ch14: Computer Vision

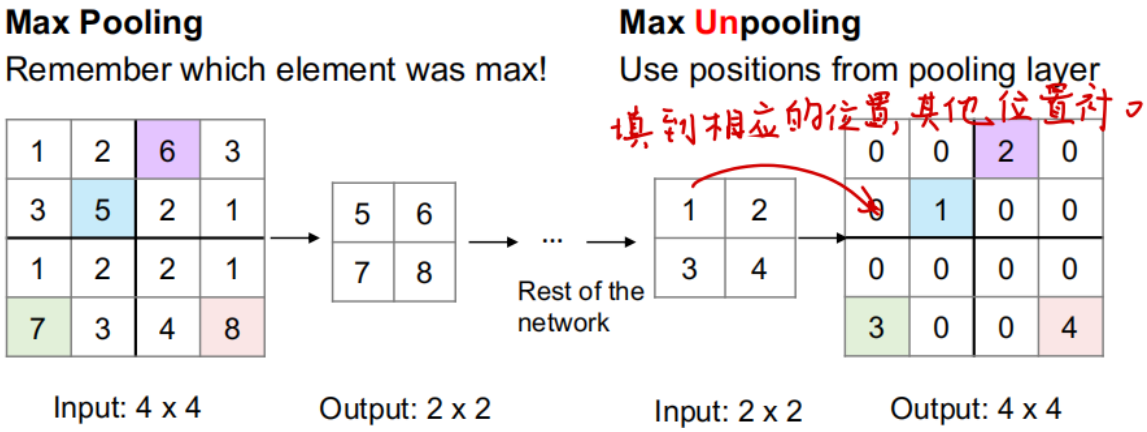
Image Segmentation

naive idea及问题（P5-P7）

- **Sliding Window**
- **Fully Convolutional**

Upsampling（P9-12）

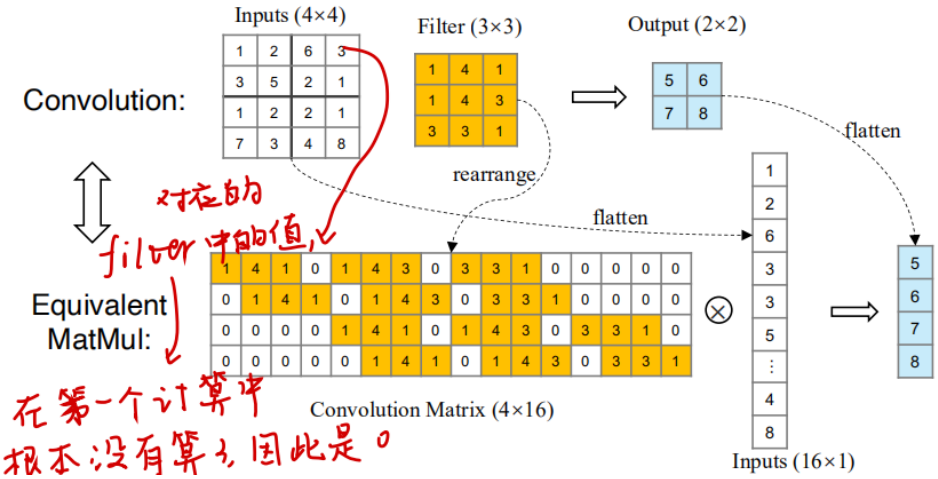
- **Unpooling**



- **Transposed Convolution**

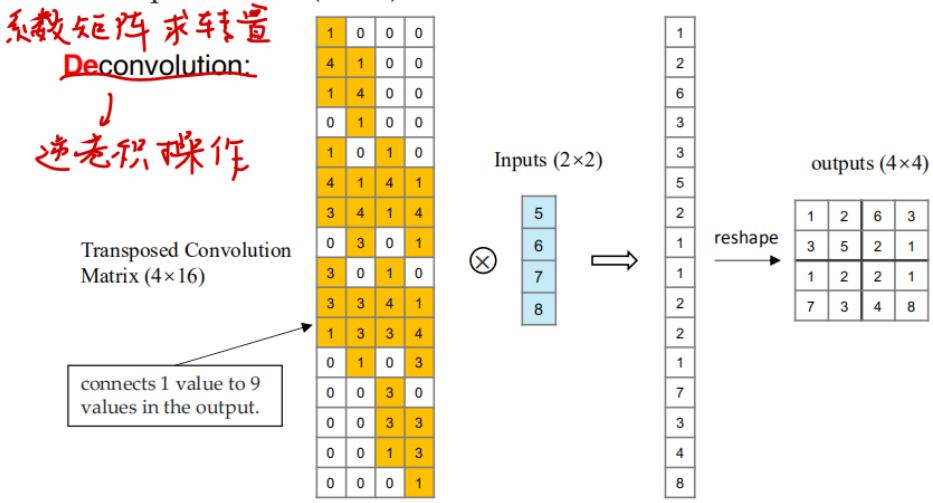
- **Convolution Matrix**

