arXiv:2304.04399v1 [cs.CV] 10 Apr 2023

CAVL：学习视觉与语言的对比式自适应表征

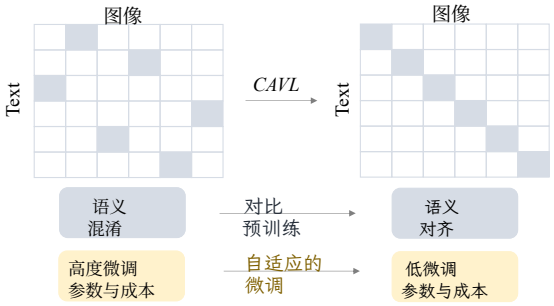
Shentong Mo*、Jingfei Xia*、Ihor Markevych 卡内基梅隆大学

摘要

视觉与语言预训练旨在共同学习视觉与语言表征，以便迁移到视觉语言下游任务。然而，在预训练阶段，语言与视觉之间存在语义混淆。此外，当前的预训练模型迁移到下游任务时，微调往往需要消耗大量计算资源。在本文中，我们提出了一种简单而有效的方法，用于学习视觉与语言的对比和自适应表征，即 *CAVL*。具体而言，在预训练过程中，我们引入成对对比损失，以学习整个句子与同一批次中每幅图像之间的对齐。在微调阶段，我们引入两个轻量级适配网络，以减少模型参数并提高训练速度，从而节省计算资源。我们在六项主要下游任务上评估了 *CAVL*，包括视觉问答（*VQA*）、视觉常识推理（*VCR*）、视觉推理自然语言（*NLVR*）、区域到短语定位（*RPG*）、文本到图像检索（*TIR*）和零样本文本到图像检索（*ZS-TIR*）。与基线相比，我们取得了更优的性能，并大幅缩短了微调时间（具体而言，缩短了 *76.17%*）。大量实验和消融研究表明，我们提出的 *CAVL* 中的对比预训练和自适应微调是有效的。

1. 引言

视觉与语言表征预训练 [24,27,40] 一直是多模态领域的热门研究方向。这是因为，通过利用现有预训练模型生成的特征，无需花费大量计算时间来建模语言与视觉分布，即可使用预训练模型，在各种任务上取得可媲美最先进水平的结果。更好的预训练模型还可用于



Model	Fine-tune Params (M)↓	Fine-tune Cost (h)↓
VisualBERT [24]	113.90	24.5
VL-BERT$_{base}$ [40]	115.04	25.6
VL-BERT$_{large}$ [40]	342.55	70.2
CAVL (ours)	**46.70**	**6.1**

图1。视觉-语言模型中存在的两个主要问题（语义混淆和高昂的微调成本）示意图，其中我们提出了 **CAVL**（对比预训练和自适应微调）来解决这些问题。

视觉和语言领域中的多种任务，例如视觉问答（VQA）[10]和视觉常识推理（VCR）[50]。在为不同视觉语言任务设计的各种架构中，一个关键点是聚合视觉和语言领域中的多模态信息。然而，在预训练阶段，视觉与语言之间存在语义混淆，也就是说，物体/实体或图像/文本之间存在错位。另一个问题是，迁移到下游任务时，预训练模型往往需要耗费大量训练时间和资源进行微调。

在这项工作中，我们提出了一个简单但有效的框架，用于学习**C**对比式与**A**自适应的**V**视觉和**L**语言表征，称为**CAVL**，该框架包括对比式预训练和自适应微调。具体而言，为了消除预训练期间语言与视觉之间的错位，我们采用成对对比损失（PwCL），学习完整句子与每张图像之间的对齐；对于正确配对，我们最大化视觉与语言嵌入的余弦相似度，同时最小化错误配对嵌入的余弦相似度。为了进一步消除对大量训练集的需求

*这些作者对本研究作出了同等贡献。

在微调阶段，我们在CAVL中引入了两个轻量级适配网络，以学习自适应表示。一个适配器使用捷径模块来获得任务特定特征，并将预训练模型中的泛化特征与输出块融合；另一个适配器则在 BERT 的每一层中的注意力和前馈模块之后应用瓶颈结构。我们的CAVL不仅大幅减少了训练集参数，而且还将性能保持在具有竞争力的水平。

我们在四个主要下游任务上进行了广泛的实验：VQA、VCR、视觉推理自然语言和区域到短语定位。与此前最先进模型相比 [7, 24, 27, 38, 40, 42],我们的 CAVL 在迁移到视觉语言下游任务时取得了相当甚至更好的性能。对比预训练帮助我们的 CAVL更好地理解图像与文本之间的关系，从而改善了下游任务的结果。关于自适应微调的消融研究也证明了所提出的自适应微调在节省计算资源方面的有效性和效率。

总体而言，我们在本文中的主要贡献可总结如下：

- 我们提出了一种简单但有效的方法，用于在预训练期间学习视觉表征与语言表征之间的对齐，即 CAVL。
- 我们在 CAVL中提出了两个轻量级适配网络，以进一步减轻微调阶段对大量计算资源的需求。
- 我们的 CAVL 在迁移到六项主要视觉语言下游任务时取得了更优异的结果。
- 大量消融研究表明，自适应微调能在取得相当性能的同时，有效减少训练参数。

2. 相关工作

视觉表征预训练。近年来，视觉表征预训练已应用于许多下游任务，例如图像分类、目标检测和分割。通常，对比式自监督学习 [4–6,17,29–31] 是学习有意义的视觉表征的常用方法之一。大多数以往方法都利用与图像配对的文本，通过无监督、自监督、弱监督和监督方式学习视觉表征。由于语言和视觉可以共享相似的语义，CLIP [34] 是一种常用的神经网络，使用各种（图像、文本）对进行训练，从自然语言监督中学习可迁移的视觉表征。在自然语言指令和与任务无关的优化方式的指导下，给定一张图像，CLIP 可以预测与之最相关的文本片段，这与 GPT-2 [36]和 GPT-3 [3]的零样本能力类似。Huo 等人 [14] 采用了一种名为 BriVL 的跨模态对比学习框架进行图文预训练。与采用简单对比学习方法的 CLIP 不同，他们利用 MoCo [11]处理跨模态场景，通过构建大型队列式字典，在预训练时引入批次中的负样本。在这项工作中，多模态场景下视觉嵌入与语言嵌入之间的错位，在预训练阶段通过成对对比损失得到缓解。

