UniCLIP：对比语言图像预培训统一框架

Janghyeon Lee*
LG 人工智能研究
janghyeon.lee@lgresearch.ai

Jongsuk Kim*†
韩国科学技术院
jskpop@kaist.ac.kr

Hyounguk Shon†
韩国科学技术院
hyounguk.shon@kaist.ac.kr

Bumsoo Kim
LG 人工智能研究
bumsoo.kim@lgresearch.ai

Seung Hwan Kim
LG 人工智能研究
sh.kim@lgresearch.ai

李洪乐
LG 人工智能研究
honglak@lgresearch.ai

Junmo Kim
韩国科学技术院
junmo.kim@kaist.ac.kr

摘要

使用对比度目标预训练视觉语言模型已经取得了可喜的成果，这些成果既可以扩展到大型未整理数据集，也可以转移到许多下游应用中。随后的一些研究通过添加自监督项来提高数据效率，但这些研究中域间（图像-文本）对比损失和域内（图像-图像）对比损失是在单独空间上定义的，因此忽略了许多可行的监督组合。为了克服这一问题，我们提出了 UniCLIP，即对比语言-图像预训练的统一框架。UniCLIP 将域间对和域内对的对比损失整合到一个通用空间中。UniCLIP 的三个关键部分解决了整合不同领域对比度损失时出现的差异：(1) 增强感知特征嵌入；(2) MP-NCE 损失；(3) 与领域相关的相似度测量。在各种单模态和多模态下游任务中，UniCLIP 的表现优于之前的视觉语言预训练方法。我们的实验表明，UniCLIP 的每个组成部分都对最终性能做出了很好的贡献。

1 导言

深度学习的最新进展表明，在预训练大规模模型方面取得了重大进展，这些模型可以很好地转移到各种下游应用中。随着这种模式在计算机视觉和自然语言处理领域的成功，人们提出了从自然语言监督中学习图像表征的视觉语言预训练模型[9, 20]。在这些研究中，预训练是在简单的对比损失contrastive loss）条件下进行的，这种损失使得图像的嵌入及其匹配的文本描述（正对）比其他任意图像-文本对（负对）更相似。

为了实现数据效率更高的预训练目标，随后的研究 [13, 16] 为图像-文本对比损失引入了更多的自监督术语，包括增强型自监督和自监督-文本对比损失。

*等额捐款。按字母顺序排列。

†在 LG AI Research 实习期间完成的工作。

第 36 届神经信息处理系统会议（NeurIPS 2022）。

图像 [3，4]、增强文本 [28] 和屏蔽文本 [13]。将更多的正/负监督对纳入最终的对比度损失中，会使目标在数学上更令人满意[3]，从而使模型更节省数据。然而，这些研究都有一个很大的局限性，因为域内对比损失（如图像-图像对）和域间对比损失（如图像-文本对）是在不同的空间中独立定义的。这意味着对比度损失并不了解消极监督的大量可行组合，例如，在计算图像-文本监督的对比度损失时，图像-图像对并不包括在内，这在数据效率和特征多样性方面留下了巨大的改进空间。基于这一观点，我们将本文的目标设定为建立一个对比性图像-文本预训练框架，在该框架中，所有可能的域内和域间图像对的对比性学习都定义在同一个统一的嵌入空间中。

虽然这一目标听起来很直观，但在统一空间中定义多种模式之间的对比损失却面临着一些挑战。首先，在应用图像增强时，图像-文本语义之间可能会发生错位。例如，在图 1 中，"红苹果在绿苹果切片的右边 "的语义很容易被简单的图像增强所破坏，如水平翻转、灰度转换或裁剪，而这些都是图像-图像对比自监督学习中使用的基本增强[3]。我们从实验中验证了，对这种差异不加注意会阻碍训练并降低最终性能。其次，现有文献中针对多正对的对比损失[10, 15]与我们处理不同模态嵌入的训练目标不兼容。这是因为域内对，就像一张图像的两个不同的增强视图，相对来说更容易进行嵌入。

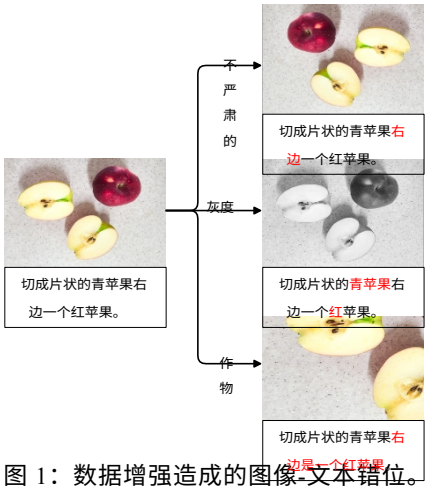

图 1：数据增强造成的图像-文本错位。对齐错误的文本用红色标出（最好用彩色浏览）。

与图像-文本配对等域间配对相比，易阳性示例更容易出现。现有的损失 [10, 15] 容易受到这种情况的影响，因为易阳性示例和难阳性示例会。最后，我们发现，在来自不同领域的嵌入之间应用相同的相似性度量，会产生不同的结果。

