视觉语言模型在视觉任务上应用

邓钰川　2021141460159

目录

1 简介

随着深度学习的出现，通过利用可端到端训练的深度神经网络 (DNN)，视觉识别研究已经取得了巨大成功。然而，从传统机器学习向深度学习的转变带来了两个新的重大挑战，即深度神经网络 (DNN) 训练收敛缓慢，需要耗费大量时间进行训练；以及在 DNN 训练中，收集大规模、任务特定和众包标注的数据是一项劳动密集和耗时的过程。

近年来，一种新的学习范式“预训练、微调和预测”在各种视觉识别任务中显示出很好的效果。在这种新的范式下，首先使用某些现成的大规模训练数据 (有或没有标注) 对 DNN 模型进行预训练 [19,27,44]，然后再使用特定于任务的标注训练数据对预训练模型进行微调。通过在预训练模型中学习全面的知识，这种学习范式可以加速网络收敛，并训练出各种下游任务的表现良好的模型。

尽管如此，“预训练、微调和预测”的范式仍需要额外的任务特定微调阶段，以获取每个下游任务的标记训练数据。受到自然语言处理方面进展的启发 [3,11,38,39]，一种名为“视觉-语言模型预训练与零样本预测”的新型深度学习范式近来引起了越来越多的关注 [23,37,55]。

在这一范式中，视觉-语言模型（VLM）通过在互联网上几乎无限的图像-文本对上进行大规模预训练，并且不需要微调，即可直接应用于下游的视觉识别任务。VLM 的预训练通常受到一定的视觉-语言目标的指导 [37,55,56]，从大规模的图像-文本对中学习图像和文本之间的对应关系 [41,42]，例如，CLIP [37] 采用图像-文本对比目标，通过将配对的图像和文本在嵌入空间中拉近，将其他图像和文本推开来进行学习。这样，预训练的 VLM 可以捕捉丰富的视觉-语言对应知识，并通过匹配任意给定图像和文本的嵌入来进行零样本预测。这种新的学习范式能够有效利用网络数据，并且允许在不进行任务特定微调的情况下进行零样本预测，其实现简单而表现出色。

2 VLM 基础

VLM 预训练 [23,37] 旨在对视觉语言模型进行预训练，以学习图像和文本之间的关联性，并在视觉识别任务 [6,19,33,40] 中实现有效的零样本预测。通过给定的图像-文本对 [41,42]，它首先采用文本编码器和图像编码器提取图像和文本特征 [11,12,19,48]，然后利用特定的预训练目标学习视觉语言的关联性 [23,37]。通过学习到的视觉语言关联性，可以以零样本的方式对 VLM 进行未见数据的评估 [23,37]，通过匹配给定图像和文本的嵌入向量。在本节中，我们介绍了 VLM 预训练的基础知识，包括用于提取图像和文本特征的常见深度网络架构，用于建模视觉语言关联性的预训练目标，以及用于评估预训练 VLM 的下游任务。

2.1 网络结构

VLM 预训练使用了一个深度神经网络，该网络从预训练数据集 $\mathcal{D} = \{x_n^I, x_n^T\}_{n=1}^N$ 中提取图像和文本特征，其中 x_n^I 和 x_n^T 分别表示图像样本和与之配对的文本样本。该深度神经网络包括图像编码器 f_θ 和文本编码器 f_ϕ，它们将图像和文本（来自图像-文本对 $\{x_n^I, x_n^T\}$）编码为图像嵌入 $z_n^I = f_\theta(x_n^I)$ 和文本嵌入 $z_n^T = f_\phi(x_n^T)$。这节介绍了 VLM 与训练中广泛采用的深度神经网络架构。

2.1.1 图像编码器结构

两种类型的网络架构已被广泛采用来学习图像特征，即基于 CNN 的架构和基于 Transformer 的架构。

基于 CNN 的架构. 如图 1 所示，针对学习图像特征，设计了多种不同的卷积神经网络，包括 AlexNet [27]、VGG [44]、ResNet [19]、EfficientNet [46] 和 FBNet [10]。作为 VLM 预训练中最受欢迎的 CNN 网络之一，ResNet [19] 采用了卷积块之间的跳跃连接，有效缓解了梯度消失和梯度爆炸问题，并使得深度神经网络成为可能。为了更好地进行特征提取和视觉-语言相关建模，一些 VLM 研究 [37] 在原始网络架构 [19,46] 上进行了特定的修改。以 ResNet 为

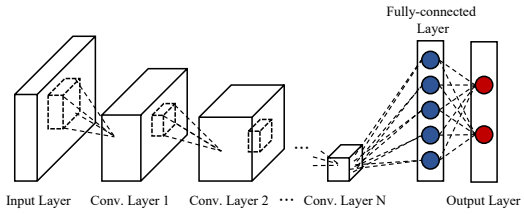

图 1: CNN 架构.