语言表征预训练。在语言建模领域的文献中，语言表征预训练主要有两种框架，包括 BERT [8]和 GPT [35]。BERT [8]是一种基于 Transformer [43] 的预训练模型，借助自注意力模块，在许多自然语言处理任务上取得了最先进水平。GPT [35] 是另一种基于 Transformer [43]的语言建模架构。GPT 模型的输入以字节为单位表示，空格除外，因此能够处理可变词汇表，并在训练时处理未知词元。在这项工作中，我们提出了一种简单而有效的 BERT 框架，并在四项主要视觉语言下游任务上评估了我们的网络。

视觉与语言表征融合预训练。已有大量工作 [7, 12, 15, 22, 24,25, 27, 33, 39–42, 48, 49, 51] 聚焦于视觉与语言表征融合预训练。通常，UNITER [7]旨在通过对多模态输入使用 Transformer，并以掩码语言建模（MLM）、掩码区域建模、图文匹配和词语—区域对齐作为预训练任务，学习联合图文嵌入。ERNIE-ViL [49]利用视觉任务的场景图与视觉—语言联合分布，预测从句子中解析出的不同类型场景图节点。LXMERT [42]是一种基于 Transformer 的模型，包含多个编码器：对象关系编码器、语言编码器和跨模态编码器。VisualBERT [24]是一个简单灵活的视觉与语言任务框架，使用 BERT 结构中的自注意力模块，将视觉中的图像嵌入与语言中的文本嵌入结合起来。VilBERT [27] 通过使用两个并行的 BERT 类模型分别处理文本片段和图像区域，对传统 BERT 进行了扩展。VL-BERT [40] 是另一种基于 BERT 的模型，它使用图像中的感兴趣区域和子词信息，并通过图像线索预测掩码词语、通过文本线索预测掩码区域来预训练模型。为解决噪声标签和领域偏差问题，CVLP [38] 在视觉分支中引入对比损失，以区分正样本和负样本。最近，DocFormer [2] 是

提出该方法以强制视觉、文本和空间特征之间进行多模态交互，从而实现视觉文档理解。然而，我们在视觉和语言分支中均采用成对对比损失，以消除预训练期间整个句子与各个图像之间的错位。

请注意，我们的 CAVL 与近期的一项工作 ConVIRT 不同 [52], 主要体现在两个方面。1）与 CLIP 类似，ConVIRT 将对比损失应用于训练一个模型，使其在自然语言监督下能够更好地执行医学图像分类，而在本文中，我们专注于解决视觉和语言领域中的错位问题。2）他们分别通过图像编码器和文本编码器生成视觉和文本嵌入。然而，我们将两个编码器结合到单个BERT 模型中，并表明成对对比学习能够使模型学习到更好对齐的图像与文本表示。此外，我们的 PwCL 损失不同于 Unicoder-VL 中提出的 VLM损失 [23] 之处有两点。1）Unicoder-VL引入全连接层来预测整幅图像与句子之间的分数，而我们的PwCL通过简单的点积计算余弦相似度，无需任何额外参数。2）VLM 损失仅采样一个负样本（图像或图像描述），并应用交叉熵损失，将其作为二分类来学习对齐。然而，我们在整个批次（批次大小为 $B^2 - B$ ）中使用负样本 B 进行预训练，从而更充分地解决错位问题。

Transformer中的适配网络。 在对BERT 模型进行微调时，适配是一种重要方法，它能够通过以更少的时间和计算资源更新少得多的参数来取得颇具潜力的结果。这促使研究人员在基于 Transformer的模型中采用高效的适配方法。MAD-X [32] 将多语言模型适配到任意任务和语言，作者提出了一种可逆适配器，并展示了其在不同语言和任务上的良好性能。Houlsby 等人 [13]在 Transformer层内部提出了一个中间层，并在冻结所有其他参数的情况下训练该中间层。在这项工作中，我们提出了两种轻量级适配器，以在特定领域任务上实现具有竞争力的性能，同时显著减少微调参数和成本。

3. 方法

3.1. 预备知识：BERT

BERT [8], 是 Bidirectional Encoder Representations from Transformers（来自 Transformer 的双向编码器表示）的缩写，是一种以词元作为输入并优化语言建模目标的模型。所有输入词元都会被映射到一组嵌入中，并经过多个“编码器风格”的模块，以生成表示。输入嵌入计算为词元嵌入、片段嵌入、和位置嵌入。然后，输入嵌入会经过多层 Transformer，其中每一层由两个模块组成：多头注意力模块和前馈模块。每个模块之后都连接一个全连接网络，并共同封装在残差加法中。

BERT 的训练过程主要分为两个步骤：预训练和微调。预训练步骤使用了两种语言建模目标：（1）掩码语言模型（MLM）。某个输入词的嵌入会被随机掩码，然后使用 BERT 根据未掩码的词预测被掩码的词。（2）下一句预测（NSP）。给定两个句子，训练 BERT 识别它们是否是同一文档中的连续句子。最后，为了将 BERT 应用于特定任务，我们引入特定于任务的输入、输出层和目标，并使用特定数据集对模型进行微调。

3.2. CAVL

在本部分中，我们提出了一种简单而有效的方法，用于学习视觉和语言的对比式自适应表征，即 CAVL。该方法由对比预训练和自适应微调组成。具体而言，对比预训练用于缓解预训练阶段视觉表征与语言表征之间的语义混淆。在微调过程中将模型迁移至下游任务时，自适应微调无需耗费大量训练时间和占用大量 GPU 内存。

3.2.1 对比预训练

对于对比预训练，我们在基于 BERT 的 Transformer 中采用自注意力机制，以对比的方式显式对齐输入文本中的元素与输入图像中的区域，如图2所示。我们的 CAVL 由三个主要组成部分构成：语言预训练、视觉预训练，以及视觉-语言预训练的对比融合。