由于域间和域内样本对之间的相似性度量存在固有差异，*即域*内样本对中的样本可以任意接近，而域间样本对中的样本则不能。

在本文中，我们提出了 UniCLIP：对比语言-图像预训练的**统一**框架，它将多种模态之间的对比目标**统一**在一个嵌入空间上。UniCLIP 的关键组件可解决上述各项挑战：(1) **增强感知特征嵌入**，使 UniCLIP 能够感知数据增强造成的错位；(2) **MP-NCE 损失**，旨在稳定易对和难对的训练；(3) 与**领域相关的相似性度量**，可调整域间对和域内对的相似性尺度差异。通过解决上述三个问题，UniCLIP 在各种单模态和多模态下游任务（如线性探测、零镜头分类、微调和图像文本检索）中的表现优于现有的视觉语言预训练方法。我们验证了 UniCLIP 的每个组件都成功地解决了统一空间中的对比学习问题，并对最终性能做出了有意义的贡献。我们的贡献总结如下：

- 我们提出了一个用于视觉语言预训练的统一框架 UniCLIP，它通过将多个领域中定义的对比损失整合到一个单一的通用空间来提高数据效率。我们研究了这种整合带来的新技术挑战。
- 我们为 UniCLIP 设计了新的组件来应对上述挑战：增强感知特征嵌入、MP-NCE 损失和领域相关相似度测量。广泛的实验表明，我们提出的每个组件都对最终性能起到了关键作用。
- 在包含各种模式的多个下游任务中，UniCLIP 的表现优于现有的视觉语言预训练方法。

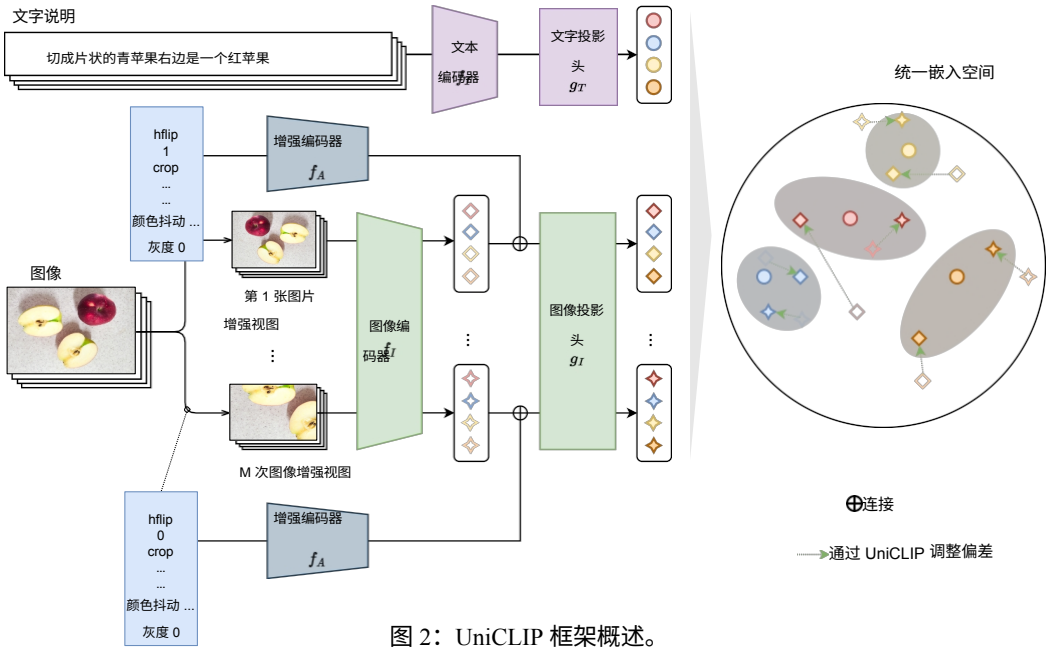

图 2：UniCLIP 框架概述。

2 方法

UniCLIP 架构（图 2）由增强编码器 f_A、图像编码器 f_I、文本编码器 f_T 以及相应的投影头 $g_{(I)}$和 $g_{(T)}$组成。对于文本标题数据，f_T和 g_T在与图像嵌入空间*相同*的嵌入空间上生成文本嵌入。图像和文本表征是通过我们的多正 NCE 损失来学习的，并在统一的嵌入空间上测量与领域相关的相似性得分。下文将详细介绍我们方法的各个要素。

2.1 建筑学

增强**编码器** 为了使增强指令 A 作为网络的输入，我们首先将其描述为一个实向量，其中包含 A 中每种基本变换应用于数据多少的信息。，对比学习中经常出现的图像增强可以按如下方式转换为实向量：

- *裁剪和调整大小*在归一化坐标系中（*即*原始图像的左上角为 $(0,0)$，右下角为 $(1,1)$），`RandomResizedCrop` 增量被编码为一个四维向量（x, y, w, h），其中 (x, y) 是裁剪图像的左上角坐标，(w, h) 是裁剪图像的大小。
- *色彩抖动*当色彩抖动增强改变了图像的亮度、对比度、饱和度和色调时，这种增强就会被编码成一个由这四个因素的变化组成的四维向量。
- *高斯模糊*：高斯模糊增强编码为高斯模糊核的标准偏差。
- *水平翻转*：`RandomHorizontalFlip` 放大系数在图像实际翻转时编码为 1，否则编码为 0。
- *灰度转换*：如果图像实际转换为灰度，则随机灰度增强编码为 1，否则为 0。

