关于语义与图片特征的融合方法

主讲人：朱泽阳

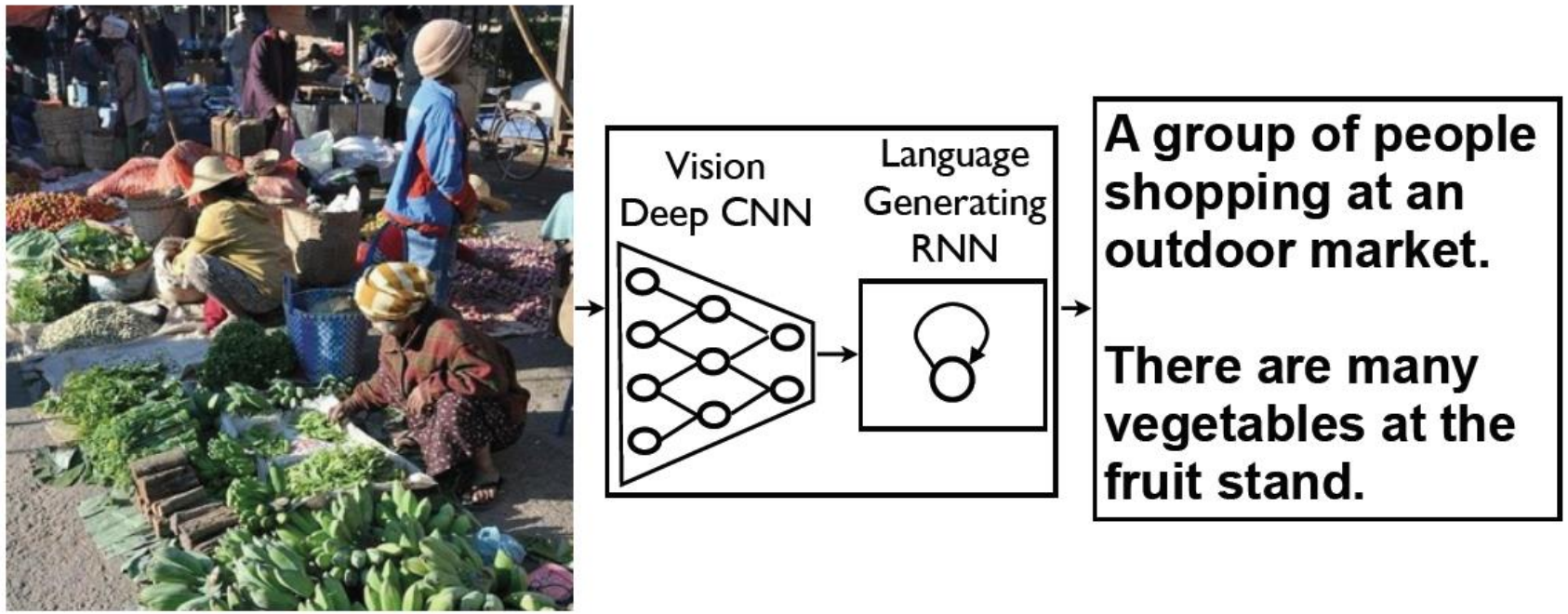
融合方法

- 拼接 $\phi_{xt} = f_{\text{MLP}}([\phi_x, \phi_t])$.
- LSTM-based 方法
- Parameter hashing方法
- Relationship方法
- FiLM (Feature-wise Linear Modulation)方法
- TIRG (Text Image Residual Gating)方法