  We can express a convolution operation using a **matrix multiplication** by rearranging the kernel matrix into a general matrix.

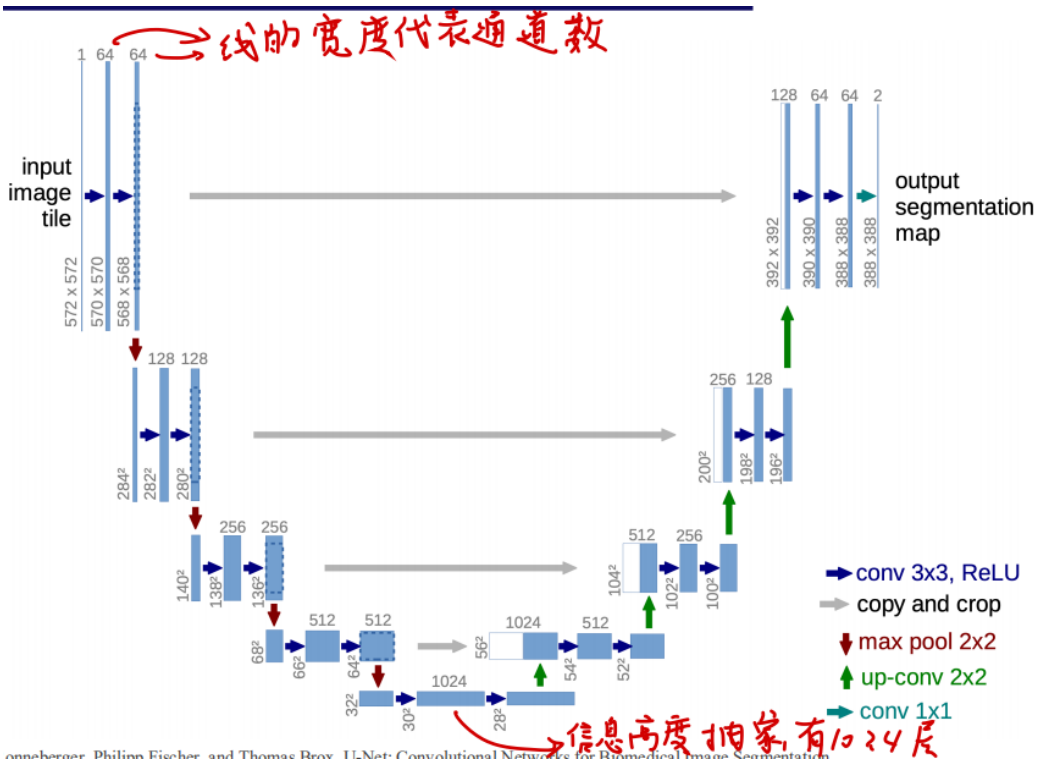


- **Transposed Convolution Matrix**

  Transpose the convolution matrix $\boldsymbol{C}$ (4×16) to $\boldsymbol{C}^{\mathrm{T}}$ (16×4).
  We can matrix-multiply $\boldsymbol{C}^{\mathrm{T}}$ (16×4) with a column vector (4×1) to generate an output matrix (16×1).