例，他们引入了 ResNet-D [20]，在 [58] 中使用了反锯齿的 rect-2 模糊池化，并将全局平均池化替换为 Transformer 多头注意力中的注意力池化 [48]。

基于 Transformer 的架构. 近年来，Transformers 已经在视觉识别任务中得到广泛的研究和探索，例如图像分类 [12,32]、目标检测 [5,61] 和语义分割 [51,59]。作为图像特征学习的标准 Transformer 架构，ViT [12] 采用了一系列的 Transformer 块，每个块由多头自注意层和前馈神经网络组成。如图 2所示，输入图像首先被分割成固定大小的补丁，然后通过线性投影和位置嵌入后被送入 Transformer 编码器中。在 VLM 研究中 [35,37,55]，在 Transformer 编码器之前加入了额外的归一化层，引入了微小的修改。

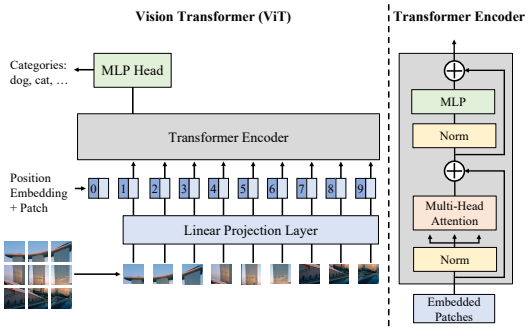

图 2: Vision Transformer 架构 [12].

2.1.2 文本编码器结构

Transformer 及其变种 [11,39,48] 已被广泛应用于学习文本特征。如图 3 所示，标准的 Transformer [48] 具有编码器-解码器结构，其中编码器由 6 个块组成，每个块都有一个多头自注意层和一个多层感知器（MLP）层。解码器也有 6 个块，每个块包括 3 个子层，包括一个多头注意力层，一个掩码多头层和一个 MLP 层。大多数 VLM 研究，如 CLIP [37]，采用标准 Transformer [48] 进行微小修改。

2.2 预训练目标

作为 VLM 的核心，各种视觉-语言预训练目标 [11,16,18,29,37,45,53,56] 旨在学习丰富的视觉-语言相关性。它们可以广泛分为三类，即对比目标、生成目标和对齐目标。

2.2.1 对比目标

对比目标通过在特征空间中拉近一对样本并将其与其他样本分开，训练 VLM 学习具有区分性的表示 [18,37,53]。

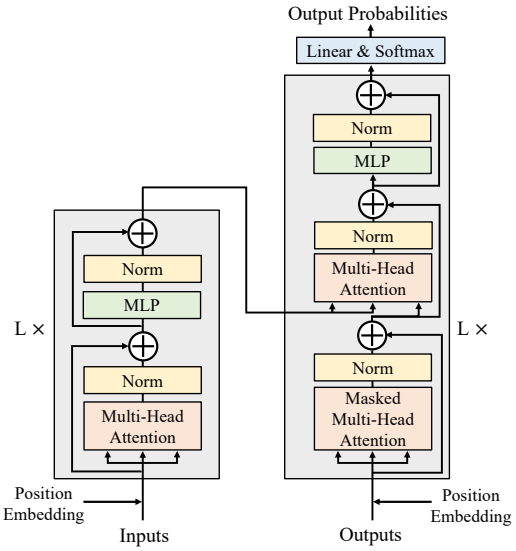

图 3: Transformer 示意图.

图像对比学习旨在通过将查询图像与其正样本键（即其数据增强）接近并与其负样本键（即其他图像）远离于嵌入空间中，来学习区分性图像特征 [7, 18]。给定 B 个图像批次，对比学习目标（比如 InfoNCE [36] 和其变体 [7, 18]）通常如下表示：

$$\mathcal{L}_I^{\text{InfoNCE}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(z_i^I \cdot z_+^I/\tau\right)}{\sum_{j=1,j\neq i}^{B+1}\exp(z_i^I \cdot z_j^I/\tau)}, \tag{1}$$