语言预训练。 对于预训练中的语言嵌入，我们输入三种类型的嵌入：1）句子中每个子词对应的词元嵌入 $\mathbf{e}^t$ ； 2）片段嵌入 $\mathbf{e}^s$ ，用于指示词元来自文本的哪一部分；3）位置嵌入$\mathbf{e}^p$ ，用于表示词元在句子中的位置。然后，我们将这三种嵌入相加，得到上下文表示 $\mathbf{e}_n$， $n \in \{1, 2, ..., N\}$，其中 N 表示句子中的子词数量。经过基于 BERT 的 Transformer 后，这些上下文嵌入变为$\mathbf{e}'_i$。我们采用了与 BERT 类似的两个目标，包括掩码语言建模（MLM）和下一句预测（NSP）。对于前一个目标，我们随机使用特殊词元屏蔽输入词元的某些部分（即， [MASK]），然后模型

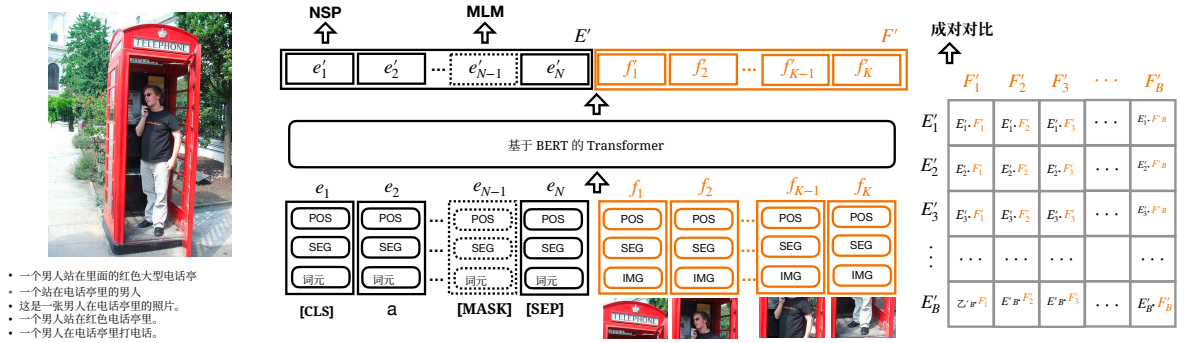


图2. 我们用于成对视觉-语言表征的 CAVL 对比预训练概览。BERT-based Transformer中的自注意力机制用于将输入文本的元素与输入图像中的区域进行隐式对齐，并采用对比学习框架，以显式方式实现整句与每幅图像之间的对齐。

经过训练来预测掩码词元。对于 NSP，我们使用嵌入 [CLS]训练模型，以判断给定的一对句子是否是上下文中的连续句子。

视觉预训练。 对于预训练中的视觉特征，我们从目标检测框架（即，Faster R-CNN）中提取图像区域作为输入 $\mathbf{f}_k$，$k \in \{1, 2, ..., K\}$，其中 K 是图像区域的数量。输入 $\mathbf{f}_k$ 也由三种视觉嵌入组成：1）每个图像区域对应一个图像特征嵌入 $\mathbf{f}^i$ ；2）一个片段嵌入 $\mathbf{f}^s$ ，用于指明图像嵌入与哪个词元嵌入相对应；3）一个位置嵌入 $\mathbf{f}^p$ ，用于表示词元与各图像区域之间的对齐。遵循 Visual-BERT 的任务无关预训练方法，对于 COCO 数据集中的一个文本片段，我们采用两条图像描述；该数据集中一张图像对应多条图像描述，如图 2所示。具体而言，我们使用其中一条图像描述作为描述图像的真实标注，同时以 50% 的概率从这两条图像描述中随机选择一条。我们的模型经过训练，能够区分图像描述是真实标注，还是从两条图像描述中随机抽取的。

视觉语言预训练。 为缓解语言和视觉之间的语义混淆，我们基于多层 Transformer 中的视觉和语言表示设计了一种成对对比学习机制。具体而言，我们计算大小为 B 的批次中每一对语言嵌入$\mathbf{E}'_b$ 和视觉嵌入 $\mathbf{F}'_b$ 之间的余弦相似度，其中 $b \in \{1, 2, ..., B\}$。然后，在同一批次中，我们联合学习这些相似度，以实现整句与每幅图像之间的对齐：最大化批次中 B 正确配对的视觉和语言嵌入的余弦相似度，同时最小化$B^2 - B$ 错误配对的嵌入的余弦相似度。我们对这些相似度分数应用成对对比损失进行优化。

具体而言，我们定义成对对比损失（PwCL），用于语言嵌入$\mathbf{E}'_i$ 与视觉嵌入$\mathbf{F}'_j$ 之间：

$$\mathcal{L}_{\text{PwCL}} = -\log \frac{\sum_{i=1}^{B}(\mathbf{E}'_i \cdot \mathbf{F}'_i)}{\sum_{i=1}^{B}\sum_{j=1}^{B} \mathbb{1}_{i \neq j}(\mathbf{E}'_i \cdot \mathbf{F}'_j)} \tag{1}$$

其中， $\mathbb{1}_{i \neq j}$ 是一个指示函数，用于检查语言嵌入$\mathbf{E}'_i$ 和视觉嵌入$\mathbf{F}'_j$ 是否对齐。通过这种方式，我们最大化正确配对中视觉嵌入和语言嵌入的余弦相似度，同时最小化错误配对中嵌入的余弦相似度。直观地说，在我们的 CAVL中学习整个句子与每张图像之间的对齐，以减轻预训练过程中存在的语义混淆。

值得一提的是，这种对比式预训练方法不同于近期强大的框架 CLIP [34]。具体而言，我们为多模态场景构建了一个图文对齐预训练框架。首先，CLIP 中的这一思路用于训练一个在自然语言监督下图像分类表现更佳的模型，而本文则侧重于解决视觉与语言领域中的多模态问题。其次，他们训练一个图像编码器和一个文本编码器，以生成视觉与语言嵌入。然而，我们将两个编码器合并为一个 BERT 模型，并表明成对对比学习能够使模型学到更好的图像和文本表征。

3.2.2 自适应微调

在本节中，我们设计了一种适配方法，并将其与其他高效方法进行比较，以便在下游任务上微调。一种思路是，模型中的冻结部分为通用问题提供基础信息，而更新部分则生成任务特定特征。受此启发，我们提出了一种方法，在预训练模型旁添加一个捷径模块，并将预训练模型和捷径模块的输出合并。