如果一个图像增强A由上述所有五种增强组成，则首先会根据上述规则将A编码为一个 11 维向量然后通过一个 MLP 得到增强嵌入 $f_A(A)$。请注意，由于增强的随机性，每个前向和每个样本的 $f_A(A)$ 都会不同。

图像编码器和图像投影头 为了让模型学会如何调整由图像增强引起的图像-文本错位，图像编码器或投影头必须将增强信息作为输入。但是，如果编码器不知道对图像进行了哪种增强，就无法充分利用增强数据。，当编码器经过水平翻转增强训练时，如果它将增强图像和图像是否翻转的标志作为输入，其形式为（图像，未翻转标志）或（翻转图像，翻转标志），那么当编码器需要对下游任务中的（翻转图像，未翻转标志）进行编码时，可能会表现出不理想的行为，因为编码器没有经过此类数据的训练，这意味着模型失去了一些泛化能力。因此，图像编码器必须与增强*无关*，图像投影头必须具有增强*感知能力*。，编码器能充分享受数据增强带来的好处，并更好地进行泛化，而投影头仍能纠正增强造成的域间错位。

为了使图像表示与增强无关，使图像嵌入与增强相关，增强信息只提供给投影头，而编码器只看到增强图像，不知道应用了哪种增强。因此，对于图像 x，图像编码器 f_I 将增强图像 $A(x)$ 作为输入，得到一个不考虑增强的图像表示 $h = f_I(A(x))$。然后，通过图像投影头 $g_{(I)}$ 从图像表示 h 和增强嵌入 $f_{(}A)(A)$得到统一嵌入空间中的增强感知图像嵌入 $z = g_{(I)}(f_{(I)}(A(x)), f_{(}A)(A_{))}$。我们采用 ViT（Vision Transformer）[7] 作为具有可学习位置嵌入的图像编码器 f_I，图像投影头 $g_{(I)}$ 由三个残差块组成。cls]标记的最后一个激活值被用作图像表征 h。

文本编码器和文本投影头 首先通过字节对编码对原始文本进行标记化，并用开始标记和结束标记进行包装，得到标记化文本 x。任何文本增强方法都可以在这里应用，就像图像嵌入的情况一样，但我们不会为一个文本创建多个增强视图，因为我们发现这样做没有太大帮助。因此，统一潜空间中的文本表示 $h = f_T(x)$ 和文本嵌入 $z = g_{(T)}(f_T(x))$ 是在没有任何增强嵌入的情况下得到的。我们使用 Transformer [27] 作为带有可学习位置嵌入的文本编码器 f_T，并使用线性层作为文本投影头 g_{T}。起始标记的最后一个激活值被用作文本表示 h。

2.2 多正对的对比损失函数

对比损失函数可根据损失对一个数据点的正负对数量进行分类。例如，triplet loss [22] 只取一个正对和一个负对，N-pair loss [24] 和 InfoNCE loss [26] 取一个正对和多个负对，MIL-NCE loss [15] 和 SupCon loss [10] 取多个正对和多个负对。由于在我们的统一框架中存在多个正对，我们首先回顾一下 MIL-NCE loss 和 SupCon loss 函数，并讨论它们的缺点。

对于一批嵌入 $\{z_{(i)}\}_{(i)}$中的*第* i *个*嵌入 z_i，设 $P_{(i)}$ 是第 i *个*样本的所有正样本索引（不包括 i 本身）的集合，N_i 是*第 i 个*样本的所有负样本索引的集合。

$$P_i = \{j \mid (z_i, z_{(j)}) \text{ 为正对，且 } j \neq i\} \tag{1}$$

$$N_i = \{j \mid (z_{(i)}, z_{(}j_{)}) \text{ 是负对}\}。 \tag{2}$$

*第 i 个*嵌入点和*第 j 个*嵌入点之间的相似性得分用 $s_{i,j} > 0$ 表示。对比损失函数会尽量增大正对的相似性得分，同时尽量减小负对的相似性得分。，如果 批次中的每个样本只有一个阳性样本，即 $P_i = \{p_{(i)}\}$，那么*第* i *个*样本的 InfoNCE 损失 [26] 或 NT-Xent 损失 [3] 可描述为

$$L_i^{\text{InfoNCE}} = -\log \frac{s_{i,pi}}{s_{i,pi} + \sum_{n \in N_i} s_{i,n}}. \tag{3}$$

MIL-NCE 损失 *第* i *个*嵌入的 MIL-NCE 损失 [15] 定义如下

$$L_i^{\text{MIL-NCE}} = -\text{对数}\frac{\sum_{p \in P_i} s_{i,p}}{\sum_{p \in P_i} s_{i,p} + \sum_{n \in N_i} s_{i,n}}. \tag{4}$$

配置 MIL-NCE 损失函数的目的是最大化所有正比例分总和 $\sum_{p\in P(i)} s(i,p)$，最小化所有负例相似性得分总和 $\sum_{n\in N(i)} s(i,n)$。然而，当存在相似性得分大到足以支配分子和分母的易正对时，硬正对就无法从 $L^{\text{MIL-NCE}}$ 中获得足够的梯度，

因为 MIL-NCE 损失将负对与正对得分之和进行比较 $\sum_{p\in P_i} s_{(i,p)}$，而不是每个正对 $s_{i,p}$。对于 $\in P_i$，从 $L^{\text{MIL-NCE}}$ 到 $s_{i,q}$ 的梯度为

$$\frac{\partial L_i^{\text{MIL-NCE}}}{\partial s_{i,q}} = -\frac{\sum_{n\in N_i} s_{i,n}}{\sum_{p\in P_i} s_{i,p} + \sum_{n\in N_i} s_{i,n}}, \tag{5}$$