其中 z_i^I 是查询嵌入向量, $\{z_j^I\}_{j=1,j\neq i}^{B+1}$ 是键嵌入向量集合，其中 z_+^I 代表 z_i^I 的正样本键，其余为 z_i^I 负样本键。τ 是一个温度超参数，用于控制学习表示的密度。

图像-文本对比学习在通过将成对的图像和文本嵌入拉近并将其他对象推开的方式来学习区分性图像-文本表示 [23, 37]。通常，这是通过最小化一个对称的图像-文本信息归一化互信息（NT-Xent）损失函数来实现的 [37]。即，$\mathcal{L}_{\text{infoNCE}}^{IT} = \mathcal{L}_{I\to T} + \mathcal{L}_{T\to I}$, 其中 $\mathcal{L}_{I\to T}$ 将查询图像与文本关键字形成对比，而 $\mathcal{L}_{T\to I}$ 则是将查询文本与图像关键字对比。给定一批大小为 B 的图像-文本对, $\mathcal{L}_{I\to T}$ 和 $\mathcal{L}_{T\to I}$ 定义如下:

$$\mathcal{L}_{I\to T} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(z_i^I \cdot z_i^T/\tau\right)}{\sum_{j=1}^{B}\exp(z_i^I \cdot z_j^T/\tau)}, \tag{2}$$

$$\mathcal{L}_{T\to I} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(z_i^T \cdot z_i^I/\tau\right)}{\sum_{j=1}^{B}\exp(z_i^T \cdot z_j^I/\tau)}, \tag{3}$$

其中 z^I 和 z^T 分别表示图像和文本嵌入。

图像-文本-标签对比学习. 图像-文本-标签对比学习 [53] 将有监督对比学习 [25] 引入到图像-文本对比学习中，通过对 Eqs. 2 和 3 进行重新定义来实现：

$$\mathcal{L}_{I\to T}^{ITL} = -\sum_{i=1}^{B}\frac{1}{|\mathcal{P}(i)|}\sum_{k\in\mathcal{P}(i)}\log\frac{\exp\left(z_i^I \cdot z_k^T/\tau\right)}{\sum_{j=1}^{B}\exp(z_i^I \cdot z_j^T/\tau)}, \tag{4}$$

$$\mathcal{L}_{T\to I}^{ITL} = -\sum_{i=1}^{B}\frac{1}{|\mathcal{P}(i)|}\sum_{k\in\mathcal{P}(i)}\log\frac{\exp\left(z_i^T \cdot z_k^I/\tau\right)}{\sum_{j=1}^{B}\exp(z_i^T \cdot z_j^I/\tau)}, \tag{5}$$

其中 $k \in \mathcal{P}(i) = \{k | k \in B, y_k = y_i\}$ [53] 和 y 是 (z^I, z^T) 子类标签。通过 Eqs. 4 and 5, 图像-文本-标签信息归一化互信息（NT-Xent）损失函数定义如下: $\mathcal{L}^{ITL}_{\text{infoNCE}} = \mathcal{L}^{ITL}_{I \to T} + \mathcal{L}^{ITL}_{T \to I}$.

2.2.2 生成目标

生成目标通过训练网络通过图像生成 [1,18]、语言生成 [11,31,56] 或跨模态生成 [45] 来学习语义特征。

遮蔽图像建模通过遮蔽和重构图像来学习跨补丁相关性 [1,17]。它随机遮蔽输入图像的一组补丁，并训练编码器在未遮蔽补丁的条件下重构遮蔽补丁。给定一批大小为 B 的图像，损失函数可以表示为:

$$\mathcal{L}_{MIM} = -\frac{1}{B}\sum_{i=1}^{B} \log f_\theta(\ \overline{x}_i^I \mid \hat{x}_i^I\ ), \tag{6}$$

其中 $\overline{x}_i^I$ 和 $\hat{x}_i^I$ 分别表示 x_i^I 中的遮蔽补丁和未遮蔽补丁。

遮蔽语言建模是 NLP 领域中广泛采用的预训练目标 [11,31]。它会随机遮蔽输入文本中一定比例的标记（例如，在 BERT [11] 中为 15%），然后用未遮蔽的标记进行重构:

$$\mathcal{L}_{MLM} = -\frac{1}{B}\sum_{i=1}^{B} \log f_\phi(\ \overline{x}_i^T \mid \hat{x}_i^T\ ), \tag{7}$$