融合方法

LSTM-based方法

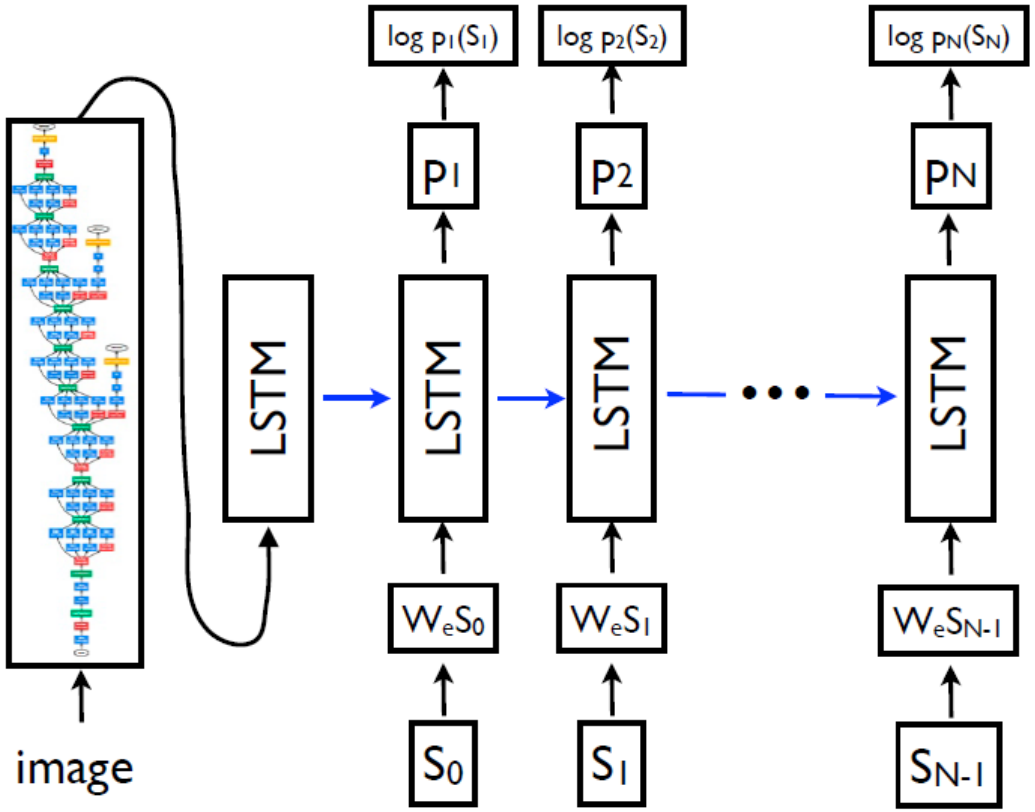
任务：Image caption



Show and tell: A neural image caption generator. In CVPR, 2015.

融合方法

LSTM-based方法



为什么这篇论文能中？

当时LSTM和CNN正流行。结合最好的框架来解决当时“热”问题（image caption）。可能因为是Vinyals（大佬）的论文，所以中了。但是不得不感叹该作者的很多论文都写的非常好，简洁易懂。

Show and tell: A neural image caption generator. In CVPR, 2015.

融合方法

Parameter hashing方法

任务：Visual Question Answering (VQA)

VQA的挑战:它是一种整体的场景理解，需要一个系统在语义的许多不同层次上捕捉各种各样的信息，如物体、动作、事件、场景、气氛以及它们之间的关系。不同的问题需要不同类型和层次的理解，才能找到正确的答案。

Q: What type of animal is this?
Q: Is this animal alone?

Q: Is it snowing?
Q: Is this picture taken during the day?

Q: What kind of oranges are these?
Q: Is the fruit sliced?

Q: What is leaning on the wall?
Q: How many boards are there?

Image question answering using convolutional neural network with dynamic parameter prediction. In CVPR, 2016.