U-Net（P15）

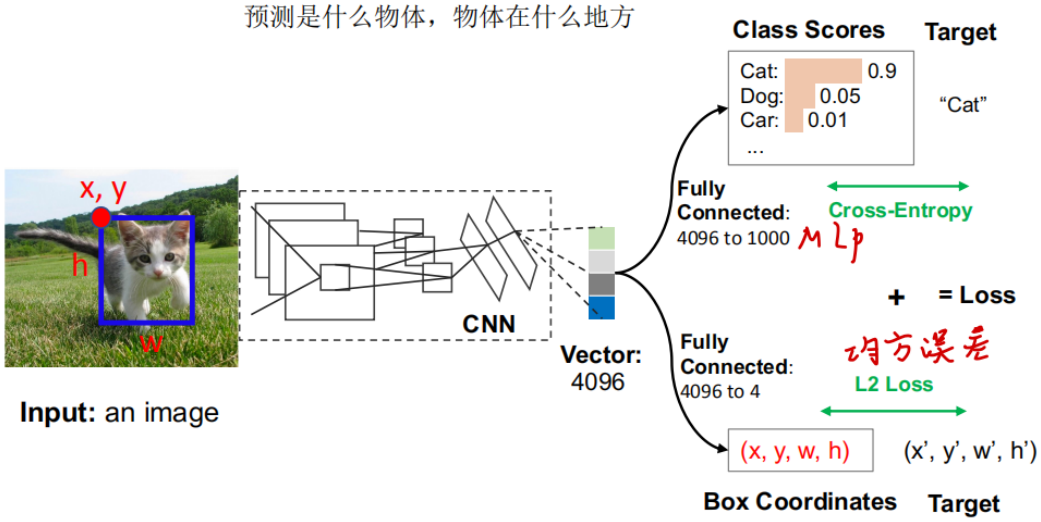


Object Detection

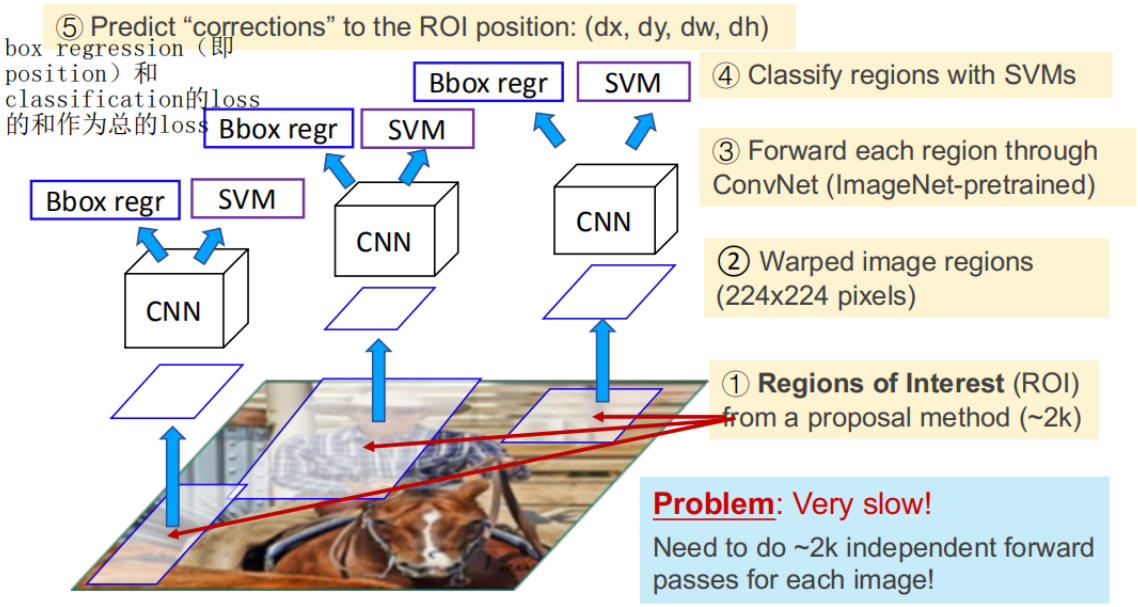
Single Object（P19）

- **Object detection = Classification + Localization**