其中 $\overline{x}_i^T$ 和 $\hat{x}_i^T$ 分别表示 x_i^T 中的遮蔽标记和未遮蔽标记。B 表示批次大小。

遮蔽跨模态建模结合了遮蔽图像建模和遮蔽语言建模 [45]。给定一个图像-文本对，它会随机遮蔽图像中的一部分补丁和文本中的一部分标记，然后学习在未遮蔽图像补丁和未遮蔽文本标记的条件下进行重构，具体如下:

$$\mathcal{L}_{MCM} = -\frac{1}{B}\sum_{i=1}^{B}[\log f_\theta(\overline{x}_i^I | \hat{x}_i^I, \hat{x}_i^T) + \log f_\phi(\overline{x}_i^T | \hat{x}_i^I, \hat{x}_i^T)], \tag{8}$$

其中 $\overline{x}_i^I / \hat{x}_i^I$ 表示 x_i^I 中的遮蔽/未遮蔽补丁, $\overline{x}_i^T / \hat{x}_i^T$ 表示 x_i^T 中的遮蔽/未遮蔽文本标记。

图像到文本生成旨在基于与文本 x^T 配对的图像，自回归地预测文本 x^T [56]:

$$\mathcal{L}_{ITG} = -\sum_{l=1}^{L} \log\ f_\theta(x^T \mid x^T_{<l}, z^I), \tag{9}$$

其中，L 表示要预测的 x^T 的标记数量，z^I 是与 x^T 配对的图像的嵌入。

2.2.3 对齐目标

对齐目标通过全局图像-文本匹配 [2,13] 或在嵌入空间上的局部区域-单词匹配 [29,54] 来对齐图像-文本对。

Image-Text Matching 模型对图像和文本之间的全局相关性进行建模 [2,13]，可以通过一个评分函数 $\mathcal{S}(\cdot)$ 来度量图像和文本之间的对齐概率，并使用二分类损失进行建模:

$$\mathcal{L}_{IT} = p \log \mathcal{S}(z^I, z^T) + (1-p)\log(1 - \mathcal{S}(z^I, z^T)), \tag{10}$$

其中 p 等于 1 表示图像和文本配对，等于 0 表示不配对。

局部区域-单词匹配局部区域-单词匹配旨在对图像-文本对 [29,54] 中的局部跨模态相关性（即“图像区域”和“单词”之间的关系）进行建模，用于密集视觉识别任务，如目标检测。它可以表示为:

$$\mathcal{L}_{RW} = p \log \mathcal{S}^r(r^I, w^T) + (1-p)\log(1 - \mathcal{S}^r(r^I, w^T)), \tag{11}$$

其中 (r^I, w^T) 表示一个区域-单词对，$p = 1$ 表示该区域和单词配对，否则 $p = 0$。$\mathcal{S}^r(\cdot)$ 表示一个局部评分函数，用于衡量“图像区域”和“单词”之间的相似度

2.3 评估设置和下游任务

本节介绍 VLM 评估中广泛采用的设置和下游任务。设置包括零样本预测和线性探测，下游任务包括图像分类、对象检测、语义分割、图像-文本检索和动作识别。

2.3.1 零样本预测

作为评估视觉语言模型（VLMs）泛化能力的最常见方式 [23,28,35,37,55]，零样本预测直接将预训练的 VLMs 应用于下游任务，无需进行任何特定任务的微调 [37]。

图像分类 [19,44] 的目标是将图像分类到预定义的类别中。视觉语言模型通过比较图像和文本的嵌入来实现零样本图像分类，其中通常会使用“提示工程”生成与任务相关的提示，例如“一个 [label] 的照片”。 [37]。

语义分割 [6] 的目标是为图像中的每个像素分配一个类别标签。预训练的视觉语言模型通过比较给定图像像素和文本的嵌入来实现分割任务的零样本预测。

目标检测 [15,40] 的目标是在图像中定位和分类物体，这对于各种视觉应用非常重要。通过从辅助数据集 [24,43] 中学习到的目标定位能力，预训练的视觉语言模型可以通过比较给定的目标提议和文本的嵌入来实现物体检测任务的零样本预测。

图像-文本检索 [4] 的目标是根据另一种模态的线索从一种模态中检索所需的样本，它包括两个任务，即文本到图像的检索（根据文本检索图像）和图像到文本的检索（根据图像检索文本）。

2.3.2 线性探测

线性探测（Linear Probing）在 VLM 的评估中被广泛采用 [37]。它会冻结预训练的 VLM，并训练一个线性分类器来对 VLM 编码的嵌入进行分类，以评估 VLM 的表示能力。图像分类 [19,44] 和动作识别 [21,49] 在此类评估中被广泛采用，其中在动作识别任务中，视频片段经常被子采样以实现高效的识别 [37]。