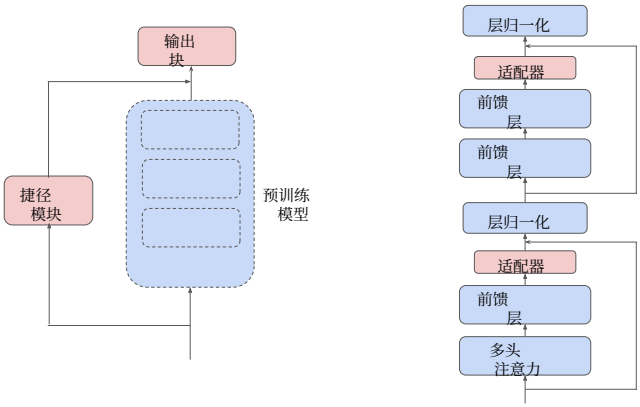


图3。CAVL 中使用适配器进行自适应微调。适配器 I（左）：该适配方法冻结预训练模型，并在微调期间训练捷径模块和输出块。适配器 II（右）：该适配方法遵循 [13]，在 BERT的每个层归一化层正前方添加一个适配器。在微调期间，它冻结前馈层和注意力模块，同时更新适配器和层归一化层中的参数。蓝色块表示冻结部分，红色块中的参数会在微调阶段更新。

增加一个输出块。预训练模型获得图像和文本之间的泛化特征，而捷径模块充当选择神经元，用于捕获各项特定任务中的特征。然后，我们应用输出块将泛化特征与任务特定特征结合起来，以获得最终结果。捷径模块和输出块的层数少于预训练模型，因此缩短了训练时间。我们将这种类型的适配器称为

适配器 I.

受 Neil等人 [13], 的启发，我们提出了另一种适配方法，称为**适配器 II**，即在每个 BERT 层中添加瓶颈结构。具体而言，在注意力模块末尾添加一个适配器，并在前馈层末尾添加另一个适配器。注意力模块和前馈层的输入分别与适配器的输出相加，再通过层归一化模块。在微调过程中，我们冻结注意力模块和前馈层。因此，适配器充当投影模块，将通用特征映射为任务特定特征。该适配器参数较少，采用瓶颈结构设计，包含一个线性层、一个 GELU函数以及另一个线性层。

我们在图3中展示了用于自适应微调的两类适配器的网络细节。我们还在第4.4节对这些适配器进行了全面的消融研究。

4. 实验

4.1. 预训练与实现细节

沿用先前的工作 [24, 27, 40],，我们采用相同的设置，以便与这些基线进行公平比较。具体而言，我们在 MS COCO [26] 和 Visual Genome [20]上对 CAVL 进行预训练。对于视觉标记，我们使用预训练的 Faster R-CNN [37] 提取图像区域（最多 100 个每张图像中检测得分高于 0.5 的感兴趣区域）。我们使用随机梯度下降和 Adam [19] 进行优化，并使用总批量大小为 512，训练 10 个周期。我们将预训练的预热步数设置为总训练集步数的 15%。预训练和微调分别在 4 个 Tesla V100-32G GPU 上耗时 88 小时和 10 小时。

4.2. 下游任务

我们在四个下游任务上评估 CAVL 预训练模型：(1) 视觉问答（VQA），(2) 视觉常识推理（VCR），(3) 视觉推理自然语言（NLVR^2），以及 (4) 区域到短语定位（Flickr30K）。除非另有说明，我们采用适配器 I 作为自适应微调方法。

视觉问答（VQA）。在 VQA 任务中，我们遵循 BUTD 中的实验方案 [1], ，目标是给定一张自然图像，在感知层面回答问题，方法是从包含 3,129 个答案的共享答案集中选出正确答案。具体而言，我们在 VQA v2.0 数据集 [10] 上进行实验，该数据集基于 COCO [26] 数据集中的图像。我们将数据集划分为训练集（83k 张图像和 444k 个问题）、验证集（41k 张图像和 214k 个问题）以及测试集（81k 张图像和 448k 个问题）。结果见表1。与先前方法相比，我们的 CAVL 在开发测试集和标准测试集上的准确率均更高。这表明，视觉表征与语言表征之间的成对对比损失有助于学习整个句子与每张图像之间的对齐。

视觉常识推理（VCR）。在 VCR 任务中，我们需要为给定问题选出正确答案，并提供理由说明，以便对给定图像进行更高层次的认知和常识理解。在实验中，我们使用一张图像以及一组类别化的

模型	VQA 开发测试集	VQA 标准测试集	VCR 问答 →A val	VCR 问答 →A test	VCR 问答 → R val	VCR 问答 → R test	VCR 问答 →AR val	VCR 问答 →AR test
LXMERT [42]	72.42	72.54	-	-	-	-	-	-
ViLBERT [27]	70.55	70.92	72.40	73.30	74.50	74.60	54.00	54.80
VisualBERT [24]	70.08	71.00	70.80	71.60	73.20	73.20	52.20	52.40
VL-BERT [40]	71.72	72.18	73.80	-	74.40	-	55.20	-
UNITER [7]	72.27	72.46	-	75.00	-	77.20	-	58.20
CVLP [38]	72.77	72.90	-	-	-	-	-	-
CAVL（我们的方法）	**72.83**	**73.05**	**75.33**	**75.65**	**76.52**	**77.87**	**58.65**	**59.47**

表 1. 在 VQA 和 VCR数据集上的对比结果。

模型	Dev	测试集-P	测试集-U	测试集-U（共识）
LXMERT [42]	-	74.45	76.20	42.10
VisualBERT [24]	67.40	67.00	67.30	26.90
UNITER [7]	76.93	75.58	-	-
CVLP [38]	-	76.20	-	-
CAVL（我们的方法）	**79.16**	**78.31**	**79.87**	**46.23**

表2. 在 NLVR^2 数据集上的比较结果。

模型	R@1 Dev	R@1 Test	R@5 Dev	R@5 Test	R@10 Dev	R@10 Test
BAN [18]	-	69.70	-	84.20	-	86.40
VisualBERT [24]	70.40	71.33	84.49	84.98	86.31	86.51
MDETR [16]	78.90	-	88.80	-	90.80	-
CAVL（我们的）	**80.35**	**81.76**	**90.63**	**91.12**	**93.16**	**94.21**

