CLIP4IDC：用于图像差异字幕的 CLIP

Zixin Guo、Tzu-Jui Julius Wang、Jorma Laaksonen 芬兰阿尔托大学计算机科学系
{zixin.guo, tzu-jui.wang, jorma.laaksonen}@aalto.fi

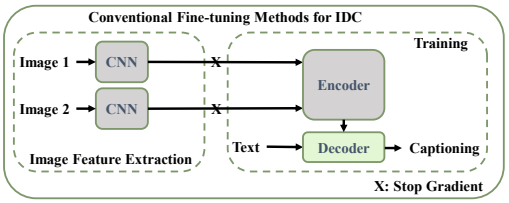
抽象的

图像差异字幕 (IDC) 旨在生成句子来描述两张相似的图像之间的差异。传统方法使用预先训练且通常冻结的视觉特征提取器来学习 IDC 模型。因此，可能出现两个主要问题：（1）用于训练这种视觉编码器的预训练数据集与下游 IDC 任务的数据集之间通常存在较大的领域差距；（2）视觉特征提取器在分别编码两幅图像时，通常不能有效地编码两幅图像之间的视觉变化。由于最近提出的 CLIP 具有出色的零样本性能，因此我们提出 CLIP4IDC 来迁移 CLIP 模型以用于 IDC 任务来解决这些问题。与直接微调 CLIP 来生成句子不同，我们引入了一个自适应训练过程来适应 CLIP 的视觉编码器，以基于文本描述捕获和对齐图像对中的差异。在三个 IDC 基准数据集 CLEVR-Change、Spot-the-Diff 和 Image-Editing-Request 上进行的实验证明了 CLIP4IDC 的有效性。

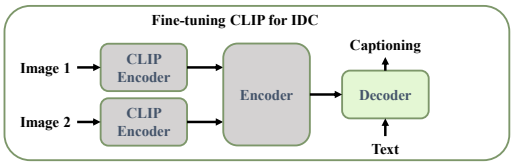
1 简介

理解和表达视觉内容的任务对机器来说很难，因为建立视觉和文本域之间的关系模型需要复杂的计算推理。作为其中一项任务，图像字幕 (IC) (Vinyals 等人，2015；Xu 等人，2015) 旨在根据给定的图像生成连贯的描述。图像差异字幕 (IDC) (Jhamtani 和 Berg-Kirkpatrick，2018；Park 等人，2019) 是从图像字幕扩展而来，描述了两幅相似图像中出现的细微变化。这更具挑战性，因为机器需要识别这对图像中的视觉对象和细微差别。

图 1a 显示了 IDC 的传统方法。首先，图像对的视觉特征是



(a) The fine-tuning strategy with a frozen (CNN) feature extractor.



(b) CLIP's fine-tuning strategy with an image encoder also fine-tuned.

图 1：不同的传统微调（FT）策略可能会导致任务准确性较差，原因是：（1）没有考虑到预训练（PT）和 FT 中的不同目标所引入的差距，以及（2）PT 和 FT 中使用的数据集的领域转变。

使用预训练模型离线提取（He et al., 2016; Ren et al., 2015）。然后字幕网络生成句子来描述对中的变化。尽管这种方法取得了很大进展（Park et al., 2019; Kim et al., 2021; Huang et al., 2021; Hosseinzadeh and Wang, 2021; Sun et al., 2022），但它们的缺点是视觉特征没有考虑到预训练和 IDC 任务之间的领域差距。Lei et al. (2021) 证明，在原始任务上训练的特征提取器的目的与后续任务的目的存在差距。例如，在图像分类任务上训练的模型提取的特征侧重于高级上下文，而丢失了 IDC 所需的细粒度信息。此外，提取的单一模态的视觉表征与文本表征不相关。

作为解决这些缺点的有效方法，在目标数据集上微调模型可以缩小任务之间的差距。姚等人（2022）

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 33–42
November 20–23, 2022.

表明，在相同的离线提取特征上进行预训练和微调的 Transformer (Vaswani et al., 2017) 模型在 IDC 中取得了最先进的结果。然而，它还没有充分利用来自大规模数据集的知识，就像视觉语言 (VL) 预训练 (Zhou et al., 2020; Li et al., 2021) (VLP) 方面的最新进展一样。特别是，对比 VLP 模型 CLIP (Radford et al., 2021) 已在众多 VL 下游任务中展示了其零样本优势 (Luo et al., 2021; Tang et al., 2021)。

我们开始在 IDC 任务上试验一种典型的 CLIP 微调策略，如图 1b 所示，其中 CLIP 的视觉编码器是在原始像素上学习和微调的。然而，不仅 CLIP 预训练和 IDC 的目标之间仍然存在差距，而且预训练收集的图像文本对和 IDC 中的图像差异对之间也存在差距。这些差距限制了模型适应 IDC 任务的能力。