3 视觉语言模型预训练

3.1 具有对比目标的 VLM 预训练

在 VLM 预训练中，对比学习被广泛应用，为学习具有辨别性的图像-文本特征设计了对比目标 [30,35,37]。

3.1.1 图像对比学习

这种预训练目标旨在学习图像模态中的有辨别性特征，通常作为充分发挥图像数据潜力的辅助目标。例如，SLIP [35] 使用了定义在公式 1中的标准 infoNCE 损失函数来学习有辨别性的图像特征。

3.1.2 图像-文本对比学习

图像-文本对比旨在通过对比图像-文本对来学习视觉语言关联性，即将成对的图像和文本嵌入拉近，而将其他嵌入推远 [37]。例如，CLIP [37] 使用了在公式 2中对称的图像-文本 infoNCE 损失函数，通过图 4中的图像和文本嵌入之间的点积来衡量它们之间的相似度。因此，预训练的 VLM 学习了图像-文本相关性，从而能够进行零样本预测和下游的视觉识别任务。

受到 CLIP 的巨大成功启发，许多研究从不同角度对对称的图像-文本 infoNCE 损失进行了改进。例如，ALIGN [23] 通过使用大规模（18 亿个）但带有噪声的图像-文本对和噪声鲁棒对比学习来扩大 VLM 预训练的规模。一些研究 [9, 30, 50] 则通过更少的图像-文本对来探索高效的 VLM 预训练方法。例如，DeCLIP [30] 引入最近邻监督来

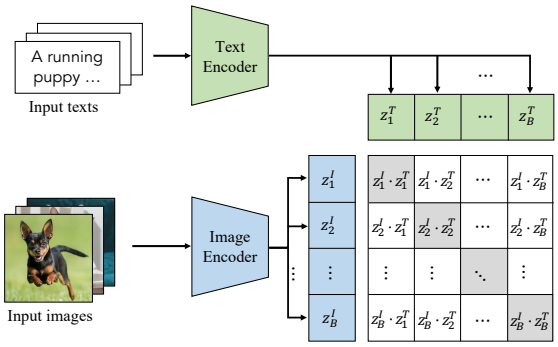

图 4: CLIP [37] 中的图像-文本对比学习说明。

利用相似对之间的信息，使得有限数据上的预训练变得有效。OTTER [50] 使用最优传输方法来伪对图像和文本，大大减少了所需的训练数据量。ZeroVL [9] 通过去偏数据采样和使用 coin flipping mixup 进行数据增强来利用有限的数据资源。

另一条后续研究的线索 [14, 52, 55] 是通过在各个语义层面上进行图像-文本对比学习，以实现全面的视觉语言关联建模。例如，FILIP [55] 将区域-词对齐引入对比学习中，能够学习到细粒度的图像-文本对应知识。Pyramid-CLIP [14] 构建了多个语义层次，并进行跨层级和同层级的对比学习，以进行有效的 VLM 预训练。

3.1.3 图像-文本-标签对比学习

这种类型的预训练将图像分类标签 [53] 引入了图像-文本对比学习中，如公式 4 所定义，将图像、文本和分类标签编码到一个共享空间中，如图 5 所示。它同时利用了带有图像标签的监督预训练和带有图像-文本对的无监督 VLM 预训练。正如 UniCL [53] 所报道的，这种预训练可以同时学习具有区分性的和任务特定的（即图像分类）特征。随后在 [57] 中对 UniCL 进行了扩展，使用约 9 亿个图像-文本对，在各种下游识别任务中取得了卓越的性能。

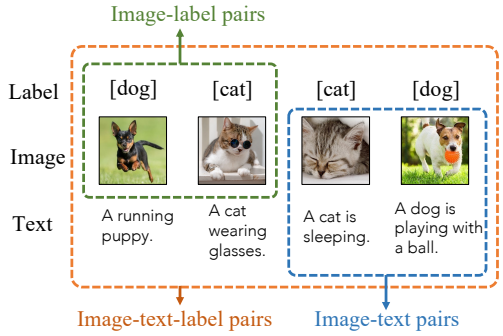

图 5: UniCL [53] 中提出的图像-文本标签空间的说明。

3.2 具有生成目标的 VLM 预训练

生成式 VLM 预训练通过学习通过遮蔽图像建模、遮蔽语言建模、遮蔽跨模态建模和图像到文本生成等方式生成图像或文本来学习语义知识。在遮蔽图像建模中，模型需要根据部分遮蔽的图像去预测缺失的部分。在遮蔽语言建模中，模型需要根据遮蔽的文本片段去预测被遮蔽的词语。在遮蔽跨模态建模中，模型需要将遮蔽的图像和文本