预测是什么物体，物体在什么地方

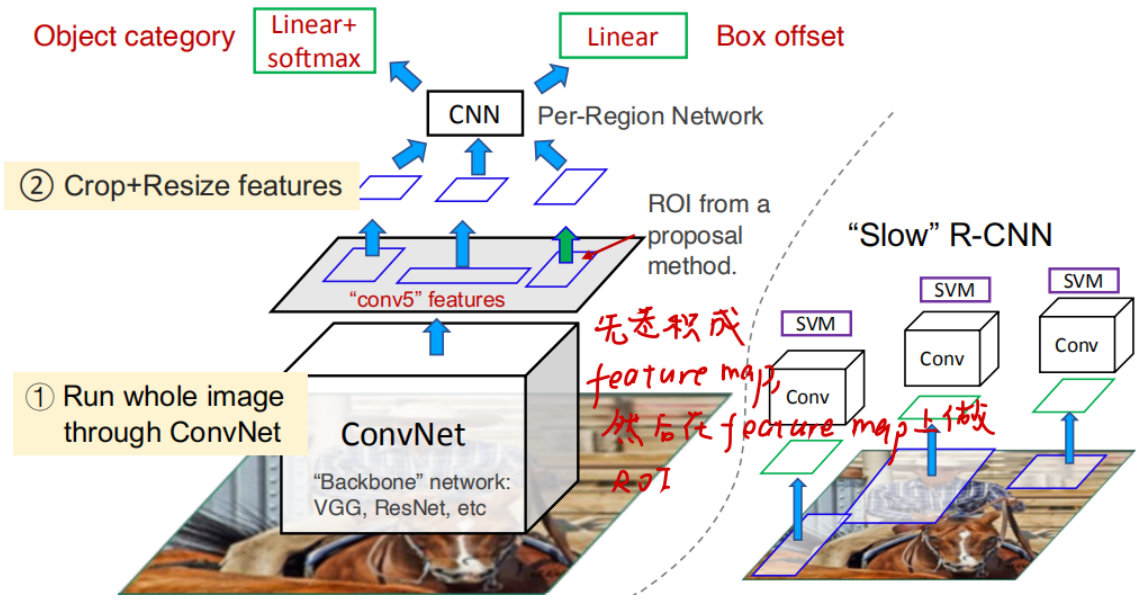


R-CNN（P22）



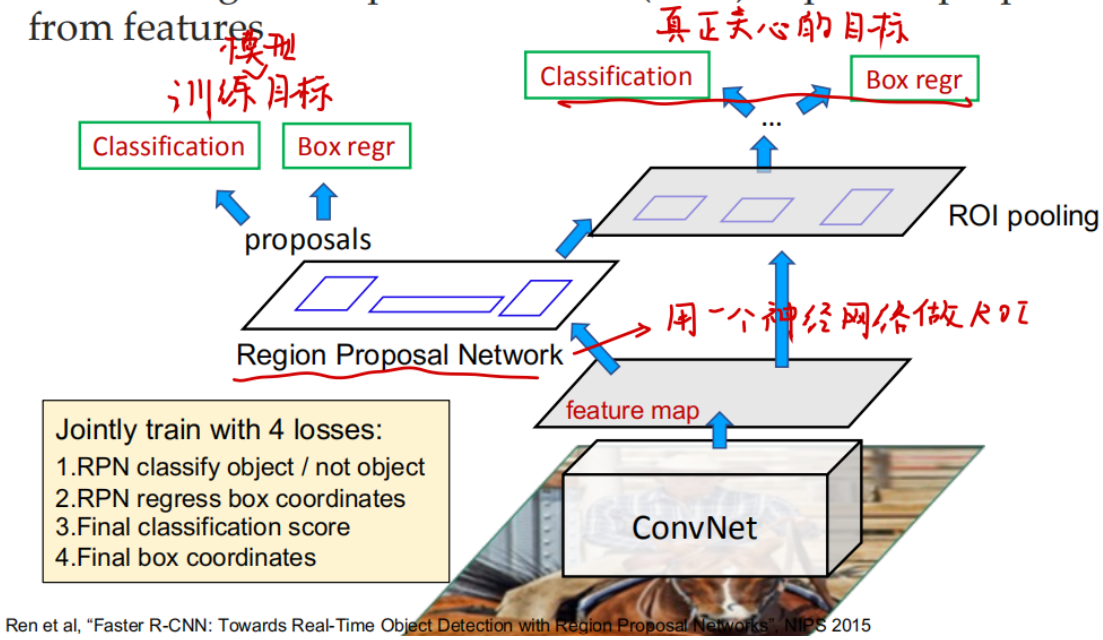
Fast R-CNN（P23）

- **Idea:** Pass the image through convnet before cropping!



