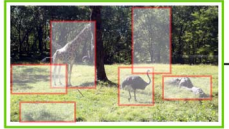

Input Image

Faster R-CNN

FC

Self-Attention Module

$\{r_1, r_2, ..., r_k\}$

Transformer

Pool

img_emb: i_0

Transformer

Cross-Attention Module

Pool

img_emb: i_1

$\{r_{c1}, ..., r_{ck}, e_{c1}, ..., e_{cn}\}$

1d CNN & Pool

cap_emb: c_1

Input Sentence

Giraffes and ostriches are sharing a grassy area.

{ '[CLS]', 'Giraffes', 'and', 'ostriches', 'are', 'sharing', 'a', 'grassy', 'area, '.', '[SEP]' }

Token Embedding + Position Embedding + Segment Embedding

Tranformer 12 layers (BERT)

$\{e_1, e_2, ..., e_n\}$

Self-Attention Module

1d CNN

Pool

cap_emb: c_0