组合起来进行预测任务。在图像到文本生成中，模型需要根据给定的图像生成相应的文本描述。这些生成式任务使得模型能够学习到视觉和语言之间的语义关联和生成能力，从而提高了其对图像和文本之间的理解和生成能力。

3.2.1 蒙版图像建模

这种预训练目标通过对图像进行遮蔽和重构来引导学习图像上下文信息，如公式 6所示。在遮蔽图像建模中（例如，MAE [17] 和 BeiT [1]），图像的某些区域被遮盖，编码器被训练在未遮盖的区域上给出遮盖区域的重构，如图 6所示。例如，FLAVA [45] 采用了类似于 BeiT 的矩形块遮盖方式，而 KELIP [26] 和 SegCLIP [34] 则使用 MAE 来遮盖训练中的大部分区域（即 75%）。

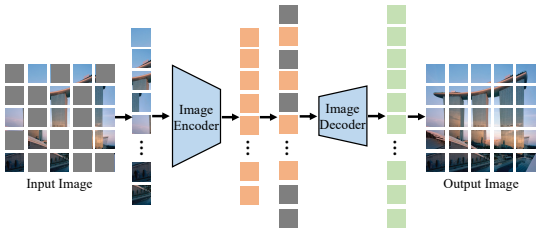

图 6: 蒙版图像建模插图。

3.2.2 遮蔽语言建模

遮蔽语言建模是 NLP 领域中被广泛采用的预训练目标，如公式 7所示，在 VLM 预训练中也证明了其在文本特征学习中的有效性。它通过对每个输入文本中的一部分标记进行遮蔽，并训练网络来预测被遮蔽的标记，如图 7所示。FLAVA [45] 遵循了 [11] 的方法，遮盖了 15% 的文本标记，并从剩余标记中重构它们以建模跨词关联。FIBER [13] 采用了遮蔽语言建模 [11] 作为 VLM 预训练目标之一，以提取更好的语言特征。

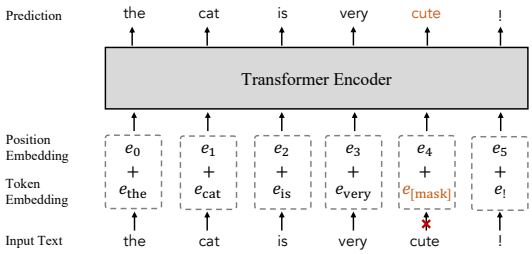

图 7: 遮蔽语言建模插图

3.2.3 遮蔽跨模态建模

遮蔽跨模态建模通过同时对图像块和文本标记进行遮蔽和重构，如公式 8所示，继承了遮蔽图像建模和遮蔽语言建模的优点。它通过遮蔽一定比例的图像块和文本标记，并训练 VLM 来基于未遮蔽的图像块和文本标记的嵌入来重构它们。例如，FLAVA [45] 遮盖了约 40% 的图像块，如 [1] 所示，以及 15% 的文本标记，如 [11] 所示，然后使用多层感知机（MLP）来预测遮盖的图像块和文本标记，捕捉丰富的视觉-语言对应信息。

3.2.4 图像到文本生成

图像到文本生成（即图像字幕生成）旨在为给定的图像生成描述性文本，通过训练 VLM 来预测准确的分词文本，以捕捉细粒度的视觉-语言相关性。它首先将输入图像编码为中间嵌入向量，然后根据公式 9将其解码为描述性文本。例如，COCA [56]、NLIP [22] 和 PaLI [8] 使用标准的编码器-解码器架构和图像字幕生成目标来训练 VLMs，如图 8所示。

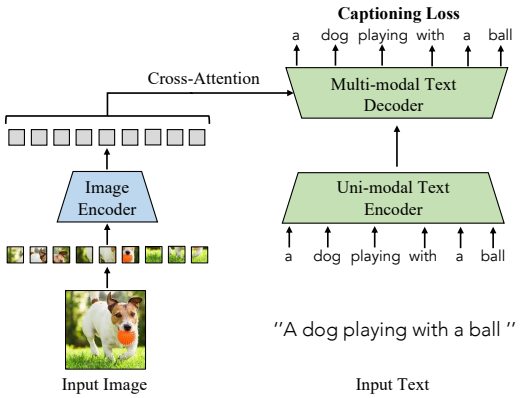

图 8: 图像到文本生成插图

3.3 具有对齐目标的的 VLM 预训练

对齐目标使得视觉语言模型（VLM）通过学习预测给定的文本是否正确描述了给定的图像来对齐配对的图像和文本。它可以大致分为全局图像-文本匹配和局部区域-词语匹配两个方面，用于 VLM 的预训练。