Faster R-CNN（P24）

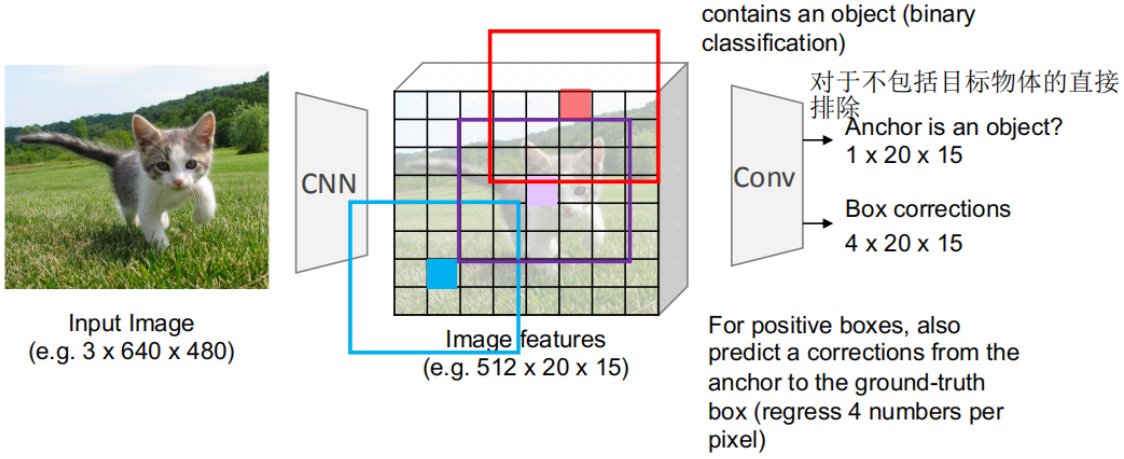
- Insert a Region Proposal Network (RPN) to predict proposals from features



Ren et al, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NIPS 2015

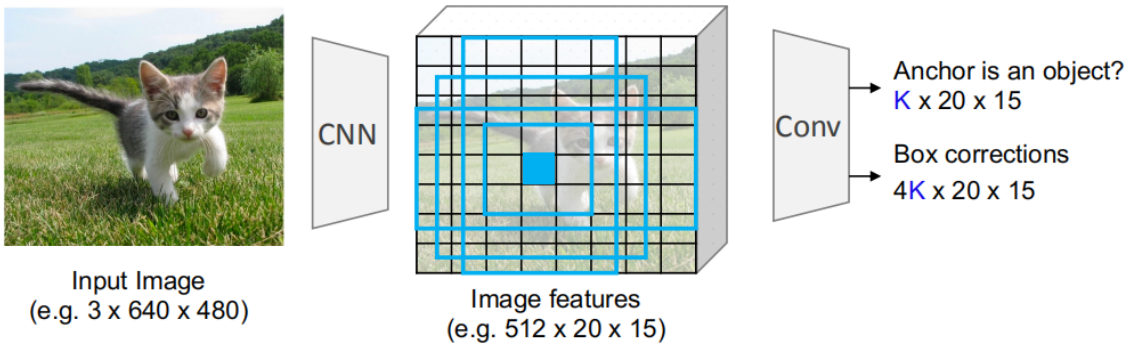
Region Proposal Network（P25-26）

- Imagine an **anchor box** of fixed size at each point in the feature map



- In practice use K different anchor boxes of different size / scale at each point