融合方法

Parameter hashing方法

$$\mathbf{f}^o = \mathbf{W}_d(q)\mathbf{f}^i + \mathbf{b}$$

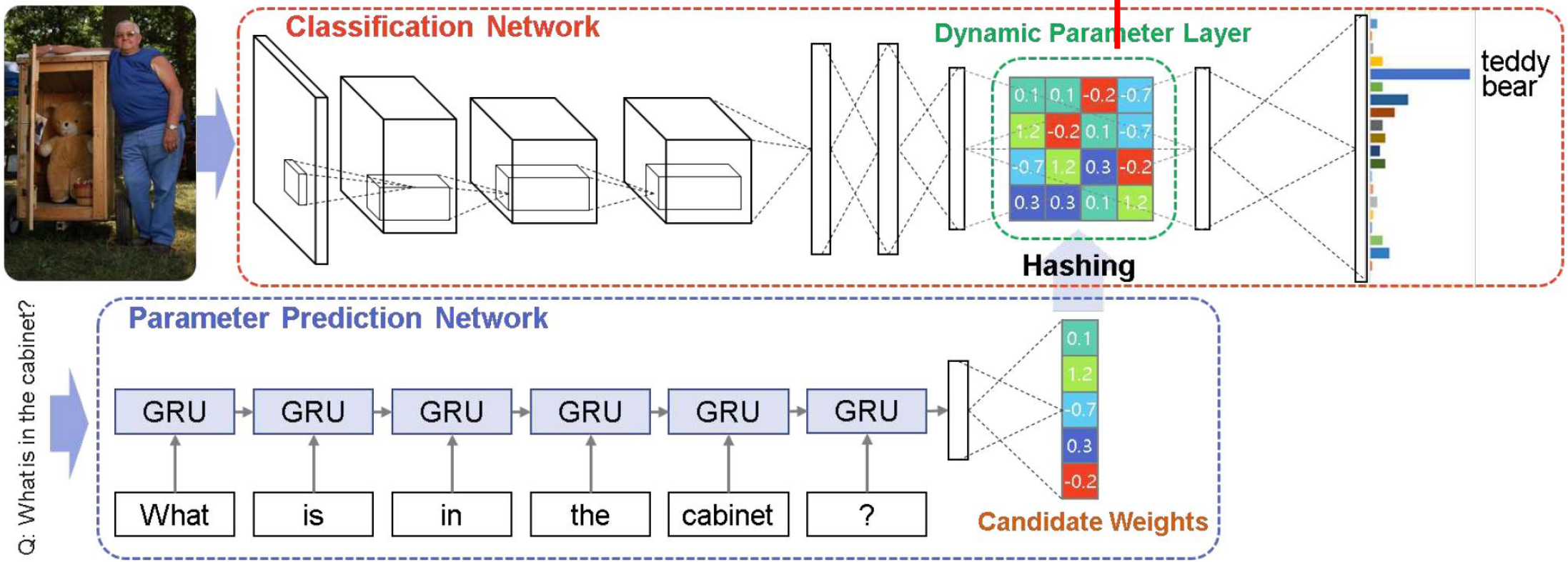
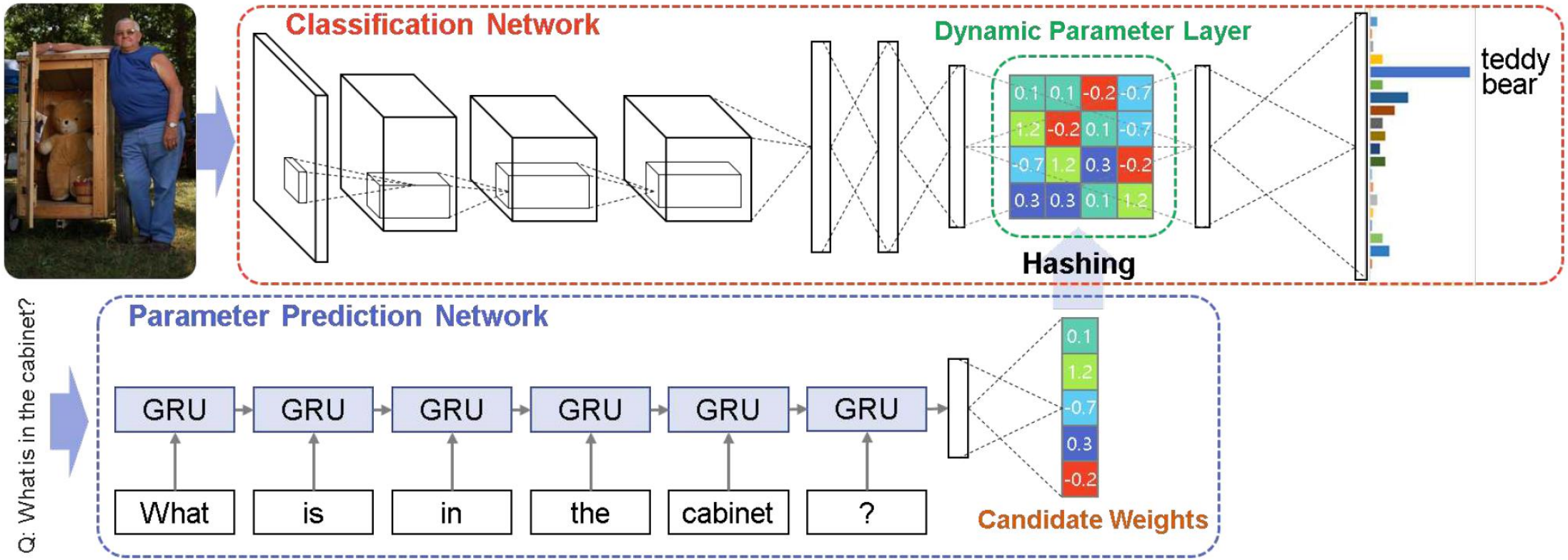


