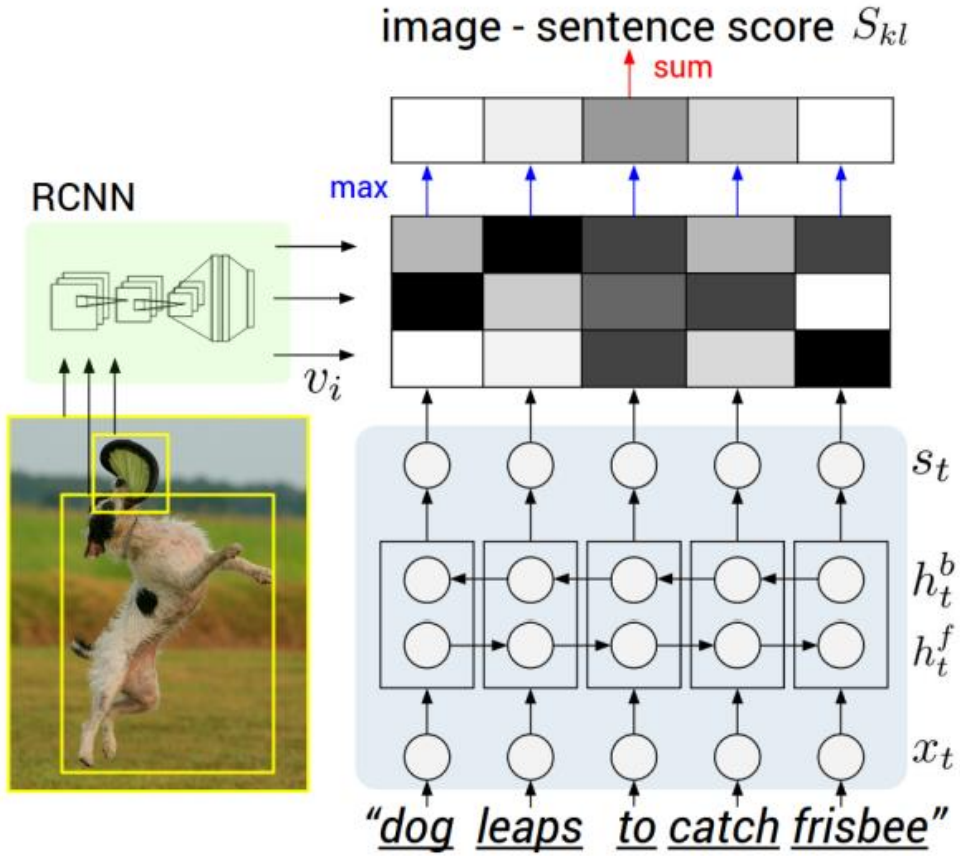
$$\mathcal{C}_G(\theta) = \sum_k \Big[\underbrace{\sum_l max(0, S_{kl} - S_{kk} + \Delta)}_{\text{rank images}} + \underbrace{\sum_l max(0, S_{lk} - S_{kk} + \Delta)}_{\text{rank sentences}} \Big].$$

起源

$$S_{kl} = \sum_{t \in g_l} max_{i \in g_k} v_i^T s_t$$

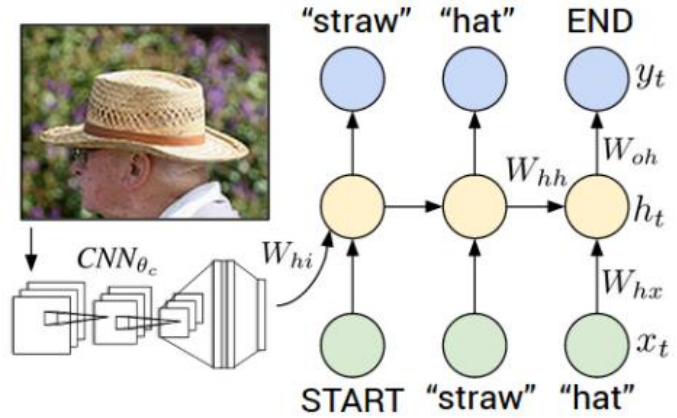
$$\mathcal{C}(\theta) = \sum_k \Big[\underbrace{\sum_l max(0, S_{kl} - S_{kk} + 1)}_{\text{rank images}} + \underbrace{\sum_l max(0, S_{lk} - S_{kk} + 1)}_{\text{rank sentences}} \Big].$$