每个关注像素点周围有K个大小不同的box，选择不同的范围

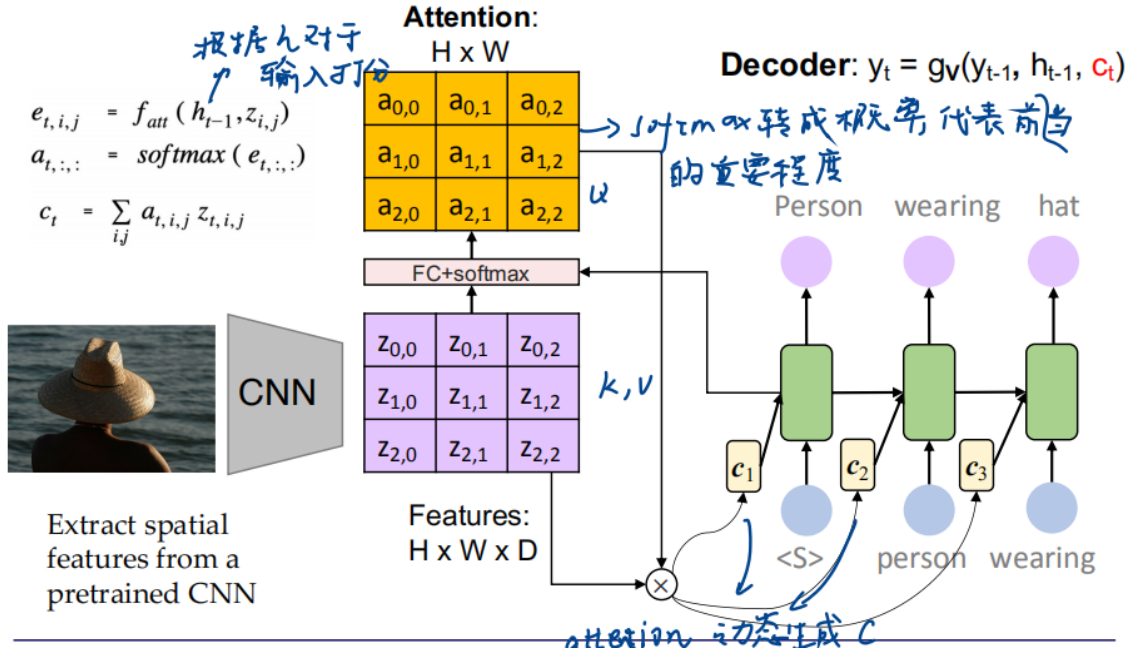


- Sort the K*20*15 boxes by their “objectness” score, take top ~300 as our proposals