为了解决这些问题，我们研究如何有效地将预先训练好的 CLIP 迁移到 IDC。图 2 显示了所提出的 CLIP4IDC 模型的概览。与直接微调用于 IDC 任务的 CLIP 相比，CLIP4IDC 采用“ *adapt*-and-*fine-tune* ”策略。对于 *adapt*，CLIP 编码器学习捕捉图像对中的细粒度差异，而不是分别为这两幅图像生成高级语义信息。在此阶段，图像对和句子的视觉和文本表示被学习以与检索损失对齐。对于 *fine-tune*，学习过的视觉编码器后面跟着一个从头开始训练的字幕 Transformer。

在合成和真实基准数据集 CLEVR-Change (Park et al., 2019) 和 Spot-the-Diff (Jham-tani and Berg-Kirkpatrick, 2018) 上分别进行了广泛的实验。此外，还报告了在混合真实合成数据集 Image-Editing-Request (Tan et al., 2019) 上的结果。CLIP4IDC 在这三个数据集的所有指标上都优于强基线。这项工作的主要贡献是：

1) 与在预提取特征上进行训练的传统方法相比，我们在原始像素上对 CLIP 进行微调以实现 IDC。这既保留了预训练特征的表现力，又使其适应新的任务领域。

2）我们提出了 CLIP4IDC，它由适应和微调阶段组成，以缩小差距在预训练 CLIP 期间，以及针对 IDC 对其进行微调时，在目标和数据域之间进行调整。通过相互检索视觉差异和描述来学习适应性。

3）大量实验表明，CLIP4IDC 在三个数据集的 IDC 任务中所有指标均优于多个强基线。[1]

2 CLIP4IDC

如图 1a 所示，标准 IDC 方法基于预提取的特征生成句子。瓶颈在于三个方面：1) 特征提取中的梯度流停止，2) 预训练和 IDC 微调之间的目标和数据域不匹配，3) 视觉特征是“ 纯视觉的 ”，即它们位于视觉域，远离文本域。在以下部分中，我们将介绍 CLIP4IDC，这是一种基于 CLIP 的方法，用于解决这些瓶颈。

2.1 CLIP 微调方法

图 1b 显示了对 IDC 的 CLIP 进行微调的端到端方法。具体来说，图像表示由使用 CLIP 初始化的视觉编码器生成（Dosovitskiy 等人，2020 年），并输入到 Transformer 编码器中，以专注于解释图像对中的差异。使用 Transformer 解码器来描述给定视觉上下文的变化。

2.2 模型架构

图 2 概述了 CLIP4IDC 模型，其中包含视觉和语言编码器。

语言编码器。给定文本标题 T，使用由 N_G 个 Transformer 层组成的语言编码器 G，表示为

$$G(T) = G(\{E_{\text{bos}}, E_{t_1}, ..., E_{t_m}, E_{\text{eos}}\} + p_{\text{T}}), \quad (1)$$

其中 $E_* \in \mathbb{R}^{d_T}$ 是每个标记的线性投影，$p_T \in \mathbb{R}^{(m+2)\times d_T}$ 是学习到的位置嵌入，用于保留位置信息。E_{bos} 和 E_{eos} 是标记嵌入，分别表示文本的开始和结束。语言编码器的输出 $g \in \mathbb{R}^{d_T}$ 是通过收集标记嵌入 E_{eos} 的输出生成的。