细粒度标注模型



$$b_v = W_{hi}[CNN_{\theta_c}(I)]$$
$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + \mathbb{1}(t=1) \odot b_v)$$
$$y_t = softmax(W_{oh}h_t + b_o).$$

Caption生成模型



2015 CVPR, Deep Visual-Semantic Alignments for Generating Image Descriptions

起源

检索式

简单模型

复杂模型

	Flickr8K				Flickr30K				MSCOCO 2014					
Model	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	METEOR	CIDEr
Nearest Neighbor	—	—	—	—	—	—	—	—	48.0	28.1	16.6	10.0	15.7	38.3
Mao et al. [38]	58	28	23	—	55	24	20	—	—	—	—	—	—	—
Google NIC [54]	63	41	27	—	66.3	42.3	27.7	18.3	66.6	46.1	32.9	24.6	—	—
LRCN [8]	—	—	—	—	58.8	39.1	25.1	16.5	62.8	44.2	30.4	—	—	—
MS Research [12]	—	—	—	—	—	—	—	—	—	—	—	21.1	20.7	—
Chen and Zitnick [5]	—	—	—	14.1	—	—	—	12.6	—	—	—	19.0	20.4	—
Our model	57.9	38.3	24.5	16.0	57.3	36.9	24.0	15.7	62.5	45.0	32.1	23.0	19.5	66.0

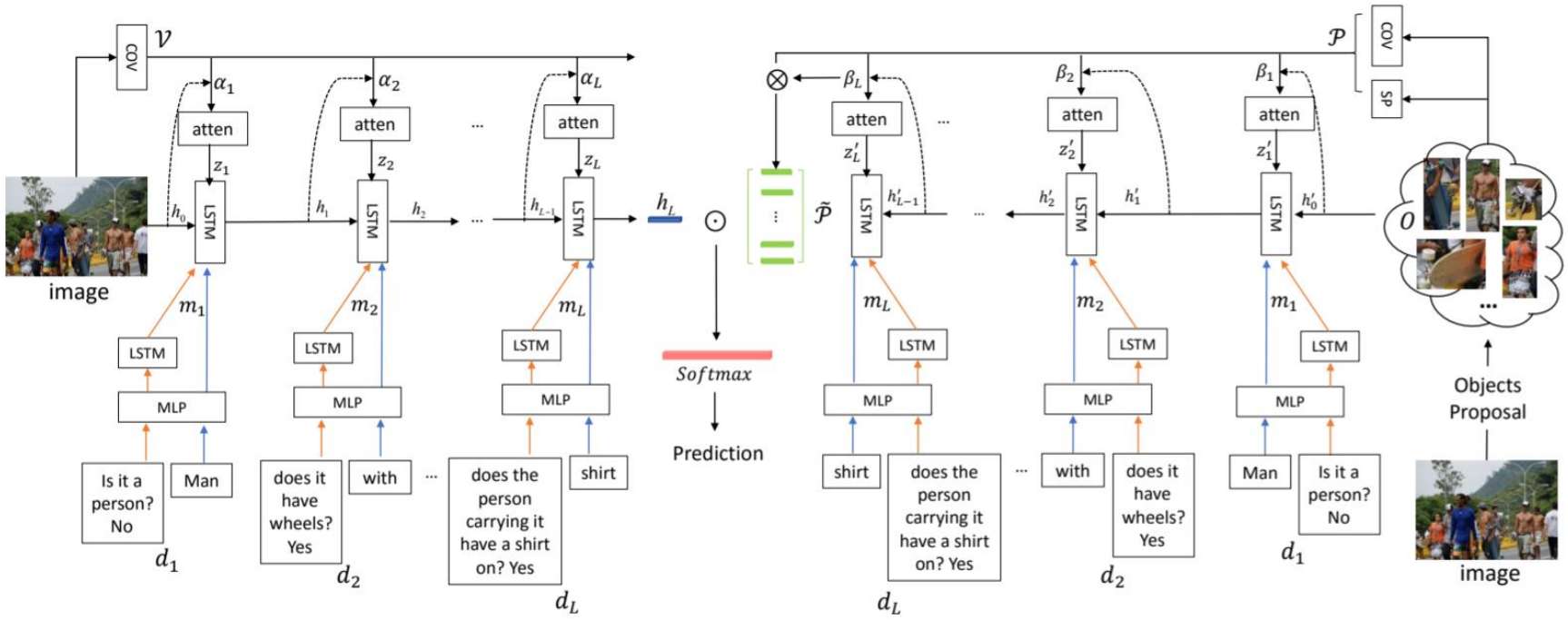
Table 2. Evaluation of full image predictions on 1,000 test images. **B-n** is BLEU score that uses up to n-grams. High is good in all columns. For future comparisons, our METEOR/CIDEr Flickr8K scores are 16.7/31.8 and the Flickr30K scores are 15.3/24.7.