Image Captioning

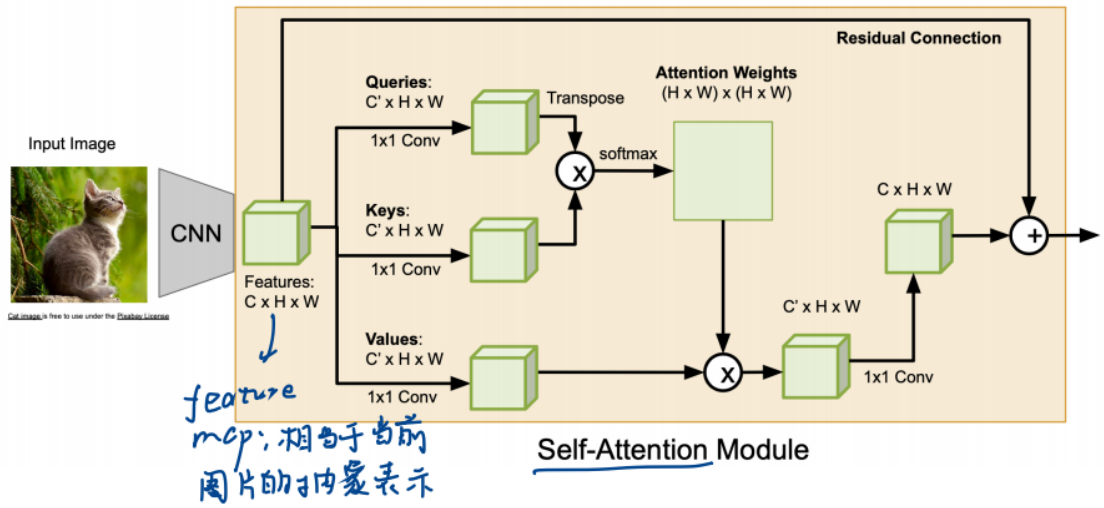
Encoder-Decoder with Attention（P31）

- Each context vector will attend to different image regions

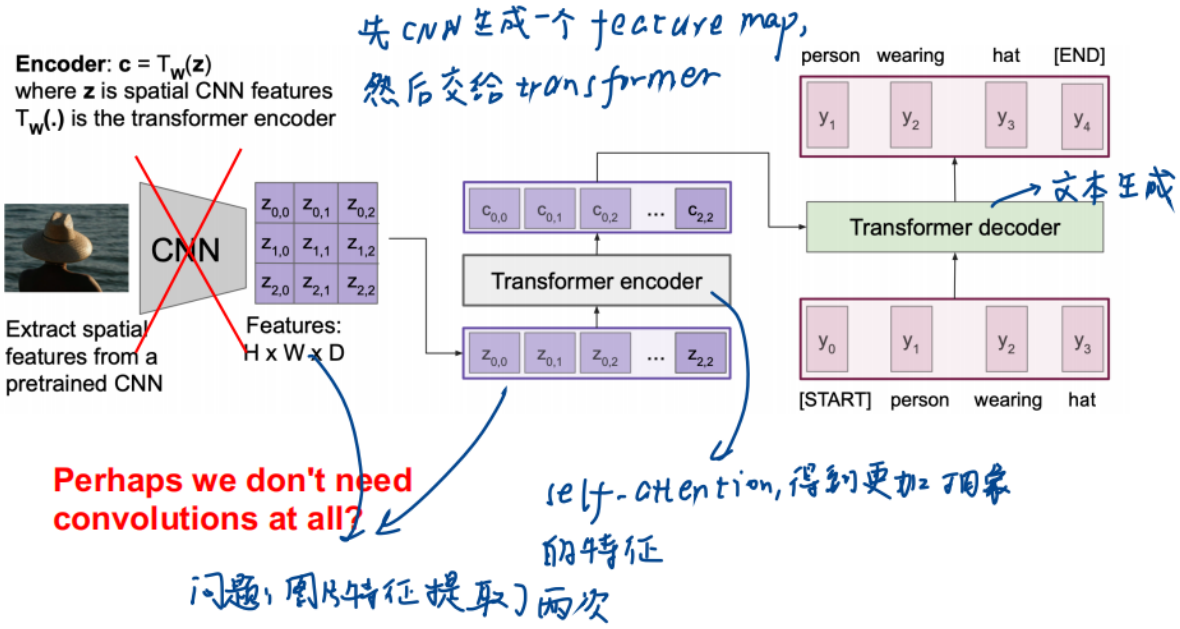


CNN with Self-Attention（P33-34）

- We can also apply self-attention to CNN features.