视觉编码器。图像对 (X^1, X^2) 中的每个图像都使用 CLIP 的初始卷积层进行拼接，拼接成 n 个图像块，尺寸为

[1] https://github.com/sushizixin/CLIP4I DC

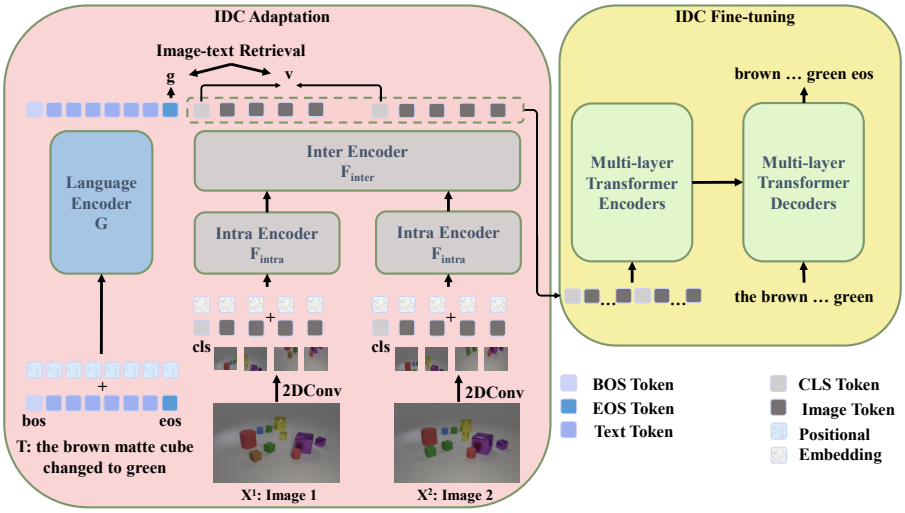


图2：CLIP4IDC 的详细架构。

情态 d_I 为：

$$X^1 = \{x_{\text{cls}}, x_1^1, ..., x_n^1\} + p_{\text{I}}, \quad (2)$$

$$X^2 = \{x_{\text{cls}}, x_1^2, ..., x_n^2\} + p_{\text{I}}, \quad (3)$$

其中 x_{cls} 是学习到的类嵌入，用于表示图像的全局上下文，$p_{\text{I}} \in \mathbb{R}^{(n+1)\times d_I}$ 是位置嵌入。$\{\cdots\}$ 是嵌入的序列。视觉编码器 F 被构建来捕捉图像对中的细微变化。F 由 CLIP 的权重初始化，由 *intra* 和 *inter* 个 Transformer 模块组成。具体来说，包含 N_{intra} 个 Transformer 层的 *intra* 模块 F_{intra} 从图像对中学习单模态上下文。带有 N_{inter} 层的 *inter* 模块 F_{inter} 被构建来关注每对上下文之间的细微差异。这些程序被表述为：

$$F(X^1, X^2) = F_{inter}(\{F_{intra}(X^1) + e_1, \quad (4)$$
$$F_{intra}(X^2) + e_2\} + p),$$

其中 $p \in \mathbb{R}^{2(n+1)\times d_I}$. e_1 和 $e_2 \in \mathbb{R}^{d_I}$ 是用于表示第一和第二幅图像的特殊标记嵌入。之后，将可学习的线性投影 $W \in \mathbb{R}^{d_I \times d_T}$ 应用于视觉表示 $F(X^1, X^2)$，在此基础上生成最终的视觉表示 $F'(X^1, X^2)$。

2.3 IDC 专属适配

接下来，我们提出了两个新颖的 IDC 特定借口任务，即图像对到文本（IP-T）和文本到图像对（T-IP）检索，以便更好地调整视觉表示以进行字幕制作。

在针对实际 IDC 任务对 CLIP 进行微调之前，我们通过 IP-T 和 T-IP 检索将视觉特征调整到 IDC 任务的领域。我们的调整方法遵循对比方法，其中编码图像对更接近编码差异字幕。尽管存在其他类型的适应策略，例如更注重匹配域分布的策略（Tzeng 等人，2014），但我们只关注证明添加这样的适应步骤是否有用。我们从图像对的 x_{cls} 嵌入中聚合出一个组合视觉表示 $v \in \mathbb{R}^{d_T}$，表示为：

$$v = f(\{F'(X^1, X^2)_1, F'(X^1, X^2)_{n+2}\}), \quad (5)$$

其中 f 是均值池化操作。下标是嵌入在表示中的位置（1 索引）。给定一批 B 个图像对和差异标题，目标是匹配图像对的差异表示与差异描述之间的 $B \times B$ 个相似性。损失函数定义为：

$$\mathcal{L}_{i2t} = \frac{-1}{B}\sum_{i}^{B} \log \frac{\exp(\text{s}(v_i, g_i)/\tau)}{\sum_{j=1}^{B}\exp(\text{s}(v_i, g_j)/\tau)}, \quad (6)$$

$$\mathcal{L}_{t2i} = \frac{-1}{B}\sum_{i}^{B} \log \frac{\exp(\text{s}(v_i, g_i)/\tau))}{\sum_{j=1}^{B}\exp(\text{s}(v_j, g_i)/\tau)}, \quad (7)$$

$$\mathcal{L} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i}, \quad (8)$$

其中 $\mathcal{L}_{i2t}$ 和 $\mathcal{L}_{t2i}$ 分别是 IP- T 和 T-IP 检索的损失函数。s$(\cdot,\cdot)$ 表示余弦相似度函数，τ 是可学习的温度参数，用于平滑梯度。

2.4 字幕

在实际的字幕生成阶段，视觉编码器使用从上一个自适应阶段获得的权重进行初始化，并将视觉编码器的输出 $F'(X^1, X^2)$ 输入到字幕生成模型中。如图 2 所示，字幕生成模型包含多层 Transformer 编码器和解码器。