2015 CVPR, Deep Visual-Semantic Alignments for Generating Image Descriptions

Attention

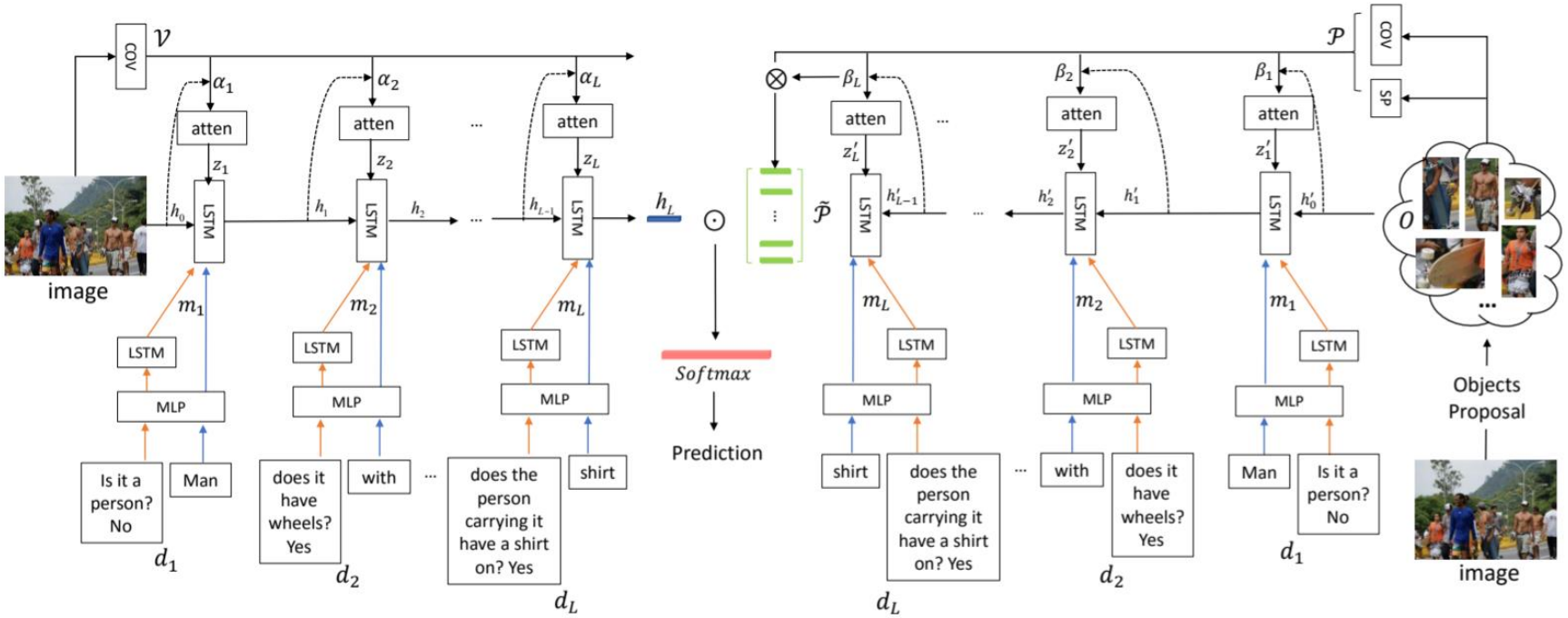
Image-level Attention

$$e_{ti} = \tanh\left(W_v v_i + W_h h_{t-1}\right)$$
$$\alpha_{ti} = \mathrm{softmax}\left(\mathrm{e_{ti}}\right)$$
$$z_t = \sum_{i=1}^{K} \alpha_{ti} e_{ti}$$
$$h_t = \mathrm{LSTM}\left(m_t, z_t, h_{t-1}\right)$$



