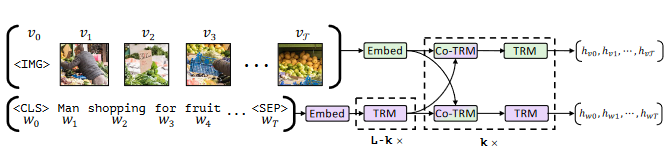
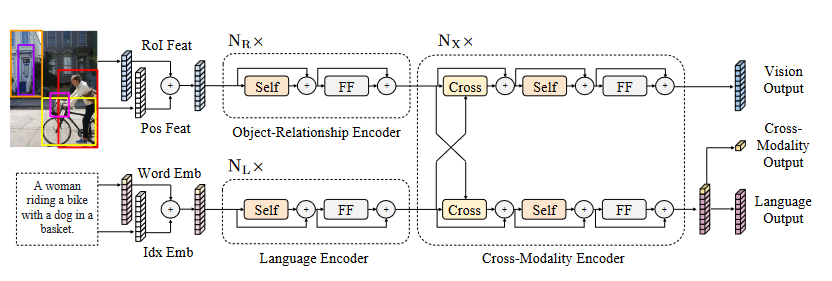
**MODELS’ IMPLEMENTATION DETAILS**

**ViLBERT**

****

* **Architectural details**
  + **BERT language stream:** Bert-base-uncased [12 layers]
    - **Self-attention’s dimensions:** 768
    - **Intermediate feed forward layer’s dimensions:** 3072
  + **BERT visual stream:** [6 layers]
    - **Self-attention’s dimensions:** 1024
    - **Intermediate feed forward layer’s dimensions:** 1024
  + **ViLBERT co-attention:** [6 layers]
    - **Intermediate dimension:** 1024
  + **Final MLP:**
    - **Hidden layer dimensions:** 2048
* self.ImageClassifFC = nn.Sequential(
* nn.Linear(self.config.bi\_hidden\_size, self.config.bi\_hidden\_size \* 2),
* GeLU(),
* BertLayerNorm(self.config.bi\_hidden\_size \* 2 , eps=1e-12),
* nn.Linear(self.config.bi\_hidden\_size \* 2 , num\_labels),
* )
* **Adaptation mechanisms:**
  + **Image classification:** The final *multi-modal pooled representation* ***of the visual stream*** (vector of length 1024) is connected to a linear layer of 2048 neurons with GeLU activation function and BertLayerNorm. A final output layer of **size = n classes** is treated as the logits.
* **Hyperparameter choice:**
  + **Image classification**
    - **Criterion:** Cross-entropy loss
    - **Optimizer:** AdamW
      * **Base learning rate:** 0.00002
      * **Learning rate scheduler:** Reduce learning rate on plateau
    - **Number of training epochs:**
    - **Batch size:**

**LXMERT**

****

* **Architectural details**
  + **BERT language stream:** Bert-base-uncased [9 layers]
    - **Self-attention’s dimensions:** 768
    - **Intermediate feed forward layer’s dimensions:** 3072
  + **BERT visual stream:** [5 layers]
    - **Self-attention’s dimensions:** 768
    - **Intermediate feed forward layer’s dimensions:** 3072
  + **ViLBERT co-attention:** [5 layers]
    - **Intermediate dimension:** 768
    - **Intermediate feed forward layer’s dimensions:** 3072
  + **Final MLP:**
    - **Hidden layer dimensions:** 1536

        self.logit\_fc = nn.Sequential(

            nn.Linear(hid\_dim, hid\_dim \* 2),

            GeLU(),

            BertLayerNorm(hid\_dim \* 2, eps=1e-12),

            nn.Linear(hid\_dim \* 2, num\_classes)

* **Adaptation mechanisms:**
  + **Image classification:** The final *multi-modal pooled representation* ***of the full input*** (vector of length 768) is connected to a linear layer of 1536 neurons with GeLU activation function and BertLayerNorm. A final output layer of **size = n classes** is treated as the logits.
* **Hyperparameter choice:**
  + **Image classification**
    - **Criterion:** Cross-entropy loss
    - **Optimizer:** AdamW
      * **Base learning rate:** 0.00002
      * **Learning rate scheduler:**
        + Reduce learning rate on plateau.
        + Warmup linear schedule.
    - **Number of training epochs:**
    - **Batch size:**