因此，当 $\sum_{p\in P_i} s_{i,p}$ 已经很大时，梯度将消失为零，因为即使正对的分数 $s_{i,q}$ 很小，易正对的分数也很大。换句话说，在 MIL-NCE 损失中，*易正对阻碍了难正对的训练*。在我们的统一框架中，这个问题会更加突出，因为硬阳性和易阳性经常与来自域内和域间的监督同时存在。

SupCon Loss *第* i *个*嵌入的 SupCon loss [10] 描述如下

$$L_i^{\text{SupCon}} = \mathbb{E}_{p\in P_i}\left[-\log \frac{s_{i,p}}{\sum_{p'\in P_i} s_{i,p'} + \sum_{n\in N_i} s_{i,n}}\right]. \tag{6}$$

在这种情况下，每个正分 $s_{i,p}$ 都会与负分对进行比较，但分母中的正分之和仍然会产生不良的。对于相似性得分较高的简单正向配对，可以通过降低其得分来减少损失，从而减少分母。对于 $q\in P_i$、

$$\frac{\partial L_i^{\text{SupCon}}}{\partial s_{i,q}} = \frac{s_{i,q} - \frac{1}{|P_i|}\sum_{p\in P_i} s_{i,p} + \sum_{n\in N_i} s_{i,n}}{\sum_{p\in P_i} s_{i,p} + \sum_{n\in N_i} s_{i,n}}, \tag{7}$$

因此，由于分母中 $s_{i,q}$ 项的更新相对较大，硬阳性的训练效果会好于 MIL-NCE 损失。但是，如果我们假设正向得分的总和远大于负向得分的总和，那么

$$\frac{\partial L_i^{\text{SupCon}}}{\partial s_{i,q}} \propto s_{(i,q)} - \frac{1}{|P_i|}\sum_{p\in P_i} s_{i,p} + \sum_{n\in N_i} s_{i,n} \approx s_{i,q} - \mathbb{E}_{p\in P_i}[s_{i,p}]。 \tag{8}$$

由于梯度并不总是负的，L^{SupCon} 将试图降低易正对 (z_i, z_q) 的相似性得分，因为 $s_{(i,q)}$ 将大于平均正对得分，而不是增加或至少保持它。说，*硬正对会阻碍* SupCon 损失中易*正分数的收敛*。

多正 NCE 损失 由于分母中的正分之和会导致易正对和难正对干扰，因此我们可以直接使用多正 版本的 InfoNCE 损失，让每个正对都独立贡献损失，如下所示。

$$L_i = \mathbb{E}_{p\in P_i}\left[-\log \frac{s_{i,p}}{s_{i,p} + \sum_{n\in N_i} s_{i,n}}\right] \tag{9}$$

从梯度中可以看出

$$\frac{\partial L_i}{\partial s_{i,q}} = -\frac{\sum_{n\in N_i} s_{i,n}}{|P_{(i)}|\, s_{i,q} + \sum_{n\in N_i} s_{i,n}}, \tag{10}$$

由于梯度总是负的，因此不会出现易正对相似性递减的问题。

有了这种多正向版本的 InfoNCE 损失，我们会重新考虑将 i 从正向集合中剔除公式 1 中的 P_i。如果对比损失可以处理多个正对，那么就没有理由

从损失定义中排除三元对（$z_{(i)}, z_i$）。由于 z_i 与 z_i 本身最 相似，三元对也必须被用作强正对，这将导致

$$L_i = E_{p\in Pi\cup\{i\}}\left[-\log \frac{s_{i,p}}{s_{i,p}+\sum_{n\in Ni} s_{i,n}}\right]. \tag{11}$$

在此，我们为统一的对比 ve 学习框架提出了一种多正向 NCE 损失，称为 MP-NCE 损失，它是公式 11 的加权版本，定义为

$$L_{\text{MP-NCE}} = E_{p\in Pi\cup\{i\}}\left[-w_{D(i,p)}\log \frac{s_{i,p}}{s_{i,p}+\sum_{n\in Ni} s_{i,n}}\right], \tag{12}$$

其中，$D(i, p)$表示第 i 个和第 p 个数据采样的域组合，$w_{D(()(i,p)())}$是特定域的平衡超参数，它使得每个域间和域内监督对损失的贡献相同。，当我们对数据集中的每个原始图像-文本对使用三幅图像的增强视图和一个相应的文本时，一批数据中共有 9N 个图像-图像正对、6N 个图像-文本正对和 N 个文本-文本正对，因此如果（$z_i, z_{(p)}$）是图像-图像对、图像-文本对和文本-文本对，$w_{D()()(i,p)()}$将分别设为 $1/9$、$1/6$ 和 1。

虽然我们提出的 MP-NCE loss 是在多正向环境中使用的，但即使是在单正向环境中，例如图像自监督对比学习，也应考虑使用 MP-NCE loss，将微不足道的一对（z_i, z_i）也视为正向，因为与骨干网络相比，MP-NCE 涉及的计算开销可以忽略不计。