2018 CVPR, Parallel Attention: A Unified Framework for Visual Object Discovery through Dialogs and Queries

Attention



Propose-level Attention

$$p_i = [u_i, s_i, c_i]$$

$$e'_{ti} = \tanh\left(W_p p_i + W'_h h'_{t-1}\right)$$

$$\beta_{ti} = \text{softmax}\left(e'_{ti}\right)$$

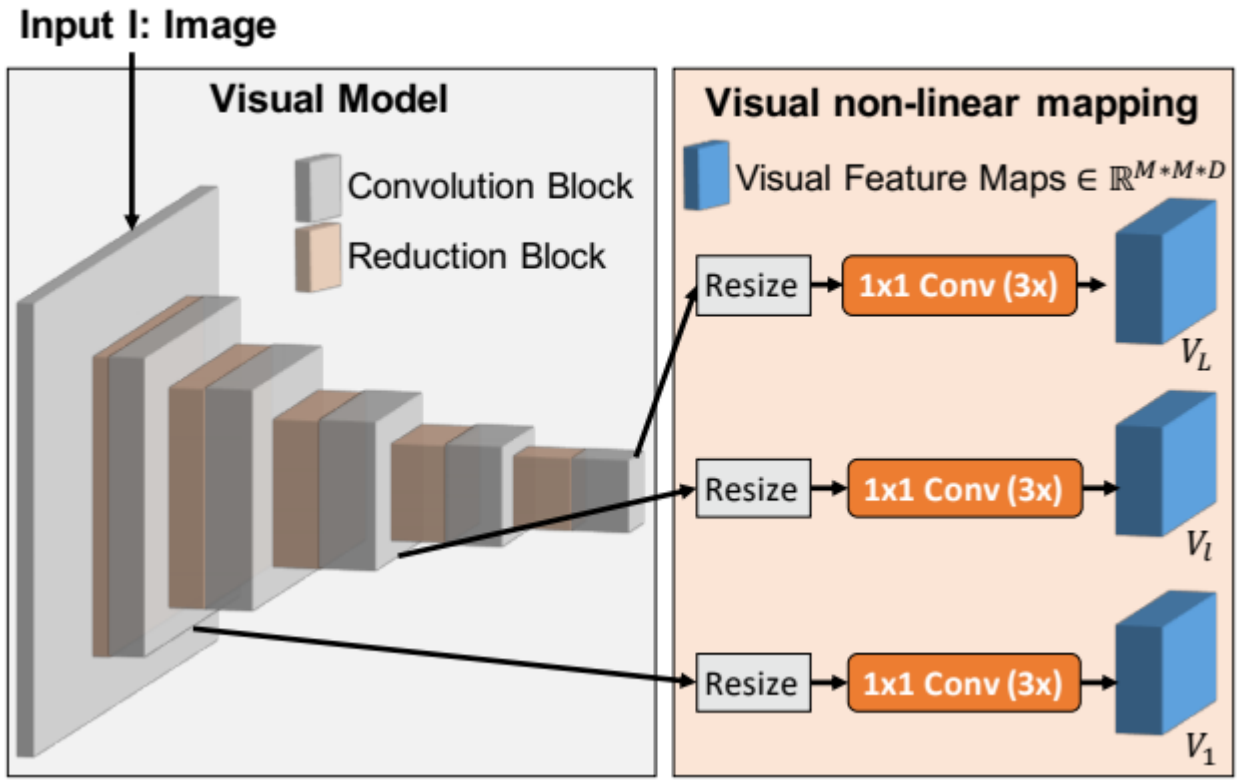
$$z'_t = \sum_{i=1}^{K} \beta_{ti} e'_{ti}$$

