ImageCaption

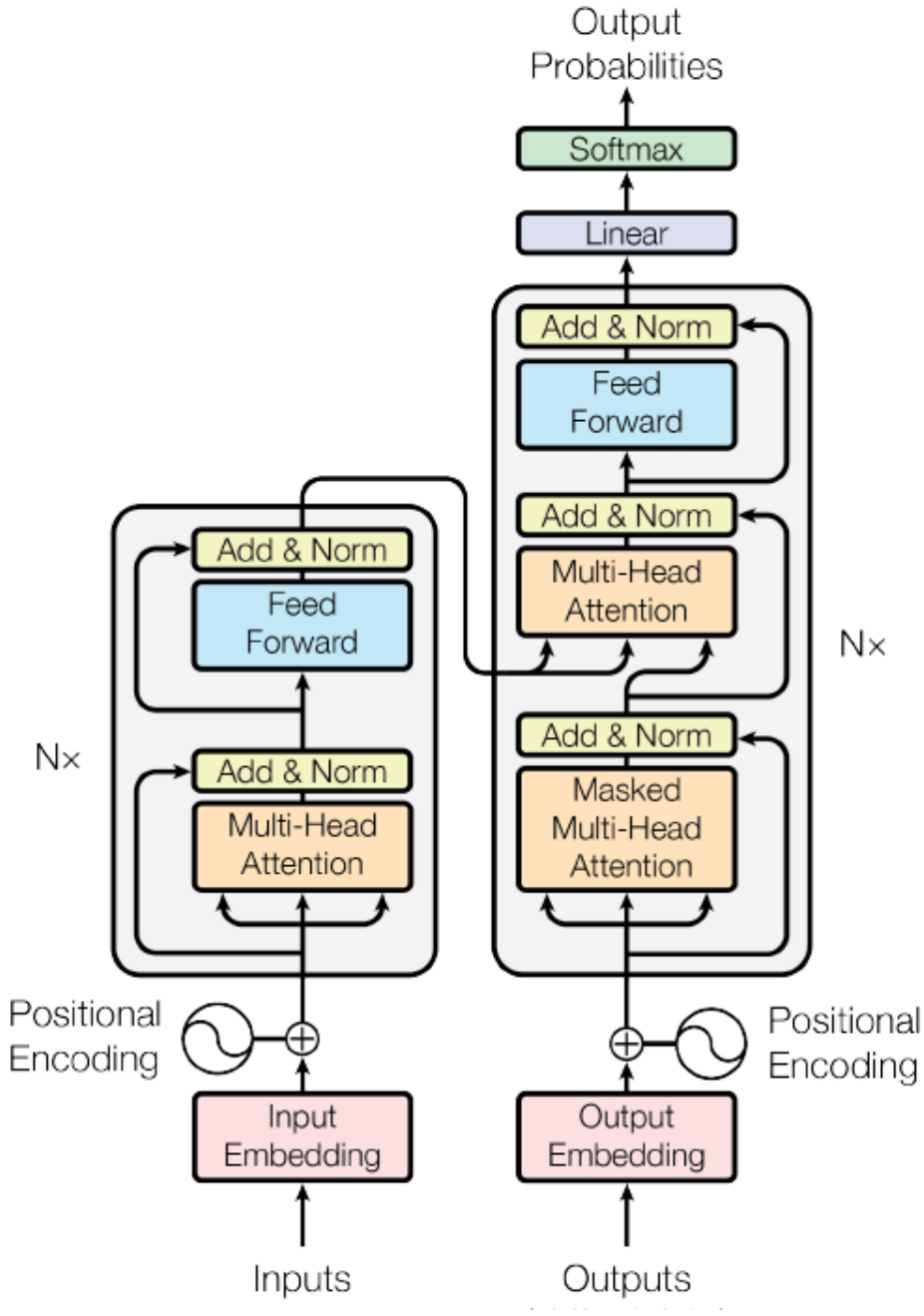
模块

encoder decoder

不用encoder 和 decoder时，会有N to M的句子不对等问题；encoder 和 decoder直接引入意义单元，来解决这个问题

Transformer

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
N×
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
N×
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs

(shifted right)

payless

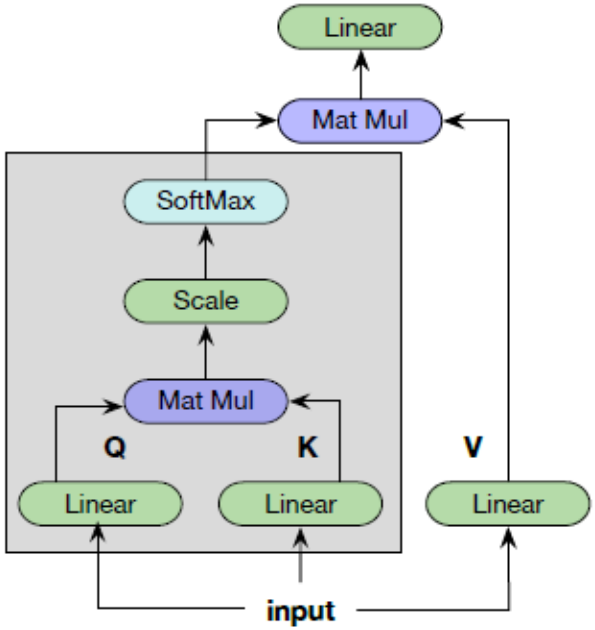
使用了GLU

$$h(x) = (W_0 x + b_0) \otimes \sigma(W_1 x + b_q)$$

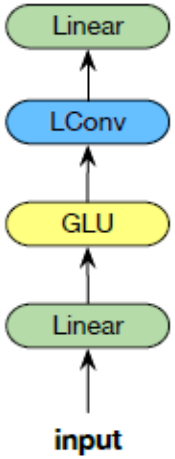
使用深度卷积网络，参数量从d^2k降低到dk；在使用参数共享，将参数量降低到Hk;

对共享的参数，在k维度上用softmax进行normalization

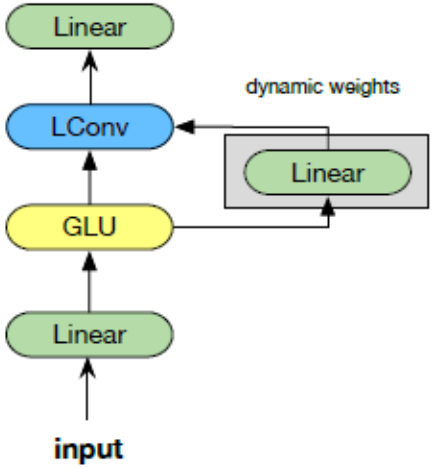
$$\mathrm{softmax}(W)_{h,j} = \frac{\exp W_{h,j}}{\sum_{j'=1}^{k} \exp W_{h,j'}}$$



