VSE++：使用最大违规改进视觉语义嵌入

摘要

我们提出了一种学习跨模态检索的视觉语义嵌入的新技术。受到最大违规、结构化预测中最大违规的使用以及对损失函数进行排序的启发，我们对用于多模态嵌入的常见损失函数进行了简单的更改。再加上微调和增强数据的使用，可以显着提高检索性能。我们使用消融研究并与现有方法进行比较，在 MS-COCO 和 Flickr30K 数据集上展示我们的方法 VSE++。在 MS-COCO 上，我们的方法在标题检索方面比最先进的方法高出 8.8%，在图像检索方面比最先进的方法高出 11.3%（R@1）。

一、 介绍

联合嵌入可以实现图像、视频和语言理解方面的广泛任务。示例包括用于形状推断的形状图像嵌入[20]、双语单词嵌入[38]、用于 3D 姿势推断的人体姿势图像嵌入[19]、细粒度识别[25]、零-镜头学习[9]，以及通过合成进行模态转换[25, 26]。这种嵌入需要从两个（或更多）域映射到公共向量空间，其中语义相关的输入例如文本和图像）被映射到相似的位置。因此，嵌入空间代表了底层的域结构，其中位置和方向通常具有语义意义。

视觉语义嵌入一直是图像标题检索和生成 [13, 15] 以及视觉问答 [22] 的核心。例如，视觉问答的一种方法是首先通过一组标题描述图像，然后找到最接近的标题来回答问题 ([1, 37])。对于从文本合成图像，可以从文本映射到联合嵌入空间，然后再映射回图像空间[25, 26]。

在这里，我们重点关注跨模态检索的视觉语义嵌入；即检索给定标题的图像，或检索查询图像的标题。正如检索中常见的那样，我们通

过 R@K 来衡量性能，即 K 处的召回率——正确的查询所占的比例 item 在嵌入空间中最接近查询的 K 个点中检索（K 通常是一个小整数，通常为 1）。更一般地说，检索是评估图像和语言数据联合嵌入质量的自然方法[11]。

基本问题之一是排名；正确的目标应该比语料库中的其他项目更接近查询，这与学习排序问题[18]和 maxmargin 结构化预测 [3, 17] 不同。本文中的公式和模型架构与[15]的关系最为密切，通过三元组排名损失进行学习。与这项工作相反，我们提倡一种新颖的损失，使用增强数据和微调，这使得字幕检索性能比众所周知的基准数据的基线排名损失显着增加。我们比 MS-COCO 的最佳报告结果高出近 9%。我们还表明，通过使用我们更强的损失函数，可以放大更强大的图像编码器（经过微调）的好处。我们将我们的模型称为 VSE++。为了确保可重复性，我们的代码是公开可用的[1]。

我们的主要贡献是将最大违规纳入损失函数中。这是受到分类任务中使用最大违规挖掘 ([5, 7, 23]) 以及使用最大违规改进人脸识别图像嵌入的启发 ([27, 33])。使用最大违规最小化损失函数相当于最小化使用均匀采样的修改后的非透明损失函数。我们扩展了这个想法，在多模态嵌入的损失中明确引入最大违规，而不需要任何额外的挖掘成本。

我们还注意到，我们的表述补充了最近为该问题提出新架构或相似函数的其他文章。为此，我们展示了对[31]的改进。在可以通过修改损失来改进的其他方法中，[32]提出了一种嵌入网络来完全取代用于排名损失的相似性函数。 [24] 使用了图像和标题的注意力机制，作者顺序地、选择性地关注单词和图像区域的子集来计算相似性。在[12]中，作者使用多模态上下文调制注意力机制来计算图像和标题之间的相似度。我们提出的损失函数和三元组采样可以扩展并应用于其他此类问题。

2.1 视觉语义嵌入

令 $\emptyset(i:\theta_\emptyset) \in R^{D_\emptyset}$ 为根据图像 i 计算的基于特征的表示VGG19 [28]或 ResNet152 [10]中 logits 之前的表示。类似地，令 $\varphi(i:\theta_\varphi) \in R^{D_\varphi}$ 为字幕嵌入空间（例如基于 GRU 的文本编码器）中字幕 c 的表示。这里，$\theta_\emptyset$ 和 θ_ψ 表示分别映射到这些的模型参数初始图像和标题表示。