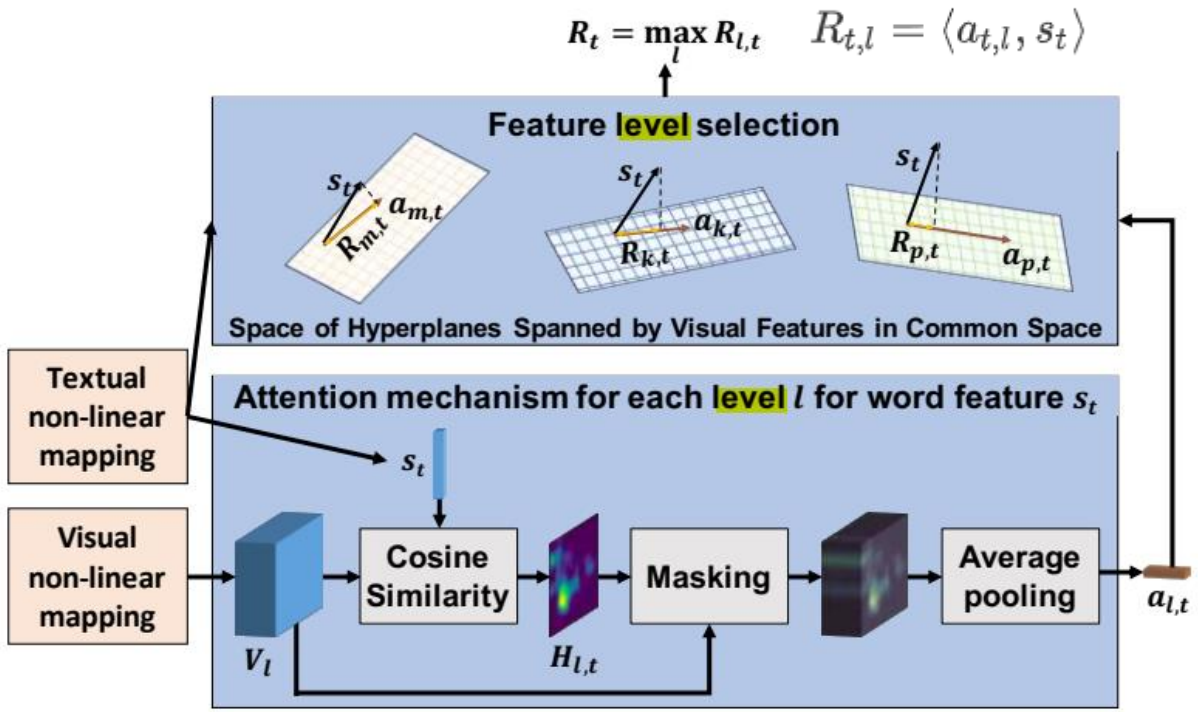
$$h'_t = \text{LSTM}\left(m_t, z'_t, h'_{t-1}\right)$$

$$\tilde{p}_i = \beta_{Li} p_i$$

2018 CVPR, Parallel Attention: A Unified Framework for Visual Object Discovery through Dialogs and Queries

Layer-Attention





最终图像和文本的相似度：$R_w(S,I) = \log\left(\left(\sum_{t=0}^{T-1}\exp\left(\gamma_1 R_t\right)\right)^{\frac{1}{\gamma_1}}\right).$