(a) Self-attention



(b) Lightweight convolution



(c) Dynamic convolution

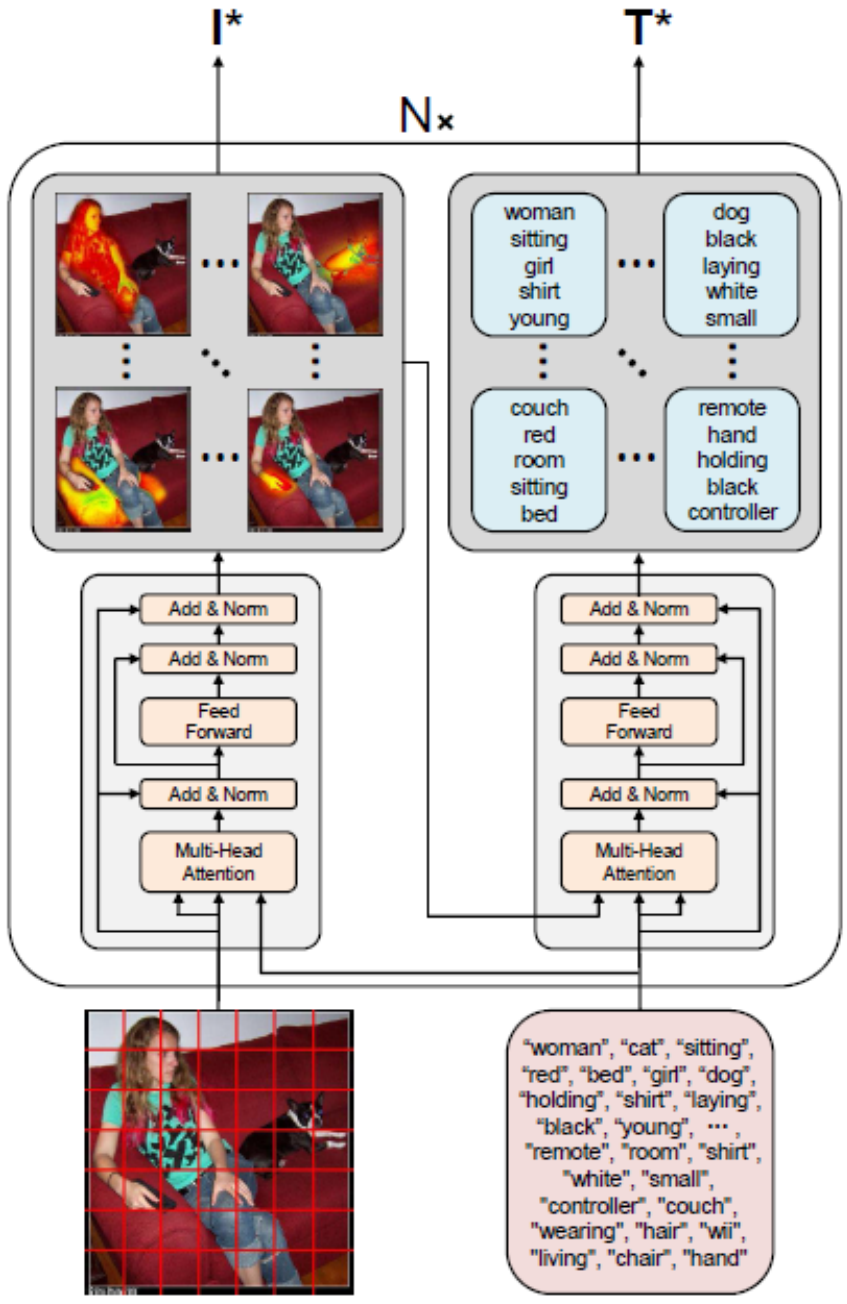
MIA

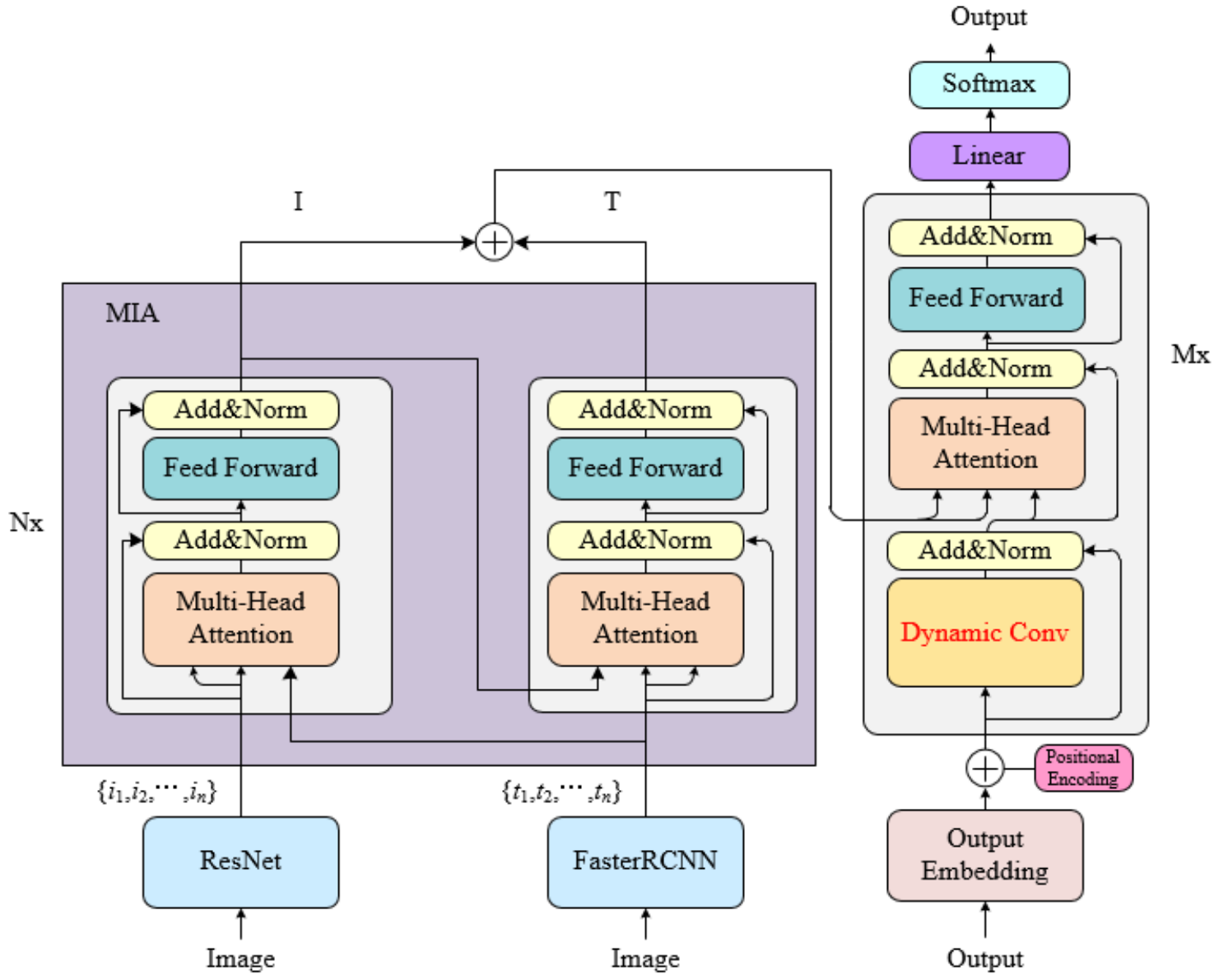
Aligning Visual Regions and Textual Concepts: Learning Fine-Grained Image Representations for Image Captioning

此方法在encoder端使用，对图像特征进行融合;

输入端的特征融合方式：

- 将不同模型提取的特征进行融合【我们的使用】
- 将模型和文本进行融合【原文】
- ~~也可以使用feature和bounding box~~


I*
T*
N×
woman
sitting
girl
shirt
young
dog
black
laying
white
small
couch
red
room
sitting
bed
remote
hand
holding
black
controller
Add & Norm
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Add & Norm
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
"woman", "cat", "sitting",
"red", "bed", "girl", "dog",
"holding", "shirt", "laying",
"black", "young", ···,
"remote", "room", "shirt",
"white", "small",
"controller", "couch",
"wearing", "hair", "wii",
"living", "chair", "hand"



MIA Transformer

每一层中:

$$\mathrm{I}_1 = \mathrm{FFN}(\mathrm{MultiHead}\left(\mathrm{T}_0, \mathrm{I}_0, \mathrm{I}_0\right))$$

$$\mathrm{T}_1 = \mathrm{FFN}(\mathrm{MultiHead}(\mathrm{I}_1, \mathrm{T}_0, \mathrm{T}_0))$$