Image question answering using convolutional neural network with dynamic parameter prediction. In CVPR, 2016.

融合方法

Parameter hashing方法



为什么这篇论文能中?

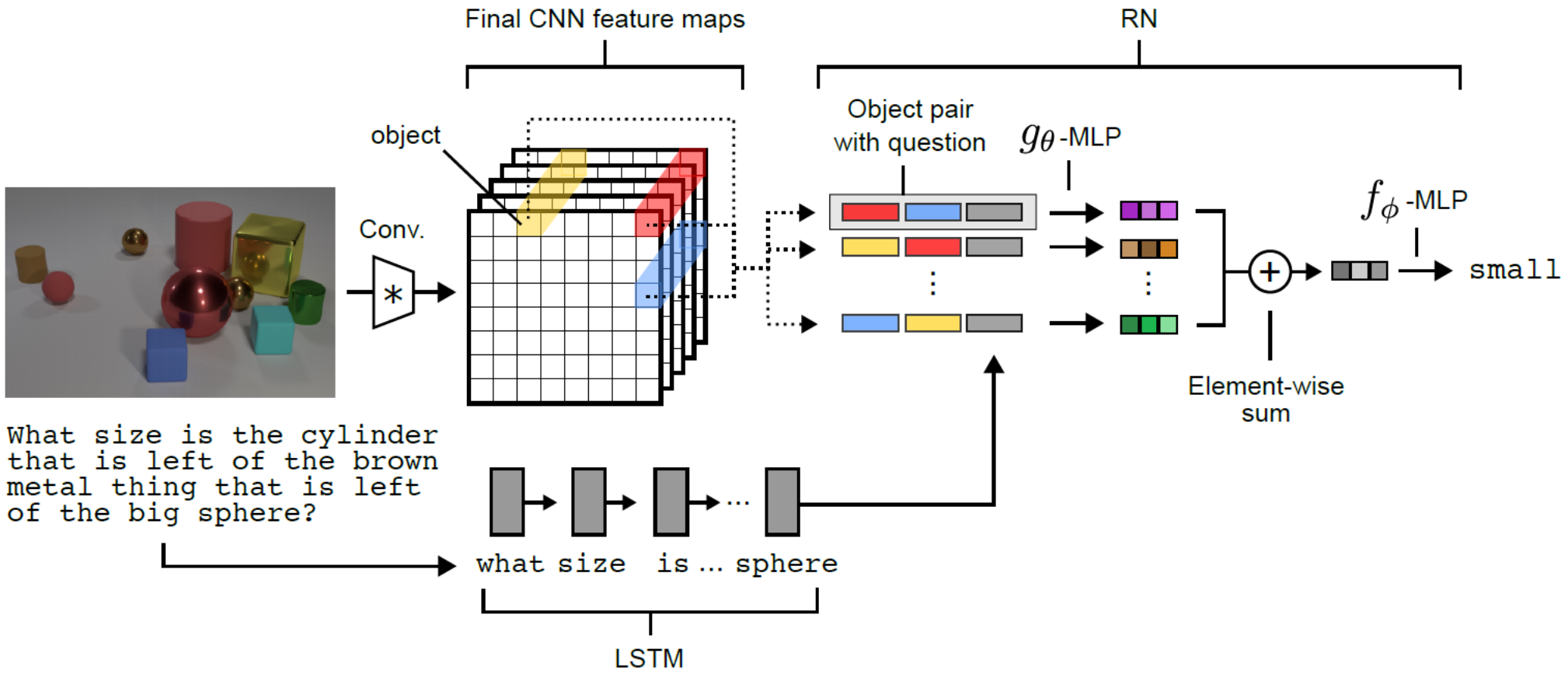
提出更具挑战的benchmark。当时关于VQA任务还只是刚刚兴起，并且当时的研究只敢做相对简单的识别问题。数据集中包含的概念都是很相似的。该方法在相对较困难的数据集上做，解决一般化问题。