2019 CVPR, Multi-level Multimodal Common Semantic Space for Image-Phrase Grounding

Co-Attention

$$Q_l \in \mathbb{R}^{d\times N} \quad V_l \in \mathbb{R}^{d\times T}$$

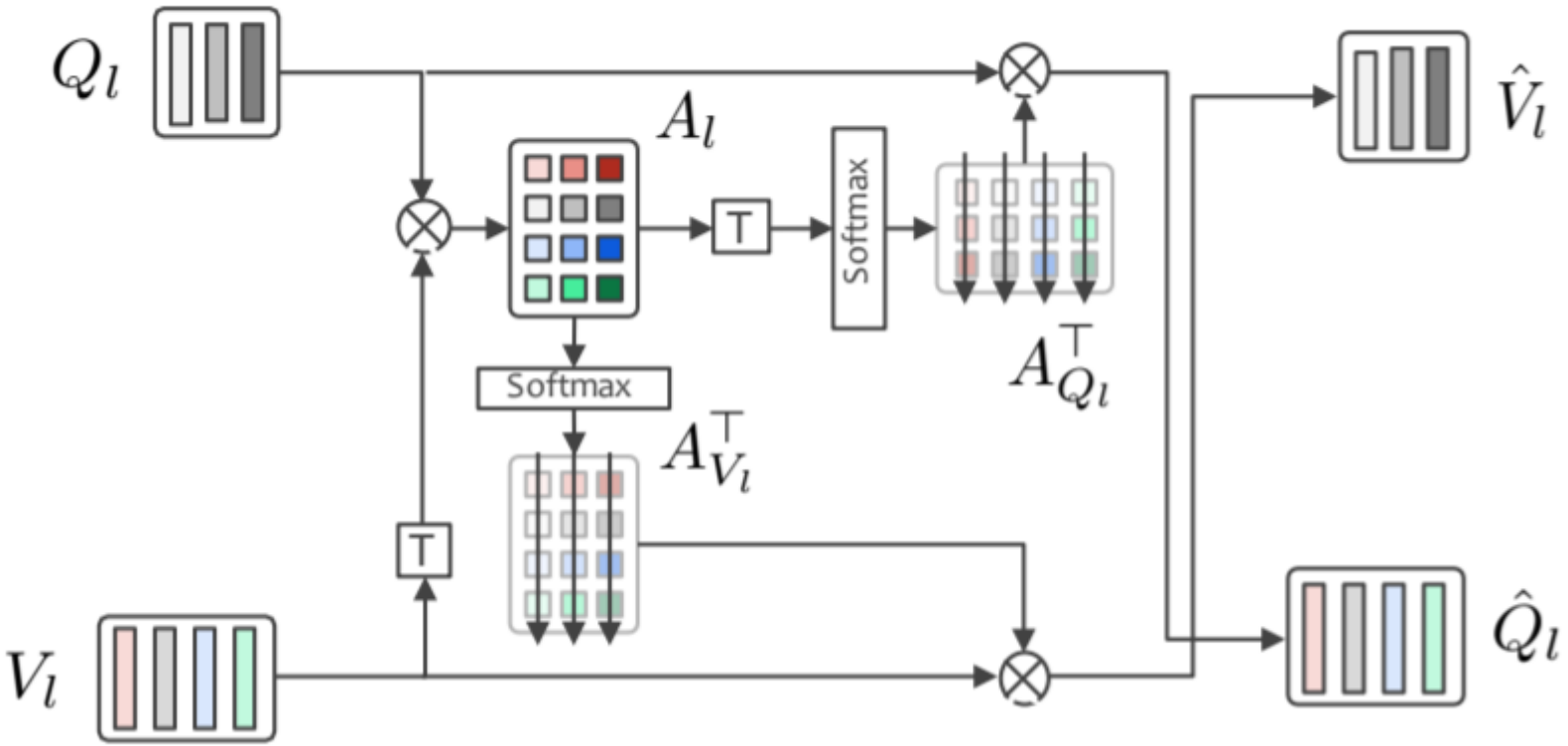
$$A_l^{(i)} = \left(W_{V_l}^{(i)} V_l\right)^{\mathrm{T}} \left(W_{Q_l}^{(i)} Q_l\right)$$

$$A_{Q_l} = \frac{1}{h}\sum_{i=1}^{h} \mathrm{softmax}\left(\frac{A_l^{(i)}}{\sqrt{d_h}}\right)$$