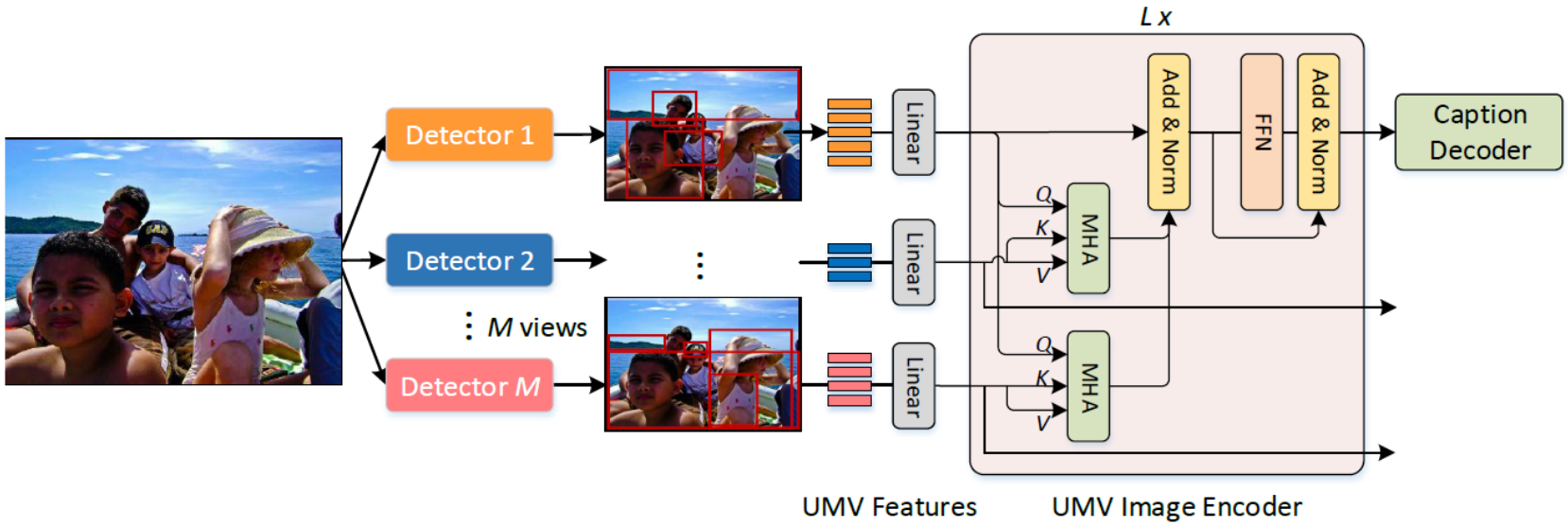
UMV

Multimodal Transformer with Multi-View Visual Representation for Image Captioning