Image question answering using convolutional neural network with dynamic parameter prediction. In CVPR, 2016.

融合方法

Relationship方法

任务：VQA



A simple neural network module for relational reasoning. In NIPS, 2017.

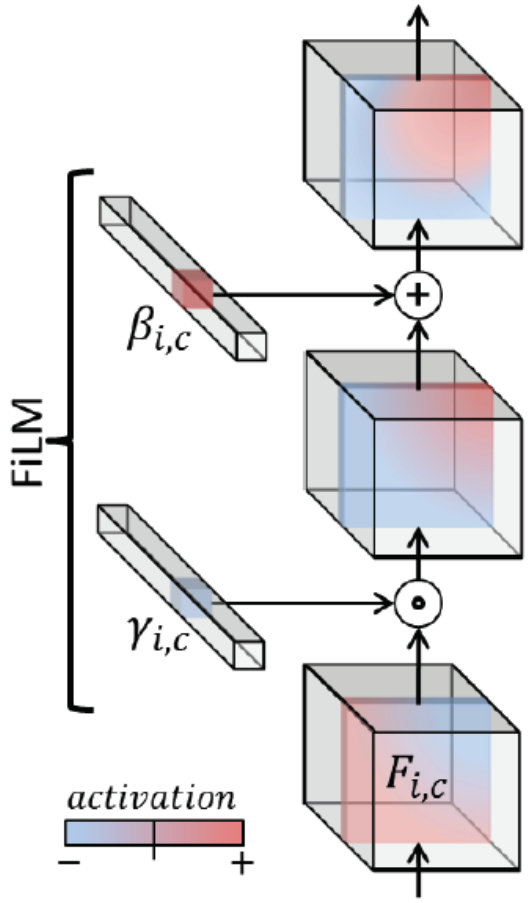
融合方法

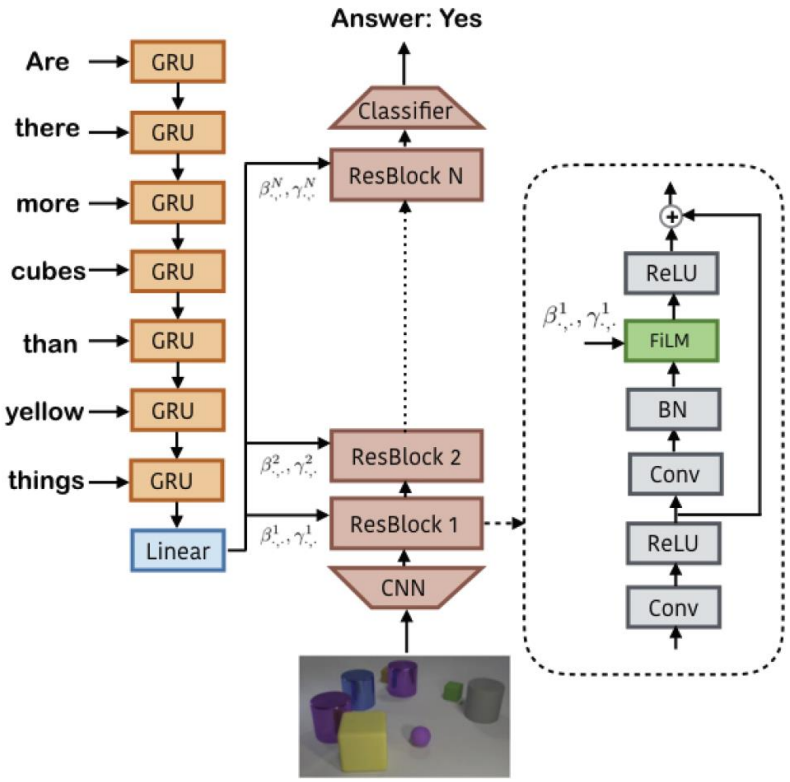
FiLM (Feature-wise Linear Modulation)方法