2.3 与领域相关的相似性得分

在 SimCLR [3] 和 CLIP [20] 中，第 i 个嵌入 z_i 和第 j 个嵌入 $z_{(j)}$之间的相似性得分 $s_{(i,j)}$ 定义为

$$s_{i,j} = \exp\left(\frac{1}{\tau}\cdot \frac{z_i^{\top} z_j}{\|z_i\|\|z_j\|}\right), \tag{13}$$

其中，τ 是一个正，通常小于 1。由于两个嵌入式的余弦相似度的值不能超出区间 $[-1, 1]$，因此余弦相似度要除以温度 τ 以扩大其范围。τ 可以是一个预定义的超参数，也可以是一个可学习的参数，允许模型为对比损失的收敛选择一个合适的尺度。

要将输入对（$z_{(i)}$，z_j）分类为正或负，我们可以定义一个阈值 b，如果 z_i 和 z_j 之间的余弦相似度大于 t *han b*，则将其分类为正，否则为负。我们可以将阈值 b 作为偏移量吸收到相似性得分中，比如

$$s_{i,j} = \exp\left(\frac{1}{\tau}\cdot \frac{z_i^{\top} z_j}{\|z_i\|\|z_j\|} - b\right), \tag{14}$$

并期望通过模型学习到最佳阈值，就像温度的情况一样。请注意，如果余弦相似度大于 b，则温度会放大得分，反之会降低得分，因此等式 14 是一个合理的相似度度量，可将阈值视为二元分类问题的决策边界。然而，不幸的是，偏移量 b 对 InfoNCE 损失（等式 3）并没有 ，因为分子和分母中的 b 相抵后的结果是

$$L_i^{\text{InfoNCE}} = -\log \frac{s_{i,pi}}{s_{i,pi}+\sum_{n\in Ni} s_{i,n}} = -\log \frac{\exp(b/\tau)s_{i,pi}}{\exp(b/\tau)s_{i,pi}+\sum_{n\in Ni}\exp(b/\tau)s_{i,n}} \tag{15}$$

对于任意 τ 和 b，这意味着$\partial^{L(\text{InfoNCE})}/\partial b$ 始终为零。

，在我们的统一框架中，当数据对从多个域中采样时，阈值会因采样数据对是域内数据对还是域间数据而不同，因为一般来说，域内正向数据对比域间正向数据对更容易分类。这促使我们引入 领域特定温度 $\tau_{Di,j)()}$ 和偏移量 $b_{D()()(i,j)()}$，并提出了一个与领域相关的相似性分数

$$s_{i,j} = \exp\left(\frac{1}{6}\cdot \frac{z_i^{\top} z_j}{\|z_i\|\|z_j\|}\right)^{i}$$

$\tau_{D(i,j)}$ - $b_{D(i,j)}$. (16)

表 1：11 个下行数据集的零拍摄图像分类性能和线性探测性能。(†) 原始论文中报告的结果。

方法	预培训数据集	宠物	CIFAR-10	CIFAR-100	太阳 397	Food-101	花卉	水果	Caltech-101	空气飞机	DTD	伊甸园 geNet	Average
零射击分类：													
CLIP-ViT-B/32	YFCC15M	19.4	62.3	33.6	40.2	33.7	6.3	2.1	55.4	1.4	16.9	31.3	27.5
SLIP-ViT-B/32	YFCC15M	28.3	72.2	45.3	45.1	44.7	6.8	2.9	65.9	1.9	21.8	38.3	33.9
DeCLIP-ViT-B/32	YFCC15M	30.2	72.1	39.7	**51.6**	46.9	7.1	**3.9**	70.1	2.5	**24.2**	41.2	35.4
UniCLIP-ViT-B/32	YFCC15M	**32.5**	**78.6**	**47.2**	50.4	**48.7**	**8.1**	3.4	**73.0**	**2.8**	23.3	**42.8**	**37.3**
DeCLIP-ResNet50† [13]	Open30M	-	-	-	-	-	-	-	-	-	-	49.3	-
UniCLIP-ViT-B/32	Open30M	69.2	87.8	56.5	61.1	64.6	8.0	19.5	84.0	4.7	36.6	**54.2**	49.7
线性探测													
CLIP-ViT-B/32	YFCC15M	71.2	89.2	72.1	70.1	71.4	93.2	34.9	84.3	29.7	60.9	61.1	67.1
SLIP-ViT-B/32	YFCC15M	75.4	90.5	75.3	73.5	77.1	96.1	43.0	87.2	34.1	71.1	68.1	71.9
DeCLIP-ViT-B/32	YFCC15M	76.5	88.6	71.6	75.9	79.3	96.7	42.6	88.0	32.6	69.1	69.2	71.8
UniCLIP-ViT-B/32	YFCC15M	**83.1**	**92.5**	**78.2**	**77.0**	**81.3**	**97.1**	**49.8**	**88.9**	**36.2**	**72.8**	**70.8**	**75.2**
UniCLIP-ViT-B/32	Open30M	85.4	95.1	81.5	79.2	84.4	97.3	67.3	91.1	39.0	77.2	74.0	79.1

对于图像-文本统一对比学习，我们有三种可能的领域组合，因此图像-图像对、图像-文本对和文本-文本对将分别有三种不同的温度和三个偏移量。

利用所提出的依赖于领域的相似性得分（等式 16）和 MP-NCE 损失（等式 12），当从多个不同领域对负数据对进行采样时，偏移量不再被抵消。具体来说，由于可以在不改变损失函数的情况下将任何实数添加到等式 15 中的余弦相似性项中，偏移量只损失了一个固有维度，因此模型能够学习*相对*阈值。换句话说，现在我们可以学习特定领域的偏移量，这样我们就可以预期较易领域组合的偏移量大于较难领域组合偏移量。

