CLIP介绍

小组：左镕畅，邓家颖

时间：2023/09/23

CONTENTS

01

什么是CLIP

CLIP 是一种将概念知识与图像语义知识相结合的多模态模型

01

它可以轻松区分“猫”和“狗”的图像

02

或者它在训练数据中从未见过的事物

03

因为它对这些短语和这些像素代表的含义有普遍的了解

CLIP 是第一个处理计算机视觉的多模态模型

由 OpenAI 于 2021 年 1 月 5 日发布。“CLIP是一种在各种（图像、文本）对上进行训练的神经网络模型。可以用自然语言指示它在给定图像的情况下预测最相关的文本片段，而无需直接针对任务进行优化，类似于 GPT- 2和 GPT-3 的零样本功能。”

CLIP 模型是建立在数亿张图像和字幕基础上的神经网络模型，可以返回给定图像的最佳标题

具有令人印象深刻的“零样本”能力，使其能够准确预测以前从未见过的整个类别！

02

CLIP 是如何工作的

CLIP 模型由两个子模型组成：

01

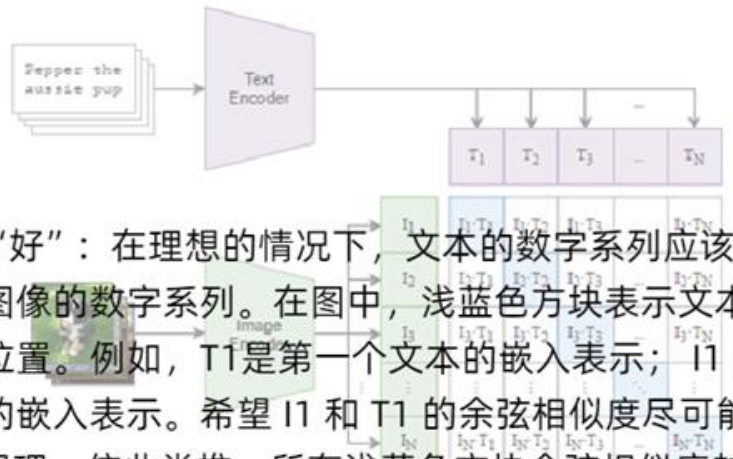
一个文本编码器，它将文本嵌入到数学空间中。

一个图像编码器，它将图像嵌入到数学空间中。

02

衡量该模型的“好”“坏”

拟合一个监督学习模型时，需要找到某种方法来衡量该模型的“好”或“坏”——目标是拟合一个尽可能“最好”和“最不坏”的模型。CLIP 模型也不例外：



01

模型的“好”：在理想的情况下，文本的数字系列应该非常接近相应图像的数字系列。在图中，浅蓝色方块表示文本和图像重合的位置。例如，T1是第一个文本的嵌入表示；I1 是第一幅图像的嵌入表示。希望 I1 和 T1 的余弦相似度尽可能高，I2 和 T2 同理，依此类推，所有浅蓝色方块余弦相似度越高，模型就越“好”！

02

模型的“坏”：在想要最大化每个蓝色方块的余弦相似度的同时，还有很多灰色方块指示文本和图像未对齐的位置。希望所有灰色方块都具有低余弦相似度。

文本和图像编码器适配

一旦模型适合，可以将图像传递到图像编码器中以检索最适合图像的文本描述，反之亦然，可以将文本描述传递到模型中以检索图像。

在所有的文本和图像对中，文本编码器和图像编码器通过同时最大化那些蓝色方块的余弦相似度并最小化灰色方块的余弦相似度来同时拟合。

03

CLIP应用

图像分类

OpenAI 最初将 CLIP 评估为零样本图像分类器。他们将其与传统的监督机器学习模型进行了比较，其表现几乎与传统的监督机器学习模型相当，而无需在任何特定数据集上进行训练。