表 3. Flickr30K实体数据集上的比较结果。

来自 VCR数据集的感兴趣区域 [50]，分别从 4 个候选答案和 4 个候选理由中选出正确项。该任务（Q → AR）可分为两个子任务：问题回答（Q→A）和答案论证（QA→R）。我们还将 VCR数据集划分为训练集（213k 个问题和 80k 张图像）、验证集（27k 个问题和 10k 张图像）以及测试集（25k 个问题和 10k 张图像）。结果见表1。尽管没有使用 VL-BERT 中规模更大的 Conceptual Captions数据集，我们的 CAVL 仍取得了有竞争力的性能^{large}。这表明，在预训练阶段提出的成对对比损失有助于消除视觉与语言之间的语义混淆。VL-BERT^{large} 也验证了在大规模数据集上进行预训练对于提升模型能力的重要性。

视觉推理自然语言（NLVR）。遵循以往工作 [24]，我们在 NLVR^2 数据集上评估预训练的 CAVL 模型，以对自然语言和图像进行联合推理。在此任务中，我们重点预测自然语言图像描述是否与一对图像相匹配。我们在表2中报告了与最先进方法的比较结果。如图所示，与以往方法相比，我们的 CAVL 在所有实验设置下均取得了新的最先进水平。这证明，在预训练阶段，将成对对比损失同时融入视觉分支和语言分支，能够有效缓解视觉与语言之间现存的语义混淆。结果还表明，我们的 CAVL 在语言与视觉联合推理方面具有优势。

区域到短语定位（RPG）。为了测试我们的 CAVL 在 RPG 上的性能，我们对 CAVL 进行微调在 Flickr30K [47] 实体数据集上预训练，该数据集包含 3 万张图像和 25 万条标注。遵循 BAN [18]和 VisualBERT [24], 的方法，我们使用来自预训练 Faster R-CNN 的图像特征。在 Flickr30K实体数据集上进行微调时，应用一个自注意力模块来生成平均注意力权重，以预测边界框与短语之间的对齐。因此，短语对应的注意力最高的框，就是模型预测的该短语的定位结果。

在该任务中，我们需要选择图像中与某个句子对应的正确边界区域。表 3 报告了与现有方法的对比结果。我们可以观察到，在相同的预训练设置下，我们的 CAVL大幅优于以往的多模态方法，这证明了我们的 CAVL 在视觉语言对齐任务中的有效性。

文本到图像检索（TIR）。遵循先前的工作 [27,28], 我们在 Flickr30K [47]数据集上评估我们的 CAVL在文本到图像检索任务中的表现，并使用召回率（Re-call@1、5、10）作为评估指标。在该任务中，我们需要将一张图像与描述其内容的图像描述进行匹配。具体而言，我们计算每个图像描述与一张图像之间的平均成对相似度（APS）分数。为了选择正确的图像-描述对，我们使用交叉熵损失对模型进行微调。在测试集的推理阶段，我们根据对齐分数对每个图像-描述对进行排序。比较结果如表4所示。可以看出，与先前的基线相比，我们的 CAVL 达到了最佳性能。借助于缓解语义混淆，我们还取得了与 ERNIE-ViL [49]相当的性能。这证明了我们的方法在文本到图像检索中的有效性。



图 4. 图像与文本预训练表示之间余弦相似度的热力图可视化。在预训练过程中，图像与文本对之间的对齐关系得以学习。

模型	TIR			ZS-TIR		
	R@1	R@5	R@10	R@1	R@5	R@10
SCAN [21]	48.60	77.70	85.20	-	-	-
PFAN [45]	50.40	78.70	86.10	-	-	-
CAMP [46]	51.50	77.10	85.30	-	-	-
CVSE [44]	52.90	80.40	87.80	-	-	-
Oscar [25]	54.78	81.05	-	-	-	-
VisualSparta [28]	57.10	82.60	88.20	-	-	-
Vinvl [51]	58.18	83.36	-	-	-	-
VilBERT [27]	58.20	84.90	91.52	31.86	61.12	72.80
SGRAF [9]	58.50	83.00	88.80	-	-	-
Unicoder-VL [23]	71.50	91.20	95.20	48.40	76.00	85.20
UNITER [7]	72.52	92.36	96.08	66.16	88.40	92.94
CAVL（我们的方法）	**74.06**	**92.56**	**96.15**	**67.07**	**88.92**	**93.36**

表 4. Flickr30K 上文本到图像检索（TIR）和零样本文本到图像检索（ZS-TIR）的对比结果。

零样本文本到图像检索（ZS-TIR）。此外，我们遵循先前的工作 [27]，并在 Flickr30K [47] 数据集上评估我们的 CAVL，用于零样本文本到图像检索。在此设置下，我们直接使用预训练权重而不进行微调，以计算零样本检索的平均成对相似度（APS）得分。我们在表4中报告了结果。我们在所有指标上也取得了新的最先进水平，表明了 CAVL 的有效性。

4.3. 可视化

在本部分中，我们将预训练的图像和文本对进行可视化，以验证我们的 CAVL 在缓解这些表示之间的语义混淆方面的有效性。具体而言，我们计算预训练图像与文本对之间的成对相似度，并将结果报告在图 4 中。如图所示，我们的 CAVL 能够在预训练期间学习整句文本与每幅图像之间的对齐关系。

4.4. 消融研究

本节中，我们针对成对对比预训练、自适应微调和批量大小对 CAVL 最终性能的影响，以及所提出的适配器（适配器 I 和 II）的效率，进行了广泛的消融研究。除非另有说明，我们均在 VQA 2.0 数据集上开展消融研究，并报告使用 5 个随机种子得到的所有结果的均值和标准差。

各模块和批量大小的影响。在表 5中，我们探究了 CAVL 中提出的各项预训练任务的影响，这些任务包括掩码语言模型、下一句预测、成对对比损失和自适应微调。可以观察到，引入成对对比损失后，与未采用语言和视觉嵌入之间成对对比损失的基线相比，CAVL 的性能更佳。这证明了 CAVL 所提出的成对对比损失的有效性。采用自适应微调的 CAVL 在取得相当性能的同时，减少了训练参数，并节省了计算资源。