3 实验

为了实现可重复性，我们在实验中使用了公开的数据集进行训练和评估，包括 CC3M [23]、CC12M [2]、DeCLIP YFCC15M [13, 25]训练和 Pets [18]、CIFAR-10、CIFAR-100 [12]、SUN397 [29]、Food-101 [1]、Flowers [17]、Cars [11]、Caltech-101 [8]、Aircraft [14]、DTD [6]、ImageNet-1k [21]、Flickr30k [19]、COCO Captions [5] 进行评估。我们将 CC3M、CC12M 和 YFCC15M 的联合数据集定义为 Open30M 数据集。

设置 在我们的实验中，对于每张原始图像和相应的文字说明，一张弱增强图像、两张强增强图像和一段文字组成一个正组。详细的增强和优化配置见附录。

3.1 主要成果

我们评估了模型在单模态和多模态下游任务中的可移植性。针对单模态基准，我们对图像分类任务进行了线耳探测和微调；针对多模态基准，我们对图像文本检索任务和零镜头图像分类任务进行了评估。

线性探测与微调 对于单模态 表 2：ImageNet-1k 微调精度实验，我们删除了在 YFCC15M 上预训练的模型的图像投影头 g_I 的图像投影头 g(I)。和增强编码器 f_A，并仅使用 IM 年龄编码器 f_I。表 1 报告了 11 个下游数据集的线性

方法	准确性
CLIP-ViT-B/32	72.27
SLIP-ViT-B/32	75.64
DeCLIP-ViT-B/32	74.34
UniCLIP-ViT-B/32	**76.54**

分类性能。我们在表 2 中报告了 ImageNet 的微调准确率。在单模态实验中，UniCLIP 在所有下游数据集上的表现始终优于其他方法。

表 3：使用在 YFCC15M 上预先训练的模型对 Flickr30k 和 COCO Captions 测试片段进行零镜头图像-文本检索。†在 Open30M 上预先训练。

方法	图像到文本检索 Flickr30k 标题 R@1	R@5	R@10	COCO 标题 R@1	R@5	R@10	文本到图像检索 Flickr30k R@1	R@5	R@10	COCO 标题 R@1	R@5	R@10
CLIP-ViT-B/32	34.9	63.9	75.9	20.8	43.9	55.7	23.4	47.2	58.9	13.0	31.7	42.7
SLIP-ViT-B/32	47.8	76.5	85.9	27.7	52.6	63.9	32.3	58.7	68.8	18.2	39.2	51.0
DeCLIP-ViT-B/32	51.4	80.2	88.9	28.3	53.2	64.5	34.3	60.3	70.7	18.4	39.6	51.4
UniCLIP-ViT-B/32	**52.3**	**81.6**	**89.0**	**32.0**	**57.7**	**69.2**	**34.8**	**62.0**	**72.0**	**20.2**	**43.2**	**54.4**
UniCLIP-ViT-B/32†	75.6	94.2	97.3	46.1	74.0	83.0	61.4	85.2	91.5	35.2	61.3	71.7

零镜头分类和图像-文本检索 表 1 显示了 11 个下游数据集的零镜头分类性能。我们使用与 [16, 20] 相同的提示模板对每个类别进行提示集合。表 3 显示了 Flickr30k 和 COCO Captions 基准上的零镜头图像文本检索结果。

3.2 消融研究

在本节中，我们将报告实验，以检验 UniCLIP 的每个组件对最终性能的贡献。我们在 CC3M 数据集上以 ViT-B/16 为骨干对 UniCLIP 的所有变体进行了 50 个 epochs 的预训练，并比较了它们在 ImageNet-1k 零点评估中的表现。

图像投影头类型 我们尝试了几种不同的图像投影头架构，包括线性层、MLP 层和残差块，当投影头将增强嵌入作为输入或不作为输入时，如表 4a 所示。结果发现，MLP 有很强的过拟合趋势，甚至比线性层的表现更差。如果在投影头中加入跳转连接，就能增强信息，提高能力，同时避免过拟合。让图像投影头具有增强感知能力，可以提高所有类型投影头的性能，因为投影头可以处理域间错位。

增强配置 表 4b 研究了增强编码对增强配置的影响。在没有增强嵌入的情况下，只使用强增强图像会严重降低性能，因为强增强会产生更多图像-文本错位。由于观察到包含一张弱增强图像比只使用强增强图像效果更好，我们选择在正集中保留一张弱增强图像作为稳定的参考样本。

表 4：不同图像投影头类型和增强配置下的 ImageNet-1k 零点拍摄精度。

(a) **图像投影头类型。**使用一个弱图像增强器和两个强图像增强器。

扩建嵌入	封头类型	精度
	MLP 3 层	24.01
	MLP 6 层	23.62
	1 个 ResBlock	**24.76**
	3 个 ResBlocks	24.46
	线性层	24.68
	MLP 3 层	24.54
✓	MLP 6 层	24.15
✓	1 个 ResBlock	27.67
✓	3 个 ResBlocks	**27.84**