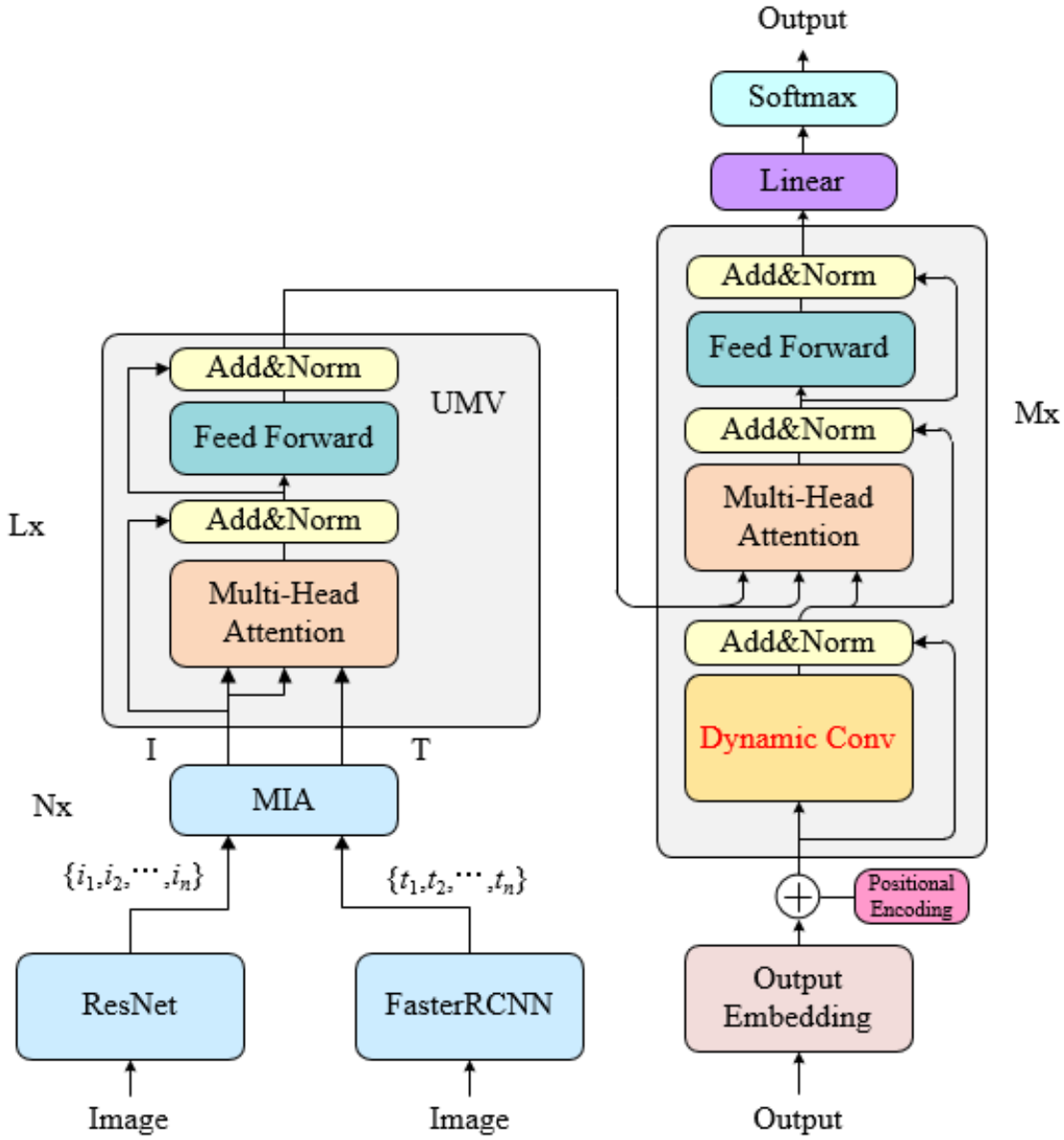
多个目标检测器结果，进行融合。

F_1 为主要视图，作为MHA的Q

$$\tilde{F_{(i)}} = MHA_{(i)}(F_{(1)}, F_{(i)}, F_{(i)})$$



在MIA模型中，数据进行提炼之后，feature的融合只是简单的进行加和，融合较简单，故引入UMV模块。

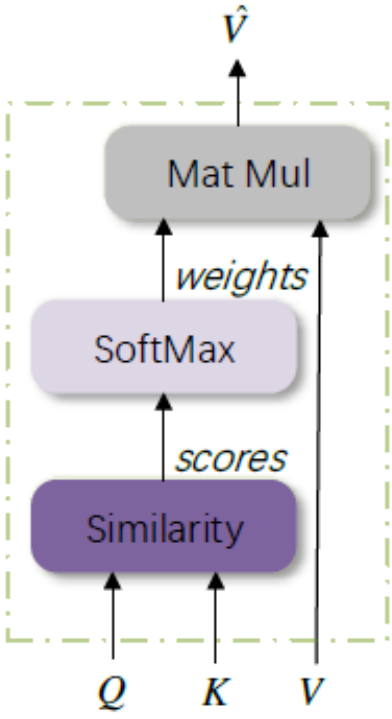


MIA_UMV Transformer

