基于预训练的多语言视觉问答

张良 2020000888

【基本介绍】

视觉问答是一项经典的视觉语言任务。给定一张图片以及一个根据图片内容提出的问题，视觉问答模型需要对该问题进行回答。因此，模型不但需要对图片和语言进行理解，还需要有一定的常识推理能力。目前已经有许多工作针对 VQA 进行，其中许多工作是在由人工标注的 VQA 数据集上训练的。由于人工标注比较昂贵，通常这些数据集虽然质量高，但是规模比较小，评测性能不佳。并且这些数据集通常是单语言的，无法做到对多种语言问题的回答。本项目采用预训练的方式，首先从大量的无标注数据中学习图片与语言基本的语义理解与常识。通过在预训练过程中加入机器翻译构造的多语言语料，模型能够经过少量微调就可以在英文和日文两个语言的 VQA 数据集上达到超过 SOTA 的性能。另外，对模型中注意力权重的可视化结果说明了模型在推理过程中能够关注到问题相关的区域。

【模型介绍】

预训练模型的结构是一个 Transformer encoder，如图 1 所示

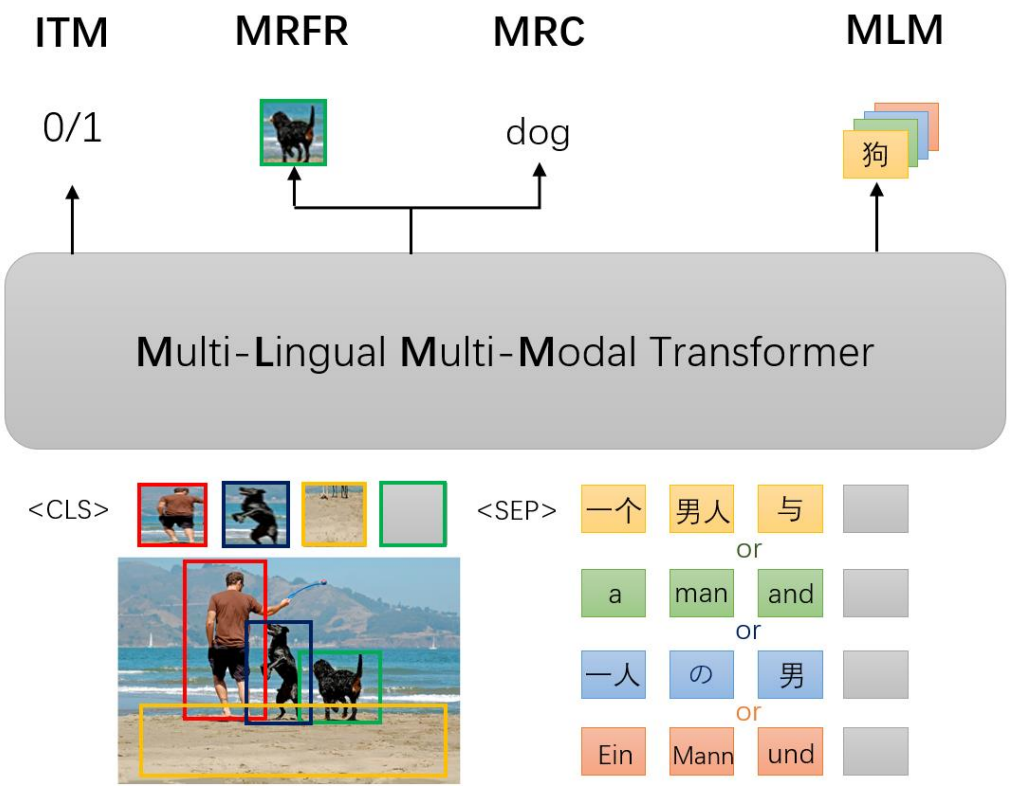

图 1 预训练模型

我们使用从网上爬取的无 VQA 相关标注的图片描述数据集进行预训练。对于图片，使用目标检测器提取图片中的可能存在的物体的区域以及他们的特征。将这些特征与文本特征和特殊标记<CLS>与分隔符<SEP>拼接起来，输入到 Transformer 模型中。对该 Transformer encoder 训练四个任务:

1. Image-Text Matching (ITM)

   从数据集 Sample50%图文不匹配的负样本，判断输入的图文是否匹配。

2. Masked Language Modeling (MLM)

   将 15%的 token 进行 mask，根据图片和上下文文本对

masked token 进行预测。

3. Masked Region Feature Regression (MRFR)

   将一部分区域进行 mask，根据文本与其他区域回归还原这个区域的特征

4. Masked Region Classification (MRC)

   将一部分区域进行 mask，根据文本与其他区域预测该区域的类别，使用 soft label

Finetune 模型

我们将 VQA 看作一个多标签分类任务，因此相比于预训练模型，只需要在 CLS 对应位置加入 MLP 进行 VQA 答案的预测，如图 2：

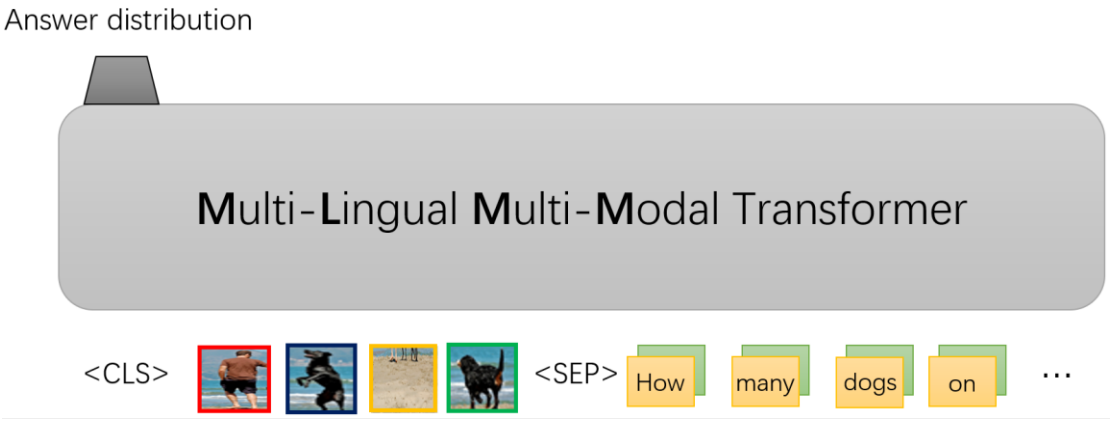

图 2 Finetune VQA 模型

为了最大限度地从预训练中获益，finetune 阶段采用和预训练阶段相同的词表，并且以相同的方式把图片区域特征和问题文本拼接起来，作为 transformer 的输入。在推理阶段，将 Answer distribution 中概率最大对应的答案视为 VQA 的答案。

【数据介绍】

采用两种语言上的 VQA 数据集进行微调和评测。

VQA2.0 是一个英文 VQA 数据集，是对 MSCOCO 数据集进行人工标注并进行平衡的数据集，包含近 20 万图片和 110 万问题-答案对。

图 3　VQA2.0 样例

VQA VG Japanese 是一个日文 VQA 数据集，包含 99,208 图片，793,664 日文问题-答案对。

【评测结果】

在 VQA2.0 和 VQA VG Japanese 两个数据集上评测的结果如表 1 所示。

数据集 / 模型	VQA2.0 test-dev	VQA VG Japanese
非预训练模型		
MCAN	70.63	-
PCATT	-	19.2
预训练模型		
UNITER	71.22	22.7
UC2	71.48	34.2
Ours	**73.21**	**35.4**

表 1　VQA 实验结果

从表 1 中的结果可以看出，我们的模型在两种语言的 VQA 数据集上的性能都显著犹豫非预训练模型。我们的模型也超过了最新的多语言多模态预训练模型 UC2，达到了 SOTA 的性能。

【可视化样例】

将问题和图片拼接输入到模型当中，取模型最后一层的注意力层，观察问题对图片的区域的注意力分布情况：

图 4　注意力分布可视化样例

从图 4 中可以看出，对于不同语言的不同问题，模型都可以观察到与问题相关的图片区域，并作出正确的回答。