feature map：相当于当前图片的抽象表示



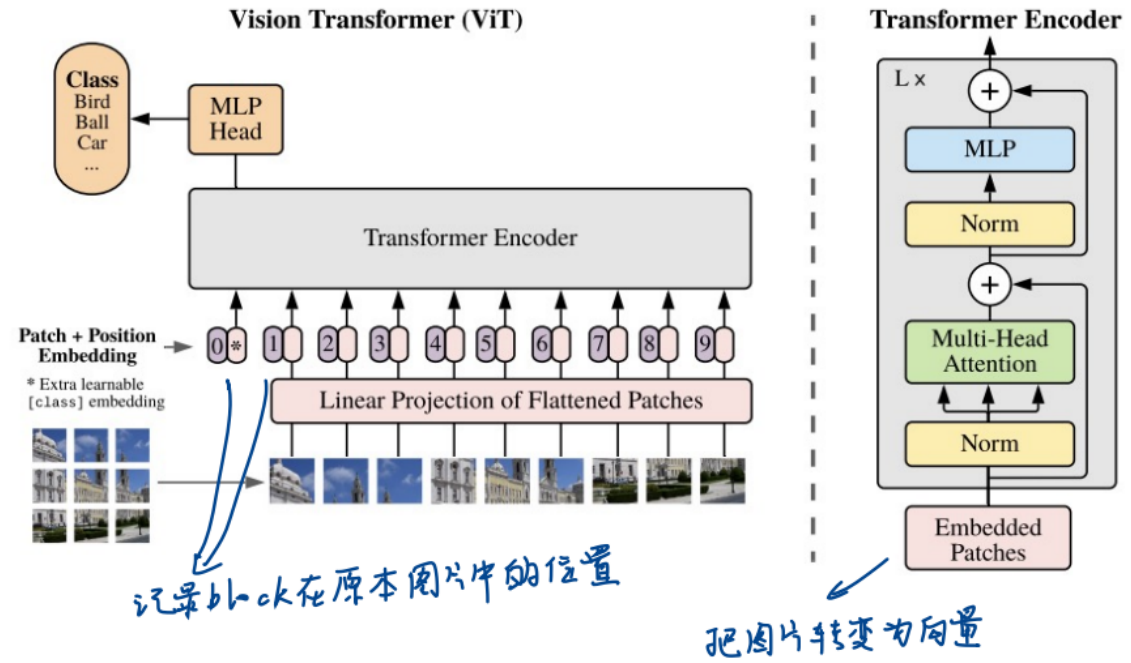
先 CNN 生成一个 feature map，然后交给 transformer

文本生成

self-attention，得到更加抽象的特征

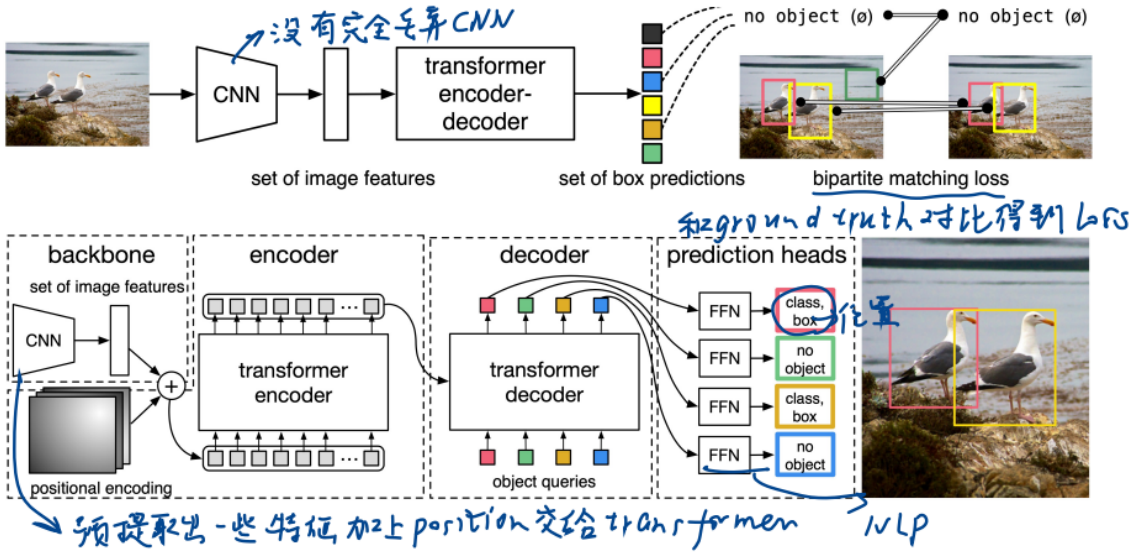
问题：图片特征提取了两次

Vision Transformer (ViT)（P36）



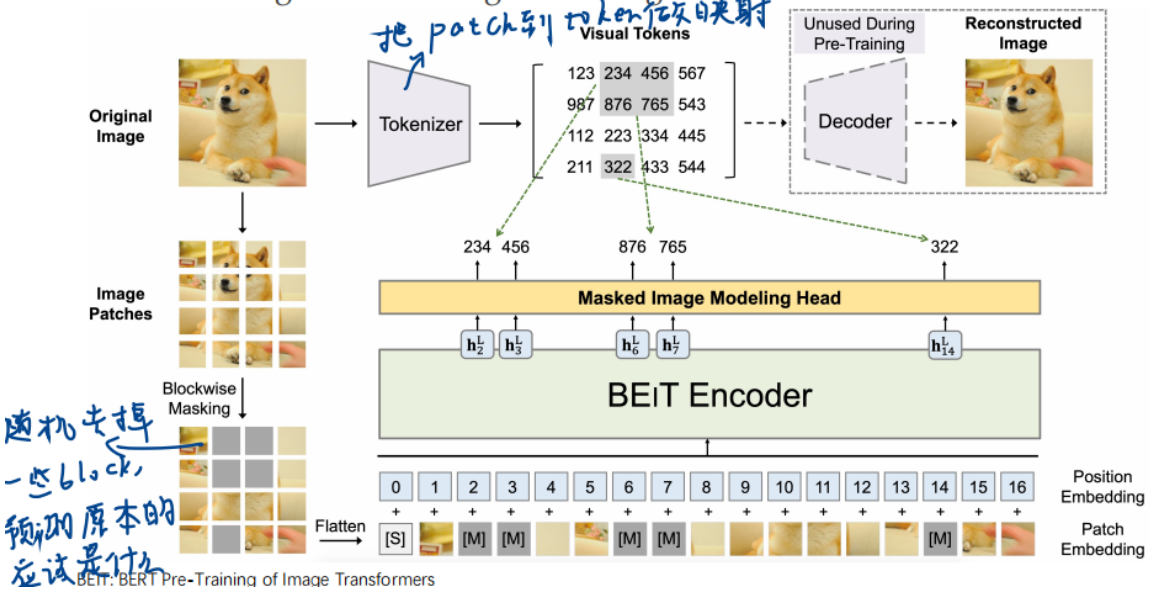
Vision Transformers for Object Detection（P36）

- DE:TR



BEIT: Pre-Trained Models for Images（P39）

- Backbone Network: Standard Transformer
- Pre-Training: Masked Image Modeling



BEIT: BERT Pre-Training of Image Transformers