我们还在表5中评估了批量大小对 CAVL 预训练模型最终性能的影响。如表所示，当批量大小为 512 时，我们的 CAVL 在 APS 上表现最佳，这表明在成对对比损失中选择批量大小至关重要。对于批量大小相同的基线，加入 PwCL 后，准确率从 70.11 提升至 71.73，其中在 VQA 任务上的提升（1.62）尤为显著。其改

MLM	NSP	PwCL	AF	批量大小	开发测试集 (↑)	标准测试集 (↑)	APS (↑)
✓	✓			64	70.11±0.12	71.03±0.15	0.43±0.08
✓		✓		64	70.82±0.13	71.56±0.15	0.52±0.06
	✓	✓		64	70.06±0.12	70.95±0.13	0.41±0.06
✓	✓	✓		64	71.73±0.15	72.08±0.17	0.68±0.04
✓	✓	✓	✓	64	71.68±0.14	72.02±0.16	0.67±0.05
✓	✓	✓	✓	128	72.18±0.13	72.32±0.16	0.72±0.04
✓	✓	✓	✓	256	72.42±0.11	72.67±0.13	0.76±0.03
✓	✓	✓	✓	512	72.83±0.05	**73.05±0.07**	**0.83±0.02**
✓	✓	✓	✓	1024	**72.88±0.08**	72.96±0.06	0.81±0.02

表5。关于 p 的消融研究 掩码语言建模，

逐对对比预训练、自适应微调和批量大小。掩码语言模型、下一句预测、PwCL 和自适应微调去噪 te 下一句预测、成对对比损失和自适应微调。

模型	微调参数 (M)↓	开发测试集 ↑	标准测试集 ↑	微调成本（小时）↓
VisualBERT [24]	113.90	70.08	71.00	24.5
VL-BERT$_{base}$ [40]	115.04	71.16	-	25.6
VL-BERT$_{large}$ [40]	342.55	71.79	72.22	70.2
CAVL（适配器 I）	98.40	**72.83±0.05**	**73.05±0.06**	10.0
CAVL（适配器 II）	**46.70**	72.67±0.08	72.86±0.11	**6.1**

表 6. 我们的 CAVL 中提出的适配器变体的消融研究。

进幅度（1.62）在 VQA 任务上尤为显著。将批量大小从 64 增至 1024，可使准确率从 71.68 提升至 72.83（ 1.15），但提升幅度小于 PwCL 带来的提升。这也印证了在视觉-语言预训练中，较大的批量大小能够引入更多负样本对，因而十分重要。表1、2和3中还展示了我们的 CAVL 在 VCR（3.45）、NLVR2（11.31）和 RPG（ 10.43）任务上取得的显著提升。

此外，我们在表格5中比较了预训练期间每个文本-图像对的嵌入之间的平均成对相似度（APS）。具体而言，我们计算语言嵌入 $\mathbf{E}'_i$ 与视觉嵌入 $\mathbf{F}'_i$ 即，$\frac{1}{B}\sum_{i=1}^{B}(\mathbf{E}'_i \cdot \mathbf{F}'_i)$之间的成对点积。我们可以观察到，加入成对对比损失的 CAVL 有助于提高每个文本-图像对之间的 APS，这验证了我们的成对对比预训练在预训练过程中缓解视觉嵌入与语言嵌入之间语义混淆的有效性。

适配器的效果。如表6所示，我们比较了各种模型在训练参数数量、微调成本以及开发测试集和标准测试集准确率方面的表现。我们发现，所有基于 CAVL 的模型都优于基线。这进一步证明了对比式视觉和语言预训练在消除预训练过程中的语义混淆方面的有效性。提出的 CAVL 联合学习同一批次中完整句子与每张图像之间的对齐，从而提升预训练模型的泛化能力。同时，我们采用两种适配器的 CAVL 所需训练参数更少，但性能与微调参数量较大的基线相当，这验证了所提出的适配器在微调较少参数以节省计算资源方面的效率。

我们还将 CAVL 与当前多模态方法中的两类适配器进行比较 [24, 40]，并根据平均微调成本评估 CAVL 在微调阶段能够节省多少时间。通常，CAVL 的两种变体（适配器 I /II）在开发测试集和标准测试集设置下均优于以往工作，同时显著减少了微调所需的参数量和成本。尽管适配器 I 的表现略优于适配器 II，但适配器 II 大幅降低了微调参数量和成本，即，分别降低 59.40% 和 76.17%。这表明 CAVL 在微调阶段能够高效地学习有效表征，用于视觉语言下游任务。

5. Conclusion

在这项工作中，我们提出了一种简单但有效的框架，用于学习视觉与语言的对比式自适应表征，称为 CAVL，该框架包括对比式预训练和自适应微调。在预训练期间，我们采用成对对比损失来减轻语言与视觉之间的语义混淆。此外，我们成功引入了两个轻量级适配器，从而无需在微调阶段耗费大量计算时间。在六项主要的视觉-语言下游任务上，我们的CAVL 相较于基线取得了更优的性能。我们开展了广泛的实验和消融研究，以证明对比式预训练和自适应微调的效率。