任务：VQA

$$\gamma_{i,c} = f_c(\boldsymbol{x}_i) \qquad \beta_{i,c} = h_c(\boldsymbol{x}_i)$$

$$FiLM(\boldsymbol{F}_{i,c}|\gamma_{i,c}, \beta_{i,c}) = \gamma_{i,c}\boldsymbol{F}_{i,c} + \beta_{i,c}.$$



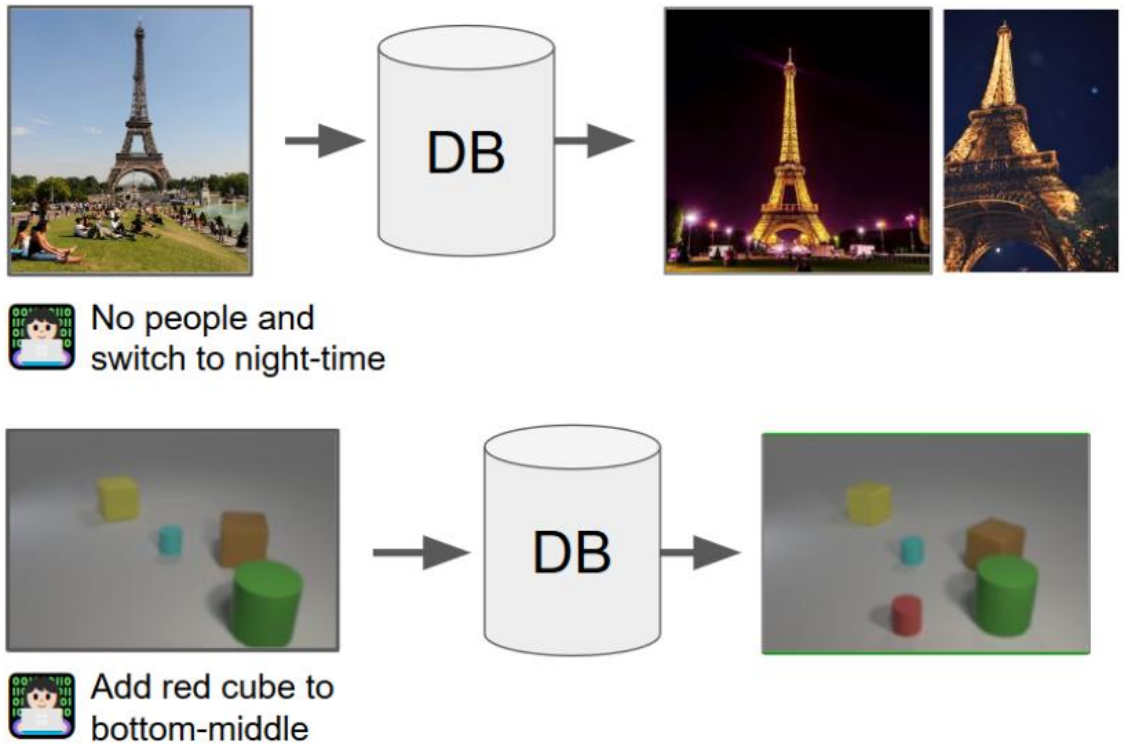


Film: Visual reasoning with a general conditioning layer. In AAAI ,2018.