然后，让到联合嵌入空间的映射由线性投影定义：

$$f(i; W_f, \theta_\emptyset) = W_f^T \emptyset(i; \theta_\emptyset) \quad (1)$$

$$g(c; W_g, \theta_\mu) = W_g^T \mu(c; \theta_\mu) \quad (2)$$

其中$W_f \in R^{D_\emptyset * D}$和 $W_g \in R^{D_\mu * D}$。我们进一步将 f (i;W_f, θ_ϕ) 和 g(c;W_g, θ_ψ) 标准化，使其位于单位超球面上。最后，我们将联合嵌入空间中的相似度函数定义为通常的内积：

$$s(i, c) = f(i; W_f, \theta_\emptyset) \cdot g(c; W_g, \theta_\mu) \quad . \quad (3)$$

设 θ ={W_f,W_g, θ_ϕ, θ_ψ}为模型参数。如果我们还微调图像编码器，那么我们还将在 θ 中包含 θ_ϕ。

训练需要最小化关于 θ 的经验损失，即累积训练数据损失 $S = \{(i_n, c_n)\}_{n=1}^N$

$$e(\theta, S) = \frac{1}{N}\sum_{n=1}^N l(i_n, c_n) \quad (4)$$

其中 l(i_n, c_n) 是单个训练样本的合适损失函数。受到图像检索中使用三元组损失[4, 8]的启发，最近的联合视觉语义嵌入方法使用了基于铰链的三元组排名损失[13,15,29,36]：

$$\lambda_{SH}(i, c) = sum_{\hat{c}}[\alpha - s(i, c) + s(i, \hat{c})]_+ + sum_{\hat{i}}[\alpha - s(i, c) + s(i, \hat{c})]_+ \quad (5)$$

其中 α 用作余量参数，[x]+max(x, 0)。该铰链损耗包括两个对称项。给定查询 i，第一个总和取自所有负标题 cˆ。第二个是对所有负图像 iˆ进行处理，给定标题 c。每一项都与负样本集的预期损失（或违规）。如果 i 和 c 在联合嵌入空间中比任何负值更接近，则铰链损失为零。在实践中，为了计算效率，通常只对小批量随机梯度下降中的负数求和（或随机采样），而不是对训练集中的所有负数求和[13,15,29]。计算这种损失近似的运行时复杂度是小批量中图像标题对数量的二次方。

当然，人们还可以考虑其他损失函数。一种是成对铰链损失，其中鼓励正对的元素位于联合嵌入空间中半径为 ρ_1的超球面内，而负对的元素不应小于 $\rho_2 > \rho_1$。这是有问题的，因为它比排名损失更能限制潜在空间的结构，并且需要使用两个很难设置的超参数。另一种可能的方法是使用典型相关分析来学习 W_f和 W_g，从而尝试在联合嵌入中保留文

本和图像之间的相关性[6, 16]。相比之下，当衡量 R@K 的性能时，对于较小的 K，基于相关性的损失不会对负项在正对局部附近的嵌入产生足够的影响，而这对于 R@K 至关重要。