$$A_{V_l} = \frac{1}{h}\sum_{i=1}^{h} \mathrm{softmax}\left(\frac{A_l^{(i)\,\mathrm{T}}}{\sqrt{d_h}}\right)$$

$$\hat{Q}_l = Q_l A_{Q_l}^{\mathrm{T}}$$

$$\hat{V}_l = V_l A_{V_l}^{\mathrm{T}}$$



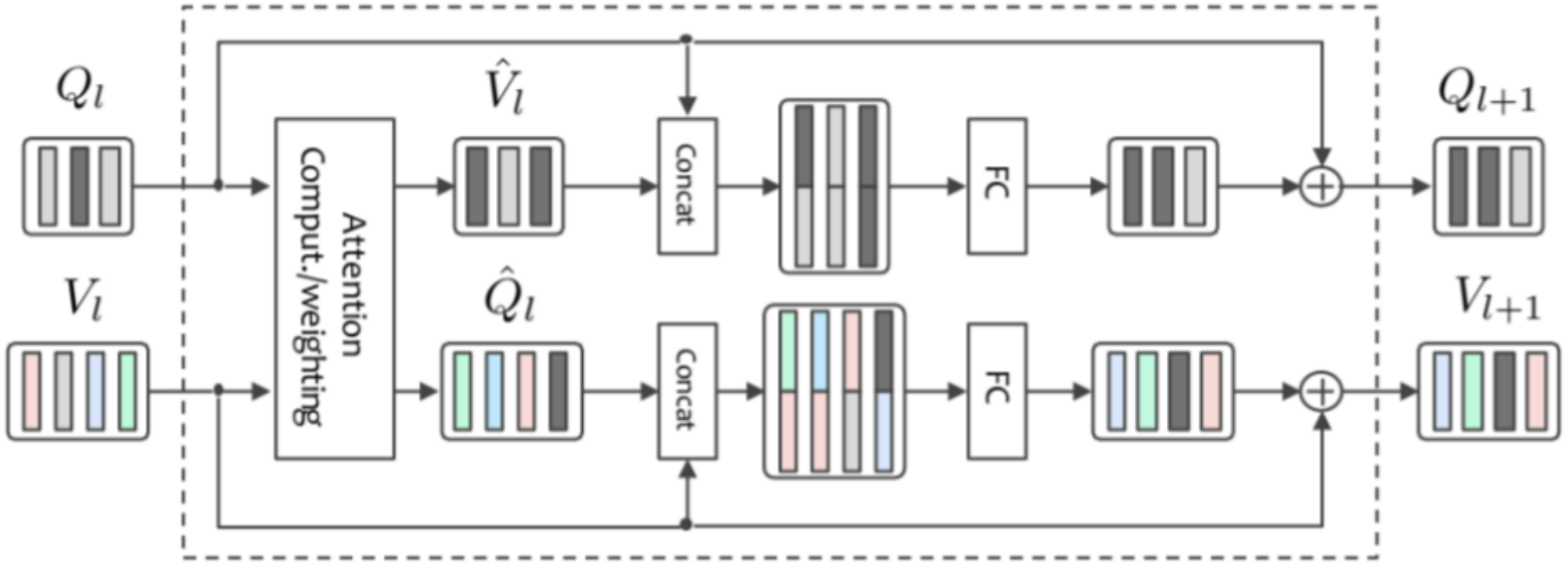
No where elements的处理

$\tilde{Q}_l \in \mathbb{R}^{d\times(N+K)}, \tilde{V}_l \in \mathbb{R}^{d\times(T+K)}$。

$$\hat{Q}_l = \tilde{Q}_l A_{Q_l}[1:T,:]^{\mathrm{T}}$$
$$\hat{V}_l = \tilde{V}_l A_{V_l}[1:N,:]^{\mathrm{T}}$$

2018 CVPR, Improved Fusion of Visual and Language Representations by Dense Symmetric Co-Attention for Visual Question Answering

