Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks

VLMo也是出自这个团队，极大地继承了VLMo、BeiT、BeiT-2。

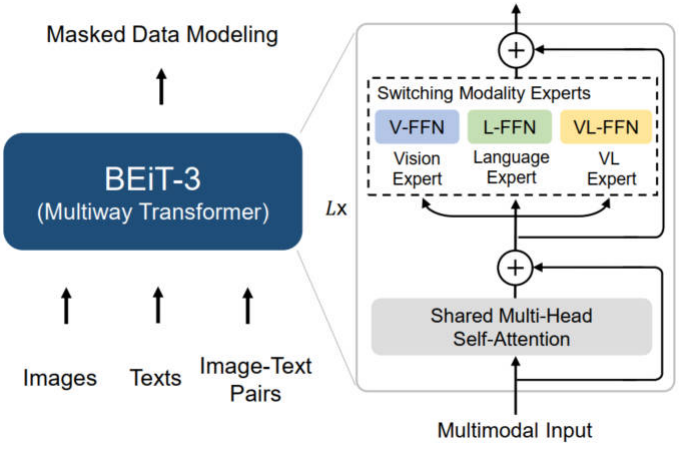
动机：

顺应一种趋势：“大一统”big convergence。即在1) backbone architecture层面，充分利用好Transformer；2) 在pre-training tasks层面，强调只使用一种objective loss，即掩码学习masked-then-predict；3) model scaling up层面，model parameters层面和dataset层面。

方法：

① 模型层面：general-purpose multimodal foundation model——BeiT-3

Backbone Network是Multiway Transformer，和VLMo如出一辙。



② 预训练目标层面：Masked Data Modeling

前提是将images和image-text pairs都当成文本来处理，具体如下，1)texts，就是English；2) images，记为imglish，视为一种语言； 3) image-text pairs，当作parallel sentences。具体的掩码学习如下，1) 对于English，采用和BERT一致的Masked Language Modeling（MLM）；2) 对于imglish，采用Masked Image Modeling（MIM），模仿BeiT中的掩码方