(b) **增强配置。**1-ResBlock 磁头用于无增强嵌入的情况，3-ResBlock 磁头用于增强嵌入的情况。

增强嵌入	扩建	准确性
✗	3 弱	24.49
✗	1 弱，2 强	**24.76**
✗	3 强	22.60
✓	3 弱	23.40
✓	1 弱，2 强	**27.84**
✓	3 强	26.43

与领域相关的相似性得分和统一监督 在表 5 中，我们可以看到使用共享相似性得分（等式 14）还是与领域相关的得分（等式 16）所带来的性能变化。我们还进行了一些实验如 SLIP [16] 和 DeCLIP [13]，根据领域组合分别形成正集和负集。正如预期的那样，在统一监督下，依赖于领域的相似性测量结果表现最佳。

表 5：ImageNet-1k 零拍准确率与领域相关的相似性得分和监督得分。

温度和偏移	监督	准确性
跨领域共享	统一	25.51
依赖领域	离职	26.59
依赖领域	统一	**27.84**

损失函数 正如第 2.2 节所分析的，SupCon 损失[10]优于 MIL-NCE 损失[15]，但性能不如多正版本的 InfoNCE 损失（等式 9），如表 6 所示。平衡权重 $w_{(D)(()(i,p)())}$ 可以提高性能，而且令人惊讶的是，我们只需在正集 P_i 中添加一对微不足道的配对 $(z_{(i)}, z_{(i)})$，就能显著提高性能，而额外的计算量几乎可以忽略不计。

图 6：不同损失函数下 ImageNet-1k 的零拍摄精度。

损失函数	准确性
MIL-NCE	22.23
超级会议	23.04
MP-NCE 无三元对（$z_{(i)}, z_{(i)}$）和 $w_{D(()(i,p)())}$ 公式 9）	24.60
MP-NCE w/o $w_{(D)(()i,p())}$ 公式 11）	26.41
MP-NCE	**27.84**

4 结论

我们提出了 UniCLIP，这是一个用于视觉语言预训练的统一框架，它通过将多个领域中定义的对比损失整合到一个单一的通用空间来提高数据效率。本文使用图像-文本数据集来验证我们的方法，因为视觉和语言是深度学习中最活跃的研究领域之一。虽然我们只在视觉-语言多模态数据集上进行了实验，但所提出的 UniCLIP 框架可以很容易地扩展到其他类型的多模态数据集，因为除了增强编码部分外，它是以一种与模态无关的方式设计的。将 UniCLIP 应用于不同类型的模态所需的所有特定模态知识就是将每种特定模态的增强功能描述为一个实向量，如第 2.1 节所述，这一点非常简单。至于 UniCLIP 框架在不同类型的多模态数据集上的应用效果如何，我们将留待今后的工作中进行研究。

致谢

这项工作得到了韩国政府资助的信息通信技术规划与评估研究所（IIPP）的资助（MSIT）。(No. 2022-0-00184, Development and Study of AI Technologies to Inexensive Conform to Evolving Policy on Ethics)

参考资料

[1] L.Bossard, M. Guillaumin, and L. V. Gool.食物-101-用随机森林挖掘辨别成分。*欧洲计算机视觉会议*，第 446-461 页。Springer, 2014.7

[2] S.Changpinyo, P. Sharma, N. Ding, and R. Soricut.概念 12m：推动网络规模图像-文本预训练以识别长尾视觉概念。*IEEE/CVF 计算机视觉与模式识别大会论文集*》，第 3558-3568 页，2021 年。7

[3] T.Chen、S. Kornblith、M. Norouzi 和 G. Hinton。视觉表征对比学习的简单框架。*国际机器学习会议*，第 1597-1607 页。PMLR, 2020.2, 4, 6

[4] X.Chen and K. He.探索简单的连体表示学习*IEEE/CVF 计算机视觉与模式识别大会论文集*》，第 15750-15758 页，2021 年。2

[5] X.Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick.Lin、R. Vedantam、S. Gupta、P. Dollár 和 C. L. Zitnick。Microsoft coco 字幕：
数据收集和评估服务器。*ArXiv 预印本 arXiv:1504.00325*, 2015。7

[6] M.Cimpoi、S. Maji、I. Kokkinos、S. Mohamed 和 A. Vedaldi。描述野生纹理In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606-3613, 2014.7

[7] A.Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Min- derer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby.一幅图像胜过 16x16 个单词：规模图像识别变换器。*国际学习表征会议*，2021 年。4

[8] L.飞飞、R. Fergus 和 P. Perona。从少量训练实例中学习生成视觉模型：在 101 个物体类别上测试的增量贝叶斯方法。*2004年计算机视觉与模式识别研讨会*，第178-178页。IEEE, 2004.7

[9] C.Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H.Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig.Sung, Z. Li, and T. Duerig.利用噪声文本监督扩展视觉和视觉语言表征学习。*国际机器学习大会*，第 4904-4916 页。PMLR, 2021.1

[10] P.Khosla、P. Teterwak、C. Wang、A. Sarna、Y. Tian、P. Isola、A. Maschinot、C. Liu 和 D. Krishnan。监督对比学习。*神经信息处理系统进展*》，33:18661-18673，2020.2, 4, 5, 9

[11] J.Krause, M. Stark, J. Deng, and L. Fei-Fei.细粒度分类的三维物体表示法。在*电气和电子工程师学会计算机视觉研讨会国际会议论文集*》，第 554-561 页，2013 年。7