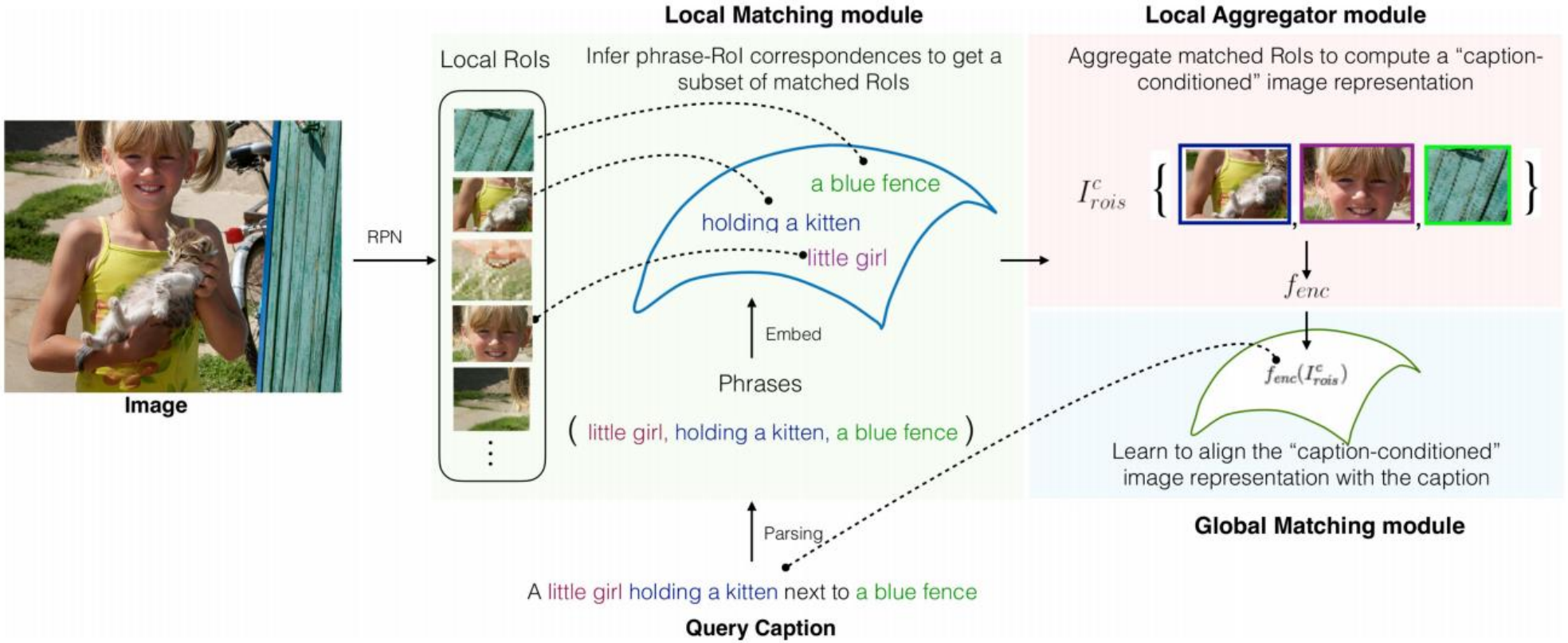
Co-Attention



$$Q_{l+1} = \mathrm{ReLU}\left(W_{Q_l}\left[Q_l, \hat{V}_l\right] + b_{Q_l}\right) + Q_l$$
$$V_{l+1} = \mathrm{ReLU}\left(W_{V_l}\left[V_l, \hat{Q}_l\right] + b_{V_l}\right) + V_l$$

2018 CVPR, Improved Fusion of Visual and Language Representations by Dense Symmetric Co-Attention for Visual Question Answering

Deal with Insufficient Alignment



2019 ICCV, Align2Ground: Weakly Supervised Phrase Grounding Guided by Image-Caption Alignment

Conclusion

1. Visual Grounding的文章每年很多，但有创新性的文章不多。
2. Visual Grounding任务定义繁杂，评估标准不统一
3. 只用弱监督的文本图片匹配可能不够，也许可以考虑与其他领域结合