3.3.1 图像-文本匹配

图像-文本匹配模型通过直接对配对的图像和文本进行对齐来建立全局图像-文本相关性，具体定义在公式 10 中。例如，给定一批图像-文本对，FLAVA（参考文献：Singh 等，2022）通过分类器和二分类损失来学习将给定的图像与其配对的文本进行匹配。FIBER（参考文献：Dou 等，2020）遵循了 Bao 等人（2021）的方法，利用成对相似度来挖掘难例负样本，以实现更好地对齐图像和文本之间的表示。

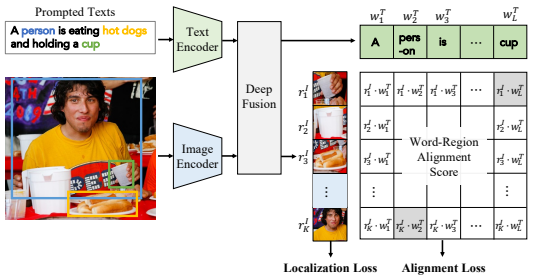

图 9: GLIP 示意图。

3.3.2 区域-词语匹配

区域-词语匹配目标通过对齐配对的图像区域和词语标记来建模局部细粒度的视觉-语言相关性，极大地改善了零样本密集预测中的目标检测和语义分割。举个例子，GLIP [29], FIBER [13] 和 DetCLIP [54] 通过将物体分类的逻辑替换为区域-词语对齐分数，即区域视觉特征和标记特征之间的点积相似度。这一过程如图 9 所示。

4 未来方向

对于 **VLM 预训练**，有三个挑战和潜在的研究方向。

- *Fine-grained vision-language correlation modelling.* 凭借本地视觉语言对应知识 [29,54], 视觉语言模型（VLM）可以更好地识别图像以外的补丁和像素，极大地增强了物体检测和语义分割等密集预测任务的效果，这些任务在各种视觉识别任务中起着重要作用。鉴于在这个方向上的 VLM 研究非常有限 [13,29,34,52,54,60]，我们期待在细粒度 VLM 预训练方面进行更多的研究，以应对零样本密集预测任务的挑战。

- *Unification of vision and language learning.* Transformer 的出现 [12, 48] 使得图像和文本学习可以在一个单一的 Transformer 中进行统一，通过以相同的方式对图像和文本进行分词处理。与现有的 VLMs（视觉语言模型）使用两个独立网络不同 [23, 37]，融合视觉和语言学习可以实现跨数据模态的高效通信，这有助于提高训练效果和效率。这个问题已经引起了一些关注 [47]，但需要更多的努力来构建可持续发展的 VLMs。

- *Data-efficient VLMs.* 现有的大部分工作都是使用大规模的训练数据和密集的计算来训练 VLM（视觉语言模型），这使得其可持续性成为一个重要问题。通过有限的图像-文本数据训练出有效的 VLM 可以很大程度上缓解这个问题。例如，不仅可以从每个图像-文本对中学习，还可以通过图像-文本对之间的监督关系学习到更有用的信息 [30, 50]。这种方法可以提高数据的利用率，减少对大规模标注数据的依赖，从而在有限的数据情况下训练出更加有效的 VLM。

5 结论

视觉识别的视觉语言模型能够有效地使用 web 数据，并允许在没有特定任务微调的情况下进行零样本预测，这很容易实现，但在广泛的识别任务中取得了巨大成功。通过视觉语言模型，我们可以更加高效准确地进行各种视觉识别任务，这将帮助我们取得更多的科学进展、提高产品质量、提升生活品质等。同时，我们也需要注意保护数据隐私和遵守国家法律法规等方面的问题。在未来，视觉语言模型还会不断地发展和完善，在更多领域中发挥巨大的作用，助力人类更好地生活和工作。

参考文献

[1] Bao, H., Dong, L., Piao, S., and Wei, F. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254* (2021).

[2] Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O. K., Aggarwal, K., Som, S., and Wei, F. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358* (2021).