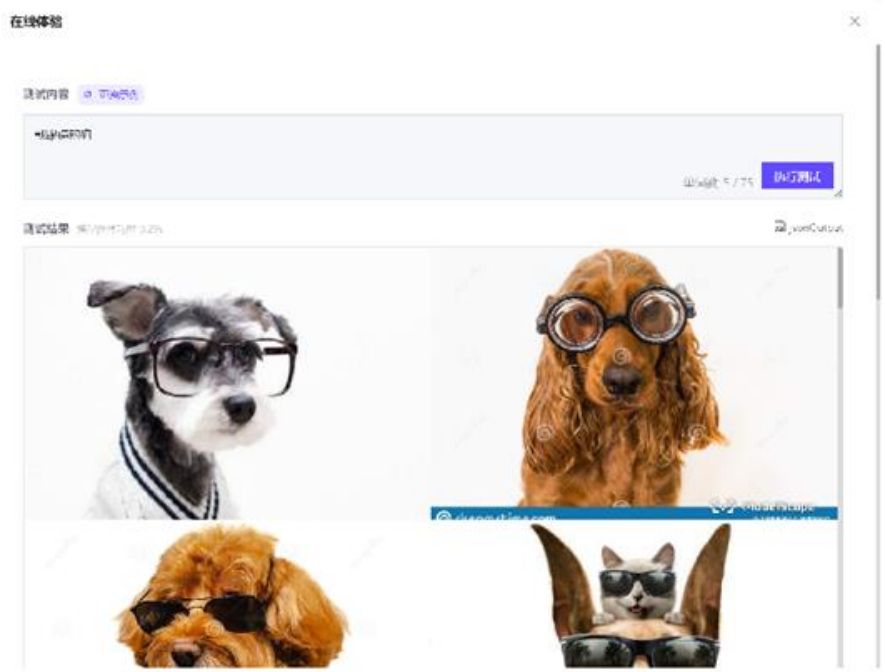
传统图像分类方法面临的一项挑战是，需要大量训练示例。可用的训练数据较少时，CLIP 在此任务上的表现相对传统图像分类方法更好。

内容审核

图像分类的一种扩展是内容审核。如果以正确的方式提出要求，CLIP 可以过滤掉图形或 NSFW 图像。

图像生成

根据文本描述生成图像；使用 CLIP 来评估其功效。



04

实验

零样本图像分类结果

cifar100

```
torch.Size([10000, 100])

Result:
zeroshot-top1: 0.8992
```

flower1

```
torch.Size([9, 4])

Result:
zeroshot-top1: 1.0
```

flower2

```
torch.Size([500, 5])

Result:
zeroshot-top1: 0.93
```

蔡徐坤


1.jpg


2.png


3.jpg


4.jpg

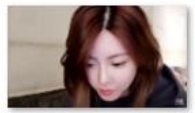

5.jpg

周淑怡


0e206d6ad31db14f92e7a4a5ea8b749f.jpg


3.png


20200330165756_98708.png


R-C.jpg


s.png

李佳琪


12.jpg


OIP-C.jpg


R-C.jpg


s.jpg

```
mine

torch.Size([14, 3])

Result:
zeroshot-top1: 0.7142857142857143
```

效果居然还挺好

补充实验1

仅添加4张蔡徐坤

```
torch.Size([18, 3])
Result:
zeroshot-top1: 0.7777777777777778
```

仅添加4张周淑怡

```
torch.Size([18, 3])
Result:
zeroshot-top1: 0.7777777777777778
```

仅添加4张李佳琪

```
torch.Size([18, 3])
Result:
zeroshot-top1: 0.6111111111111112
```

分析

- 不太可能
  - 训练集中有蔡徐坤和周淑怡，所以这两类图片表现较好

- 更有可能
  - 训练集中这三个都没有，而CLIP能work的原因也许是：通过分析姓名，能表示出姓名所代表的性别的概率；同时又能从图片中表示图片所代表的性别的概率。
  - 我认为从图片中获取性别信息应该是非常精准的，而从姓名中获取性别信息会有些困难
  - 例如“蔡徐坤”大概率是男的，“周淑怡”大概率是女的，而“李佳琪”又像男又像女，这就导致了“李佳琪”成为了一个干扰分类

补充实验2

删去李佳琪图片

```
torch.Size([10, 2])
Result:
zeroshot-top1: 1.0
```

把李佳琪图片加到蔡徐坤里

```
torch.Size([14, 2])
Result:
zeroshot-top1: 1.0
```


1.jpg


2.png


3.jpg


4.jpg


5.jpg

男

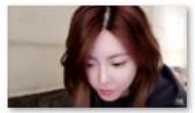
蔡徐坤

男


0e206d6ad31db14f92e7a4a5ea8b749f.jpg


3.png


20200330165756_98708.png


R-C.jpg


s.png

女

周淑怡

女


12.jpg


OIP-C.jpg


R-C.jpg


s.jpg

男

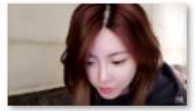
李佳琪

女?


0e206d6ad31db14f92e7a4a5ea8b749f.jpg


3.png


20200330165756_98708.png


R-C.jpg


s.png

蔡徐坤

周淑怡

李佳琪

蔡徐坤

周淑怡

李佳琪

谢谢大家

小组：左镕畅，邓家颖

时间：2023/09/23