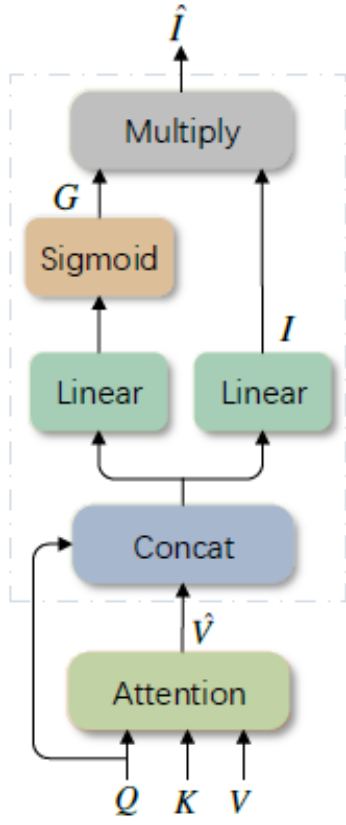
AOA

Attention on Attention for Image Captioning

将Q和attention的结果进行concat之后，送入GLU模块，输出结果替代作为attention的结果。

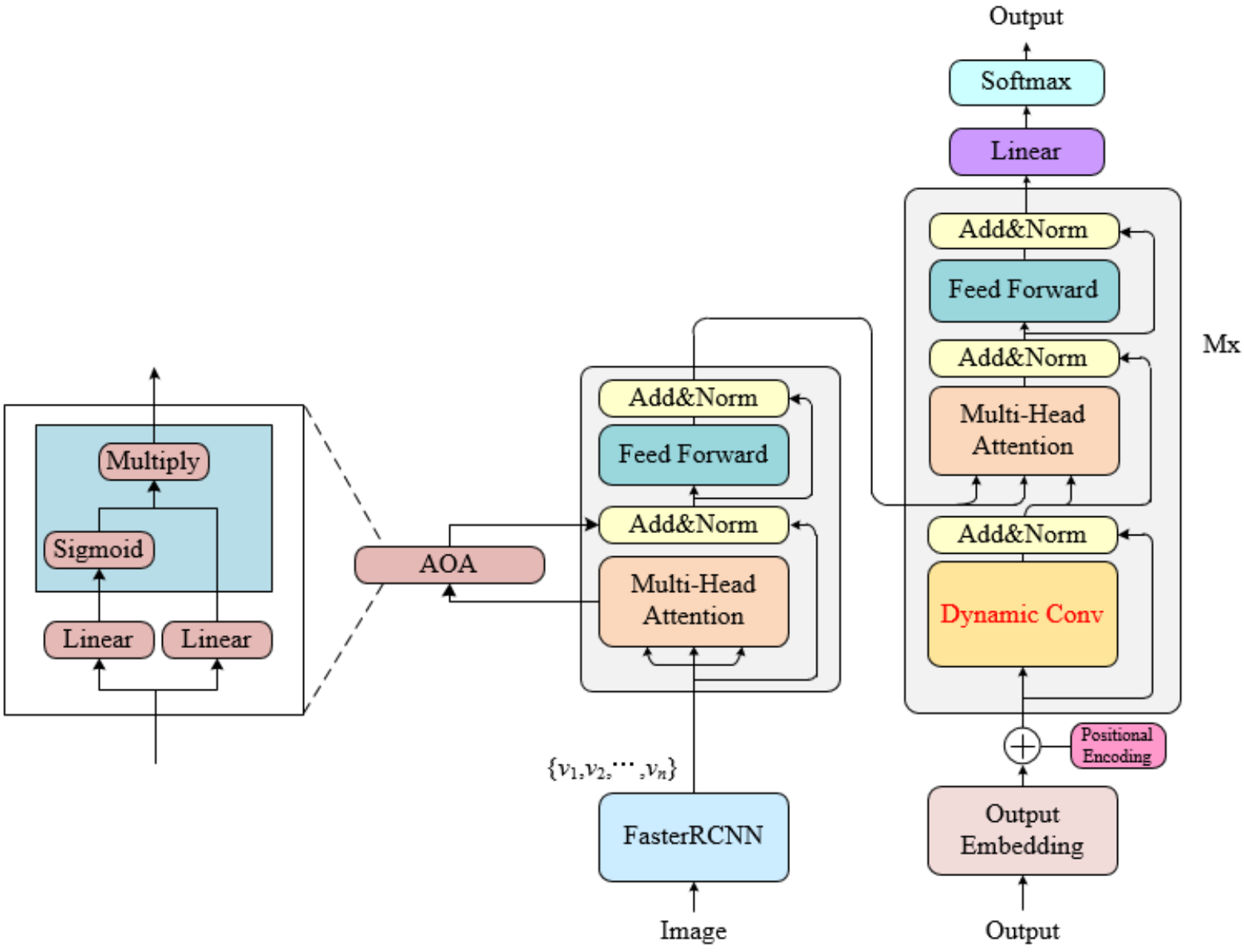


(a) Attention



(b) Attention on Attention

内部实现使用GLU门控单元。



AOA Transformer