式；3) Masked Vision-Language Modeling（MVLM），模仿BeiT-2中的self-supervision fashion。

不用多种Loss Function，并且掩码学习这种方式对于训练模型很高效。

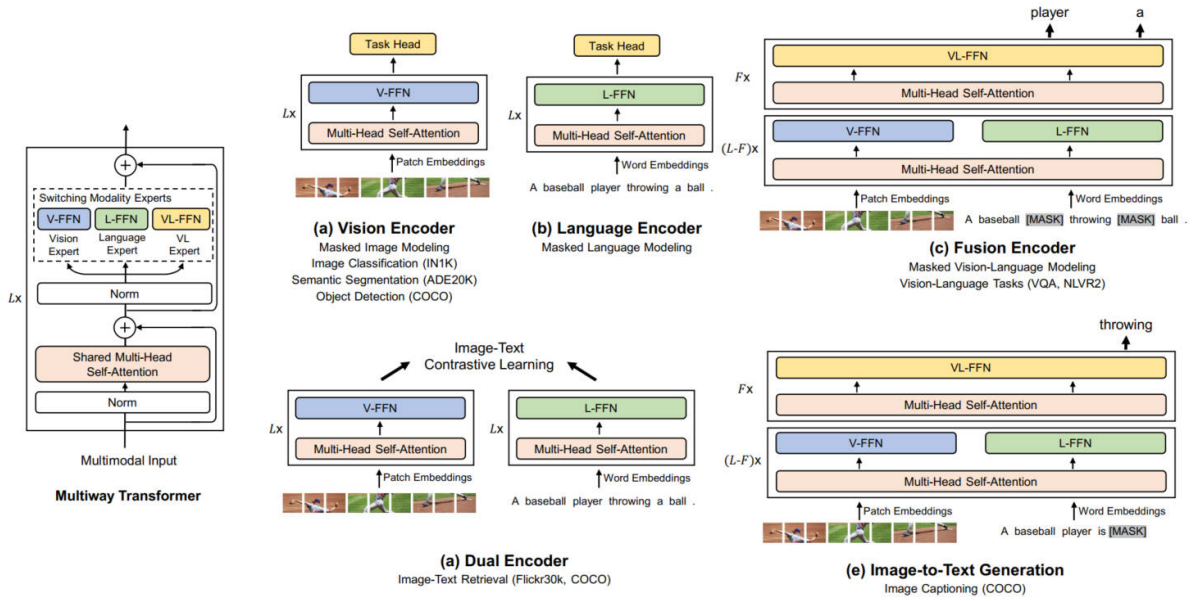
③ 规模扩大层面：Scaling Up

有两个方面，1) model parameters，达到了1.9billion； 2) dataset，使用更多的数据集。

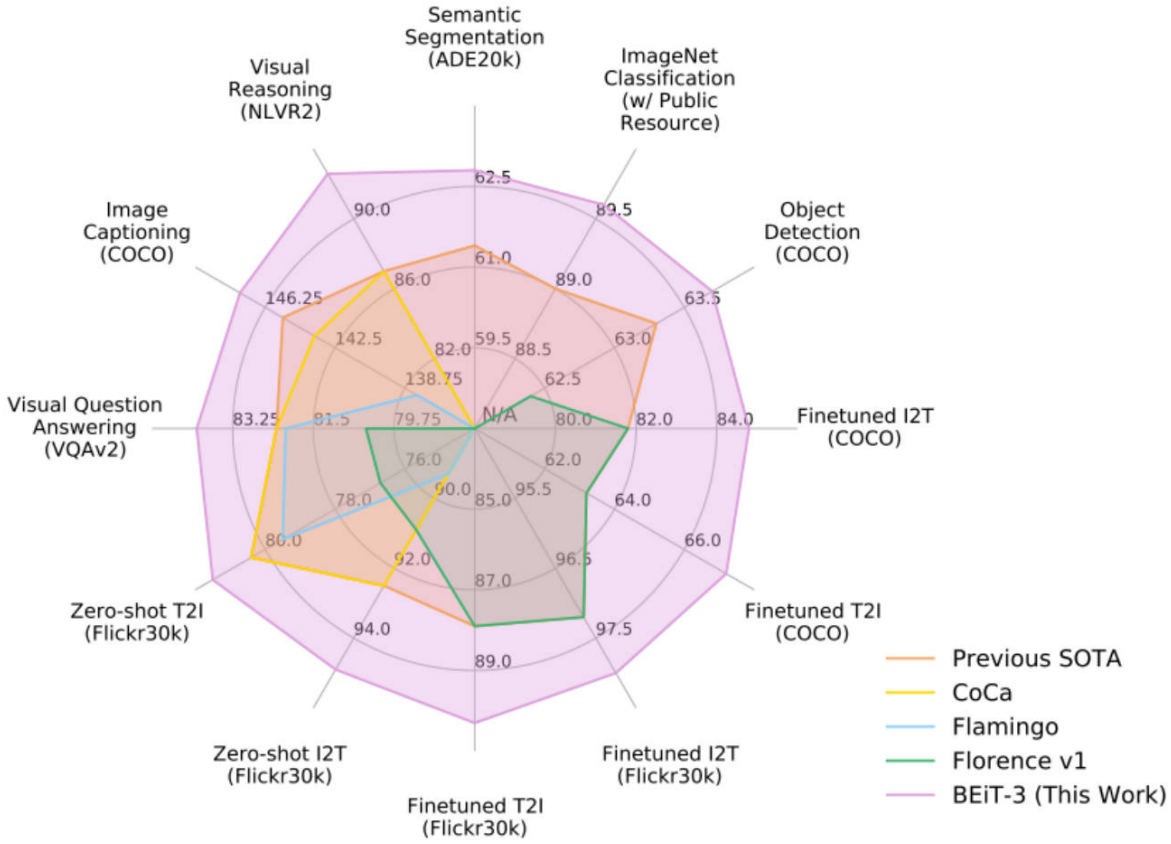
pre-training process：用256块40GB的A100训练两周。

性能：

能transfer到众多downstream tasks。



同时在很多downstream tasks取得了SOTA。


Semantic Segmentation (ADE20k)
ImageNet Classification (w/ Public Resource)
Object Detection (COCO)
Finetuned I2T (COCO)
Finetuned T2I (COCO)
Finetuned I2T (Flickr30k)
Finetuned T2I (Flickr30k)
Zero-shot I2T (Flickr30k)
Zero-shot T2I (Flickr30k)
Visual Question Answering (VQAv2)
Image Captioning (COCO)
Visual Reasoning (NLVR2)
62.5
61.0
59.5
N/A
89.5
89.0
88.5
63.5
63.0
62.5
84.0
82.0
80.0
66.0
64.0
62.0
97.5
96.5
95.5
94.0
92.0
90.0
90.0
86.0
82.0
79.75
83.25
81.5
78.0
76.0
146.25
142.5
138.75
Previous SOTA
CoCa
Flamingo
Florence v1
BEiT-3 (This Work)