融合方法

TIRG (Text Image Residual Gating)方法

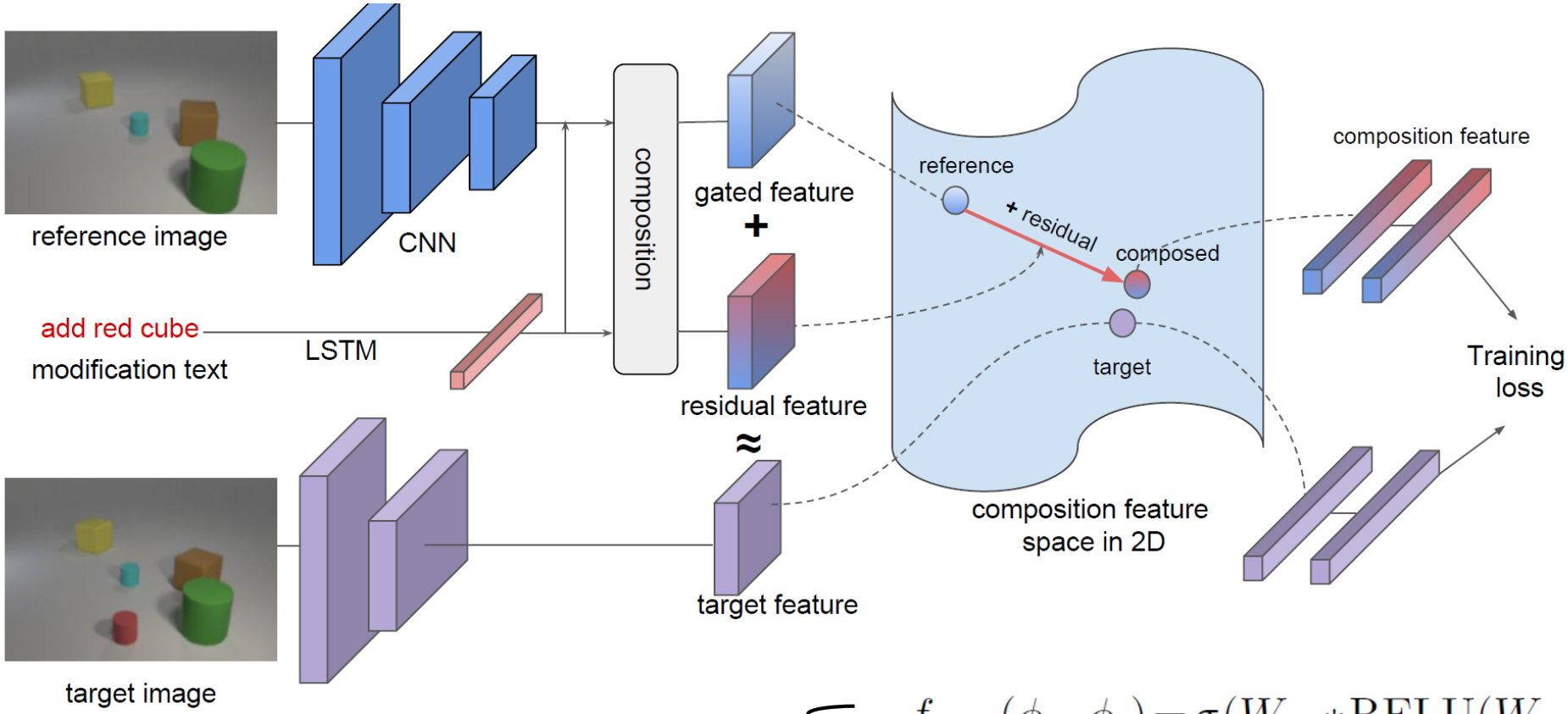
任务：Image Retrieval



Composing Text and Image for Image Retrieval - An Empirical Odyssey. In CVPR , 2019.

融合方法

TIRG (Text Image Residual Gating)方法



$$\phi_{xt}^{rg} = w_g f_{\text{gate}}(\phi_x, \phi_t) + w_r f_{\text{res}}(\phi_x, \phi_t)$$

$$f_{\text{gate}}(\phi_x, \phi_t) = \sigma(W_{g2} * \text{RELU}(W_{g1} * [\phi_x, \phi_t]) \odot \phi_x$$

$$f_{\text{res}}(\phi_x, \phi_t) = W_{r2} * \text{RELU}(W_{r1} * ([\phi_x, \phi_t]))$$

Composing Text and Image for Image Retrieval - An Empirical Odyssey. In CVPR ,2019.