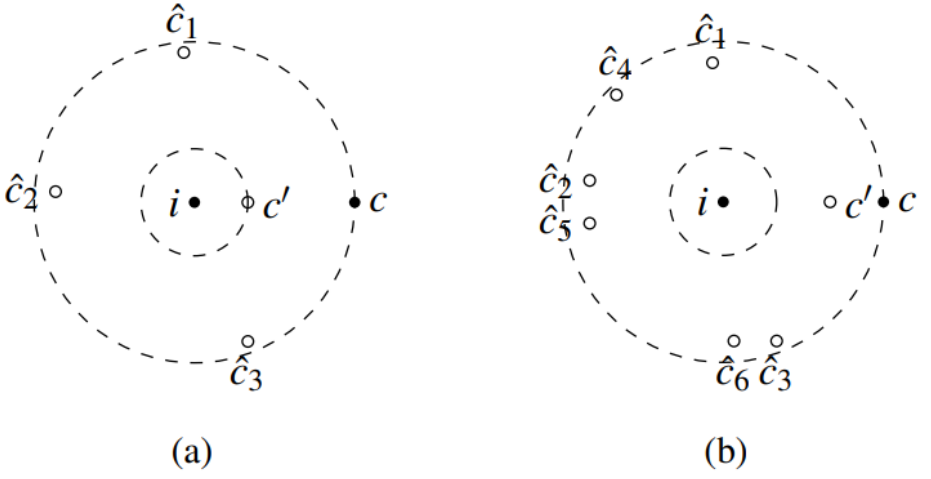

图 1：典型的正样本对和最近的负样本的图示。这里假设相似度得分是负距离。实心圆圈表示正样本对 (i, c)，而空心圆圈表示查询 i 的负样本。两侧的虚线圆以相同的半径绘制。请注意，最难的负样本更接近 (a) 中的 i。假设零裕度，与 (a) 相比，(b) 的 SH 损失更高。 MH 损失将更高的损失分配给 (a)。

2.2 最大违规

受结构化预测中使用的常见损失函数[7,30,35]的启发，我们专注于训练的最大违规，即最接近每个训练查询的负例。这对于检索尤其重要，因为它是决定 R@1 衡量成功或失败的最难的否定。给定正对 (i, c)，最难的负数由 $i = \text{argmax}_{j \neq i} s(j, c)$ 给出。为了强调最大违规因素，我们将损失定义为

$$\lambda_{MH}(i,c) = \max_{\hat{c}}[\alpha - s(i,c) + s(i,\hat{c})]_+ + \max_{\hat{i}}[\alpha - s(i,c) + s(\hat{i},\hat{c})]_+ \quad (6)$$

就像等式5一样，该损失包括两项，一项使用 i，一项使用 c 作为查询。与等式不同。如图 5 所示，该损失是根据最难的负数 c 和 i 来指定的。我们参考等式中的损失。 6 作为最大铰链 (MH) 损失，以及方程式 6 中的损失。5 为铰链总和 (SH) 损失。从 SH 损失到 MH 损失，损失函数有一系列。在 MH 损失中，获胜者获得所有梯度，而我们使用所有三元组的重新加权梯度。我们只讨论 MH 损失，因为根据经验发现它表现最好。

MH 损失优于 SH 的一种情况是，当具有较小违规的多个负数结合起来主导 SH 损失时。例如，图 1 描绘了一对正例和两组负例。在图 1(a) 中，单个否定与查询太接近，这可能需要对映射进行重大更改。然而，任何排除最大违规的训练步骤都可能导致许多小的违规负例，如图 1(b) 所示。使用 SH 损失，这些“新”负数可能会主导损失，因此模型被推回到图 1(a) 中的第一个示例。这可能会在 SH 损失中产生局部最小值，而这对于 MH 损失来说可能不会造成问题，因为 MH 损失侧重于最难的负值。

为了提高计算效率，我们不是在整个训练集中找到最难的负样本，而是在每个小批量中找到它们。这与 SH 损失的复杂度具有相同的二次复杂度。通过对小批量进行随机采样，这种近似值还具有其他优点。一是很有可能得到比整个训练集至少 90% 更难的最大违规。此外，损失对于训练数据中的标签错误可能具有鲁棒性，因为在整个训练集中对最困难的负样本进行采样的概率有点低。

结论

本文重点关注学习用于跨模式图像标题检索的视觉语义嵌入。受结构化预测的启发，与使用预期误差的当前方法相比，我们基于相对最大违规所引起的违规提出了一种新的损失[15, 31]。我们在 MS-COCO 和 Flickr30K 数据集上进行了实验，结果表明我们提出的损失显着提高了这些数据集的性能。我们观察到，改进的损失可以在导更强大的图像编

码器 ResNet152上有更好的效果。