[3] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems 33* (2020), 1877–1901.

[4] Cao, M., Li, S., Li, J., Nie, L., and Zhang, M. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713* (2022).

[5] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16* (2020), Springer, pp. 213–229.

[6] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence 40*, 4 (2017), 834–848.

[7] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (2020), PMLR, pp. 1597–1607.

[8] Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794* (2022).

[9] Cui, Q., Zhou, B., Guo, Y., Yin, W., Wu, H., Yoshie, O., and Chen, Y. Contrastive vision-language pre-training with limited resources. In *European Conference on Computer Vision* (2022), Springer, pp. 236–253.

[10] Dai, X., Wan, A., Zhang, P., Wu, B., He, Z., Wei, Z., Chen, K., Tian, Y., Yu, M., Vajda, P., et al. Fbnetv3: Joint architecture-recipe search using predictor pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 16276–16285.

[11] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[12] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[13] Dou, Z.-Y., Kamath, A., Gan, Z., Zhang, P., Wang, J., Li, L., Liu, Z., Liu, C., LeCun, Y., Peng, N., et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *Advances in Neural Information Processing Systems*.

[14] Gao, Y., Liu, J., Xu, Z., Zhang, J., Li, K., and Shen, C. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *arXiv preprint arXiv:2204.14095* (2022).

[15] Girshick, R. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (2015), pp. 1440–1448.

[16] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377* (2021).

[17] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 16000–16009.

[18] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 9729–9738.

[19] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.

[20] He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., and Li, M. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 558–567.

[21] Herath, S., Harandi, M., and Porikli, F. Going deeper into action recognition: A survey. *Image and vision computing 60* (2017), 4–21.

[22] Huang, R., Long, Y., Han, J., Xu, H., Liang, X., Xu, C., and Liang, X. Nlip: Noise-robust language-image pre-training. *arXiv preprint arXiv:2212.07086* (2022).

[23] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning* (2021), PMLR, pp. 4904–4916.

[24] Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., and Carion, N. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 1780–1790.

[25] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. *Advances in neural information processing systems 33* (2020), 18661–18673.

[26] Ko, B., and Gu, G. Large-scale bilingual language-image contrastive learning. *arXiv preprint arXiv:2203.14463* (2022).

[27] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM 60*, 6 (2017), 84–90.

[28] Lee, J., Kim, J., Shon, H., Kim, B., Kim, S. H., Lee, H., and Kim, J. Uniclip: Unified framework for contrastive language-image pre-training. In *Advances in Neural Information Processing Systems*.

[29] Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 10965–10975.

[30] Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations* (2021).

[31] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[32] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 10012–10022.

[33] Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 3431–3440.

[34] Luo, H., Bao, J., Wu, Y., He, X., and Li, T. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2211.14813* (2022).

[35] Mu, N., Kirillov, A., Wagner, D., and Xie, S. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision* (2022), Springer, pp. 529–544.

[36] Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[37] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (2021), PMLR, pp. 8748–8763.

[38] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training.

[39] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog 1*, 8 (2019), 9.

[40] Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems 28* (2015).

[41] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402* (2022).

[42] Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* (2021).

[43] Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., and Sun, J. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 8430–8439.

[44] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[45] Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 15638–15650.

[46] Tan, M., and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (2019), PMLR, pp. 6105–6114.

[47] Tschannen, M., Mustafa, B., and Houlsby, N. Image-and-language understanding from pixels only. *arXiv preprint arXiv:2212.08045* (2022).

[48] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems 30* (2017).

[49] Wang, H., and Schmid, C. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision* (2013), pp. 3551–3558.

[50] Wu, B., Cheng, R., Zhang, P., Gao, T., Gonzalez, J. E., and Vajda, P. Data efficient language-supervised zero-shot recognition with optimal transport distillation. In *International Conference on Learning Representations* (2021).

[51] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems 34* (2021), 12077–12090.

[52] Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., and Wang, X. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 18134–18144.

[53] Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., Yuan, L., and Gao, J. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 19163–19173.

[54] Yao, L., Han, J., Wen, Y., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, C., and Xu, H. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In *Advances in Neural Information Processing Systems*.

[55] Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., and Xu, C. Filip: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations* (2021).

[56] Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* (2022).

[57] Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432* (2021).

[58] Zhang, R. Making convolutional networks shift-invariant again. In *International conference on machine learning* (2019), PMLR, pp. 7324–7334.

[59] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H., et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 6881–6890.

[60] Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L. H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 16793–16803.

[61] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* (2020).