参考文献

[1] Peter Anderson、Xiaodong He、Chris Buehler、Damien Teney、Mark Johnson、StephenGould和 Lei Zhang。用于图像描述和视觉问答的自底向上和自顶向下注意力机制。载于《IEEE计算机视觉与模式识别会议论文集》，第6077–6086页，2018年。5[2] Srikar Appalaraju、Bhavan Jasani、Bhargava Urala Kota、YushengXie和R. Manmatha。Docformer：用于文档理解的端到端Transformer。arXiv预印本 arXiv:2106.11539，2021年。2[3] Tom B. Brown、BenjaminMann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakan- tan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、Cle mens Winter、Christopher Hesse、MarkChen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、IlyaSutskever 和DarioAmodei。语言模型是少样本学习器。arXiv 预印本arXiv:2005 14165，2020 2 ..[4] Ting Chen、Simon Kornblith、MohammadNorouzi和Geoffrey Hinton。视觉表征对比学习的简单框架。载于《国际机器学习会议》，第1597–1607页，2020年。2[5] Ting Chen、Simon Kornblith、Kevin Swersky、Mohammad Norouzi 和Geoffrey Hinton。大型自监督模型是强大的半监督学习器。载于《神经信息处理系统进展》，第22243–22255页，2020年。2[6] Xinlei Chen、Haoqi Fan、RossGirshick和KaimingHe。通过动量对比学习改进基线。arXiv预印本 arXiv:2003.04297，2020年。2[7] Yen-Chun Chen、Linjie Li、LichengYu、Ahmed El Kholy、Faisal Ahmed、Zhe Gan、YuCheng 和Jingjing Liu。UNITER：学习通用图文表征。arXiv预印本 arXiv:1909.11740，2019。2、6、7[8] Jacob Devlin、Ming-Wei Chang、KentonLee 和Kristina Toutanova。BERT：面向语言理解的深度双向Transformer 预训练。ar Xiv预印本 arXiv:1810.04805，2018年。2、3[9] Haiwen Diao、YingZhang、LinMa 和 Huchuan Lu。图文匹配的相似性推理与过滤。载于《AAAI人工智能会议》，2021年。7[10] Yash Goyal、Tejas Khot、Douglas Summers-Stay、Dhruv Batra 和Devi Parikh。让vqa 中的 v 发挥作用：提升图像理解在视觉问题回答中的作用。载于《IEEE计算机视觉与模式识别会议论文集》，第6904–6913页，2017年。1、5[11] Kai ming He、Haoqi Fan、Yuxin Wu、SainingXie和Ross Girshick。用于无监督视觉表征学习的动量对比。载于《IEEE计算机视觉与模式识别会议论文集》，第9729–9738页，2020年。2[12] Yicong Hong、Qi Wu、Yuankai Qi、Cristian Rodriguez- Opazo和 Stephen Gould。VLN BERT：一种循环视觉-与语言 bert导航。在 IEEE计算机视觉与模式识别会议论文集，2021年。2[13] Neil Houlsby、Andrei Giurgiu、Stanislaw Jastrzebski、Bruna Morrone、Quentinde Laroussilhe、Andrea Gesmundo、Mona Attariyan 和Sylvain Gelly。面向自然语言处理的参数高效迁移学习。载于国际机器学习会议，第2790–2799页，2019。3、5[14] Yuqi Huo、Manli Zhang、Guangzhen Liu、Haoyu Lu、Yizhao Gao、Guoxing Yang、Jingyuan Wen、Heng Zhang、Baogui Xu、WeihaoZheng、Zongzheng Xi、Yueqian Yang、An-wen Hu、Jinming Zhao、Ruichen Li、Yida Zhao、Liang Zhang、Yuqing Song、Xin Hong、Wanqing Cui、Danyang Hou、Yingyan Li、Junyi Li、Peiyu Liu、Zheng Gong、Chuhao Jin、Yuchong Sun、Shizhe Chen、Zhiwu Lu、Zhicheng Dou、Qin Jin、Yanyan Lan、Wayne XinZhao、RuihuaSong 和 Ji-Rong Wen。WenLan：通过大规模多模态预训练连接视觉与语言。arXiv预印本 arXiv:2103.06561，2021年。2[15] Chao Jia、Yinfei Yang、YeXia、Yi-Ting Chen、Zarana Parekh、HieuPham、Quoc V. Le、Yun-Hsuan Sung、Zhen Li和 Tom Duerig。利用噪声文本监督扩展视觉与视觉语言表征学习。载于国际机器学习会议，2021年。2[16] AishwaryaKamath、Mannat Singh、Yann LeCun、Ishan Misra、Gabriel Synnaeve和 Nicolas Carion。MDETR——面向端到端多模态理解的调制检测。arXiv预印本 arXiv:2104.12763，2021年。6[17] PrannayKhosla、PiotrTeterwak、Chen Wang、Aaron Sarna、Yonglong Tian、Phillip Isola、Aaron Maschinot、CeLiu 和 Dilip Krishnan。监督式对比学习。载于神经信息处理系统进展，第18661–18673页，2020。2[18] Jin-Hwa Kim、JaehyunJun和Byoung-Tak Zhang。双线性注意力网络。载于神经信息处理系统进展，第1571–1581页，2018年。6[19] Diederik PKing ma和Jimmy Ba。Adam：一种随机优化方法。arXiv 预印本 arXiv:1412 6980，2014 5。.[20] Ranjay Krishna、YukeZhu、OliverGroth、Justin Johnson、Kenji Hata、Joshua Kravitz、Stephanie Chen、Yannis Kalan-tidis、Li-Jia Li、David A. Shamma、Michael S. Bernstein和Fei-FeiLi。Visual Genome：利用众包密集图像标注连接语言与视觉。国际计算机视觉杂志，123：32–73，2017年。5[21] Kuang-Huei Lee、Xi Chen、Gang Hua、Houdong Hu和Xi- aodongHe。用于图文匹配的堆叠交叉注意力。第201–216页，2018年。7[22] Jie Lei、LinjieLi、Luowei Zhou、Zhe Gan、Tamara L. Berg、Mohit Bansal和Jingjing Liu。Less isMore：通过稀疏采样实现 ClipBERT 视频与语言学习。载于IEEE/CVF计算机视觉与模式识别会议论文集，2021年。2[23] Gen Li、Nan Duan、Yuejian Fang、Ming Gong和 Daxin Jiang。Unicoder-VL：通过跨模态预训练实现视觉与语言的通用编码器。载于第三十四届AAAI人工智能会议，第11336–11344页，2020。3、7