[12] A.Krizhevsky, G. Hinton, et al.2009.7

[13] Y.Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan.监督无处不在：高效数据对比语言图像预训练范式。*国际学习表征会议*，2022 年。1, 2, 7, 9

[14] S.Maji、E. Rahtu、J. Kannala、M. Blaschko 和 A. Vedaldi。飞机的精细视觉分类。
arXiv preprint arXiv:1306.5151, 2013.7

[15] A.Miech、J.-B.Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman.从未整理的教学视频中端到端学习视觉表征。*IEEE/CVF 计算机视觉与模式识别大会论文集*》，第 9879-9889 页，2020 年。2, 4, 9

[16] N.Mu, A. Kirillov, D. Wagner, and S. Xie.Slip：自我监督满足语言图像预训练。
arXiv preprint arXiv:2112.12750, 2021.1, 8, 9

[17] M.-E. Nilsback 和 A. Zisserman.Nilsback 和 A. Zisserman.大量类别的自动花卉分类。In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722-729.IEEE, 2008.7

[18] O.O. M. Parkhi、A. Vedaldi、A. Zisserman 和 C. Jawahar。猫和狗。*2012 IEEE 计算机视觉与模式识别会议*，第 3498-3505 页。IEEE, 2012.7

[19] B.A. Plummer、L. Wang、C. M. Cervantes、J. C. Caicedo、J. Hockenmaier 和 S. Lazebnik。Flickr30k 实体：收集区域到短语的对应关系，建立更丰富的图像到句子模型。*IEEE 计算机视觉国际会议论文集*》，第 2641-2649 页，2015 年。7

[20] A.Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin、J.Clark, et al. Learning transferable visual models from natural language supervision.*国际机器学习大会*，第 8748-8763 页。PMLR, 2021.1, 6, 8

[21] O.O. Russakovsky、J. Deng、H. Su、J. Krause、S. Satheesh、S. Ma、Z. Huang、A. Karpathy、A. Khosla、M.Bernstein, et al. Imagenet large scale visual recognition challenge.*International journal of computer vision*，115（3）

：211-252，2015.7

[22] F.Schroff, D. Kalenichenko, and J. Philbin.Facenet：用于人脸识别和聚类的统一嵌入。In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815-823, 2015.4

[23] P.Sharma, N. Ding, S. Goodman, and R. Soricut.概念性标题：用于自动图像标题的经过清理的超文本图像 alt-文本数据集。In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556-2565, 2018.7

[24] K.Sohn.具有多类 n 对损失目标的改进型深度度量学习。*神经信息处理系统进展*》，29，2016.4

[25] B.Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li.Yfcc100m：多媒体研究中的新数据。*ACM 通信*》，59（2）：64-73，2016.7

[26] A.Van den Oord, Y. Li, and O. Vinyals.具有对比预测编码的表征学习》，*arXiv e-prints*，第 arXiv-1807 页，2018 年。4

[27] A.Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,.Kaiser 和 I. Polosukhin。注意力就是你所需要的一切*神经信息处理系统进展*》，2017年第30期。4

[28] J.Wei 和 K. Zou.Eda：提升文本分类任务性能的简易数据增强技术。In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382-6388, 2019.2

[29] J.Xiao、K. A. Ehinger、J. Hays、. Torralba 和 A. Oliva。太阳数据库：探索场景类别集合。*国际计算机视觉杂志*》，119（1）：3-22，2016。7

核对表

1. 对于所有作者...
   (a) 摘要和引言中提出的主要主张是否准确地反映了论文的贡献和范围？是
   (b) 您是否描述了您工作的局限性？有
   (c) 您是否讨论过您的工作可能产生的负面社会影响？[不适用]
   (d) 您是否阅读了伦理审查指南，并确保您的论文符合这些指南？是
2. 如果您将理论结果包括在内.....
   (a) 您是否说明了所有理论结果的全套假设？[不适用]
   (b) 您是否包含所有理论结果的完整证明？[不适用]
3. 如果你做实验...
   (a) 您是否包含了重现主要实验结果所需的代码、数据和说明（在补充材料中或作为 URL）？否
   (b) 您是否指定了所有训练细节（如数据分割、超参数、如何选择）？是
   (c) 您是否报告了误差条（例如，多次运行实验后与随机种子有关的误差）？没有
   (d) 是否包括计算总量和使用的资源类型（如 GPU 类型、内部集群或云提供商）？是
4. 如果您正在使用现有资产（如代码、数据、模型）或策划/发布新资产...
   (a) 如果您的作品使用了现有资产，您是否注明了创作者？有
   (b) 您提到资产许可证了吗？[N/A]
   (c) 您是否在补充材料或 URL 中包含了任何新资产？[否]
   (d) 您是否讨论过是否以及如何征得您正在使用/收集其数据的人的同意？[不适用]
   (e) 您是否讨论过您正在使用/收集的数据是否包含个人身份信息或冒犯性内容？[不适用]

5. 如果您使用众包或进行以人为对象的研究...

   (a) 您是否包含给参与者的说明全文和屏幕截图（如适用）？[不适用]

(b) 您是否描述了任何潜在的参与者风险，并提供了机构审查委员会 (IRB) 批准的链接（如适用）？ [不适用]

(c) 您是否包括支付给参与者的估计小时工资以及用于参与者补偿的总金额？ [不适用]