[24] Liunian Harold Li、Mark Yatskar、Da Yin、Cho-Jui Hsieh 和Kai-WeiChang。VisualBERT：一种用于视觉与语言的简单且高效的基线模型。arXiv预印本arXiv:1908.03557，2019。1、2、5、6、8[25] Xiujun Li、XiYin、Chunyuan Li、Pengchuan Zhang、Xiaowei Hu、Lei Zhang、Lijuan Wang、Houdong Hu、Li Dong、Furu Wei、Yejin Choi 和 Jianfeng Gao。Oscar：面向视觉语言任务的对象语义对齐预训练。在《欧洲计算机视觉会议》，2020 年。2、7[26] Tsung-Yi Lin、Michael Maire、Serge J. Belongie、Lubomir D. Bourdev、Ross B. Girshick、JamesHays、PietroPerona、Deva Ramanan、Piotr Doll´ar 和 C.LawrenceZitnick。Microsoft COCO：上下文中的常见物体。arXiv 预印本 arXiv:1405.0312，2014。5[27] Jiasen Lu、Dhruv Batra、Devi Parikh 和StefanLee。VilBERT：面向视觉与语言任务的任务无关的视觉语言表征预训练。在《神经信息处理系统进展》，第 13–23 页，2019 年。1、2、5、6、7[28] XiaopengLu、TianchengZhao 和 KyusongLee。VisualSparta：用于大规模文本到图像检索的稀疏 Transformer片段级匹配。在《计算语言学协会第 58 届年会论文集》，2021 年。6、7[29] Shentong Mo、Zhun Sun和 Chao Li。孪生原型对比学习。在《英国机器视觉会议论文集》，2021 年。2[30] Shentong Mo、ZhunSun 和 Chao Li。通过对齐、均匀性和相关性重新思考原型对比学习。在《英国机器视觉会议论文集》，2022 2。Shentong Mo、Zhun Sun和Chao Li。[31]用于自监督视觉 Transformer 的多层级对比学习。在《2023 IEEE/CVF 冬季计算机视觉应用会议》（WACV），第 2777–2786 页，2023 年。2[32] Jonas Pfeiffer、Ivan Vuli´c、Iryna Gurevych 和SebastianRuder。MAD-X：用于多任务跨语言迁移的基于适配器的框架，2020 年。3[33] Di Qi、Lin Su、Jia Song、Edward Cui、TaroonBharti 和 Arun Sacheti。ImageBERT：利用大规模弱监督图文数据进行跨模态预训练。arXiv预印本 arXiv:2001.07966，2020。2[34] Alec Radford、Jong WookKim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。从自然语言监督中学习可迁移视觉模型。arXiv 预印本 arXiv:2103.00020，2021。2、4[35] Alec Radford、Karthik Narasimhan、TimSalimans和 IlyaSutskever。通过生成式预训练改进语言理解。2019 年。2[36] Alec Radford、Jeff Wu、Rewon Child、David Luan、Dario Amodei和 Ilya Sutskever。语言模型是无监督的多任务学习器。2019年。2[37] Shaoqing Ren、Kaiming He、Ross Girshick 和 Jian Sun。Faster R-CNN：利用区域提议网络实现实时目标检测。在《神经信息处理系统进展》，第 91–99 页，2015 年。5

[38] Lei Shi、Kai Shuang、Shijie Geng、Peng Su、Zhengkai Jiang、Peng Gao、Zuohui Fu、Gerard de Melo 和 Sen Su。对比式视觉-语言预训练。arXiv 预印本 arXiv:2007.13135，2020。2，6[39] Dandan Song、Siyi Ma、Zhanchen Sun、Sicheng Yang 和 Lejian Liao。KVL-BERT：用于视觉常识推理的知识增强型视觉与语言 BERT。arXiv 预印本 arXiv:2012.07000，2020。2[40] Weijie Su、Xizhou Zhu、Yue Cao、Bin Li、Lewei Lu、Furu Wei 和 Jifeng Dai。VL-BERT：通用视觉-语言表征的预训练。发表于国际学习表征会议，2020。1，2，5，6，8[41] Chen Sun、Austin Myers、Carl Vondrick、Kevin Murphy 和 Cordelia Schmid。VideoBERT：用于视频与语言表征学习的联合模型。arXiv 预印本 arXiv:1904.01766，2019。2[42] Hao Tan 和 Mohit Bansal。LXMERT：从 Transformer 中学习跨模态编码器表征。arXiv 预印本 arXiv:1908.07490，2019。2，6[43] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。Attention Is All You Need。发表于神经信息处理系统进展，第 5998–6008 页，2017年。2[44] Haoran Wang、Ying Zhang、Zhong Ji、Yanwei Pang 和 Lin Ma。用于图文匹配的共识感知视觉语义嵌入。发表于欧洲计算机视觉会议，第 18–34 页，2020。7[45] Yaxiong Wang、Hao Yang、Xueming Qian、Lin Ma、Jing Lu、Biao Li 和 Xin Fan。用于图文匹配的位置聚焦注意力网络。发表于第 28 届国际人工智能联合会议论文集，第 3792–3798 页，2019。7[46] Zihao Wang、Xihui Liu、Hongsheng Li、Lu Sheng、Junjie Yan、Xiaogang Wang 和 Jing Shao。CAMP：用于文本图像检索的跨模态自适应消息传递。发表于IEEE国际计算机视觉会议论文集，第 5764–5773 页，2019。7[47] Peter Young、Alice Lai、Micah Hodosh 和 Julia Hockenmaier。从图像描述到视觉指称：用于事件描述语义推断的新相似度指标。计算语言学协会会刊，2：67–78，2014。6，7[48] Fei Yu、Jiji Tang、Weichong Yin、Yu Sun、Hao Tian、Hua Wu 和 Haifeng Wang。ERNIE-ViL：通过场景图实现知识增强的视觉-语言表征。arXiv 预印本 arXiv:2006.16934，2020。2[49] Fei Yu、Jiji Tang、Weichong Yin、Yu Sun、Hao Tian、Hua Wu 和 Haifeng Wang。ERNIE-ViL：通过场景图实现知识增强的视觉-语言表征。发表于第三十四届 AAAI 人工智能会议，2021年。2，6[50] Rowan Zellers、Yonatan Bisk、Ali Farhadi 和 Yejin Choi。从识别到认知：视觉常识推理。发表于IEEE计算机视觉与模式识别会议论文集，第 6720–6731 页，2019。1，6[51] Pengchuan Zhang、Xiujun Li、Xiaowei Hu、Jianwei Yang、Lei Zhang、Lijuan Wang、Yejin Choi 和 Jianfeng Gao。

Vinvl：重新审视视觉-语言模型中的视觉表征。在 *IEEE/CVF*计算机视觉与模式识别会议论文集，第5579–5588页，2021年。 2，7[52] Yuhao Zhang、Hang Jiang、Yasuhide Miura、Christopher D Manning 和 Curtis Langlotz。从配对图像和文本中对医学视觉表征进行对比学习。 *arXiv*预印本 *arXiv:2010.00747*，2020年。 3