Model	Input	PT	B	M	C	R
Capt-Dual-Att (2019)	ResNet	–	43.5	32.7	108.5	–
DUDA (2019)	ResNet	–	47.3	33.9	112.0	–
VAM (2020)	ResNet	–	50.3	37.0	114.9	69.7
VAM+ (2020)	ResNet	–	51.3	37.8	115.8	70.4
IFDC (2021)	F-RCNN	–	49.2	32.5	118.7	69.1
DUDA+Aux (2021)	ResNet	–	51.2	37.7	115.4	70.5
VACC (2021)	ResNet	–	52.4	37.5	114.2	–
BiDiff (2022)	ResNet	–	54.2	38.3	118.1	–
IDC-PCL (2022)	ResNet	✓	51.2	36.2	128.9	71.7
CLIP4IDC	Raw	✓	**56.9**	**38.4**	**150.7**	**76.4**
CC-Full (2022)	Raw,ResNet	✓	64.3	36.4	151.4	77.1

表 1：IDC 对 CLEVR-Change 测试拆分的结果。突出显示了主要指标 CIDer。CC-Full 属于单独的组，因为它采用直接针对目标指标进行优化的策略梯度方法。

编码器分别用于视觉和文本表示。解码器经过训练，可以根据前一个基本事实词和视觉差异预测下一个标记。使用 Park 等人 (2019) 中的词级交叉熵 (XE) 损失。

3 实验

3.1 基准数据集和指标

我们在 CLEVR-Change (Park et al., 2019)、Spot-the-Diff (Jhamtani and Berg-Kirkpatrick, 2018) 和 Image-Editing-Request (Tan et al., 2019) 数据集上进行了实验。根据之前的研究，例如 (Huang et al., 2021; Hos-seinzadeh and Wang, 2021)，我们在 BLEU (B) (Papineni et al., 2002)、METEOR (M) (Banerjee and Lavie, 2005)、CIDEr-D (C) (Vedantam et al., 2015) 和 ROUGE-L (R) (Lin, 2004) 上对 *test* 分割的字幕模型进行了评估。IDC 适配通过图像对到文本 (IP-T) 和文本到图像对 (T-IP) 检索任务完成。报告的标准检索指标为：K 级召回率 (R@K)、中位数排名 (MdR) 和平均排名 (MnR)。

3.2 字幕结果

我们将 CLIP4IDC 与直接 CLIP 微调方法和采用表 1-4 中预提取特征的现有技术进行了比较。

CLEVR-Change 上的结果。表 1 显示，CLIP4IDC 在 CIDEr 上的表现优于除 CC-Full（Ak 等人，2022）之外的所有基线。请注意，CC-Full 采用策略梯度方法并直接针对生成目标字幕进行了优化，而我们提出的 CLIP4IDC 仅依赖于标准 XE 字幕损失。因此，我们认为它们的结果不具可比性，但是，我们的结果仍然相当有竞争力。正如我们将在后面的部分中看到的那样，CLIP4IDC 在真实世界数据集上的表现明显优于 CC-Full。

Model	C	T	M	A	D	DI
DUDA (2019)	120.4	86.7	56.4	108.2	103.4	110.8
VAM+ (2020)	122.1	98.7	82.0	126.3	115.8	122.6
IFDC (2021)	133.2	99.1	82.1	128.2	118.5	114.2
DUDA+Aux (2021)	120.8	89.9	62.1	119.8	123.4	116.3
BiDiff (2022)	115.9	106.8	71.8	121.3	124.9	116.1
IDC-PCL (2022)	131.2	101.1	81.7	**133.3**	116.5	**145.0**
CLIP4IDC	**149.1**	**135.3**	**91.0**	132.4	**135.5**	133.4

表 2：CLEVR-Change 测试中不同类型变化的 CIDEr 分数细分。C、T、M、A、D、DI 列代表颜色、纹理、移动、添加、删除和干扰项的变化类型，即图像对中没有变化。

Model	Input	PT	B	M	C	R
DDLA (2018)	ResNet	–	8.5	12.0	32.8	28.6
DUDA (2019)	ResNet	–	8.1	11.5	34.0	28.3
VAM (2020)	ResNet	–	10.1	12.4	38.1	31.3
IFDC (2021)	F-RCNN	–	8.7	11.7	37.0	30.2
DUDA+Aux (2021)	ResNet	–	8.1	12.5	34.5	29.9
VACC (2021)	ResNet	–	9.7	12.6	41.5	32.1
CLIP4IDC	Raw	✓	**11.6**	**14.2**	**47.4**	**35.0**
CC-Full (2022)	Raw,ResNet	✓	8.3	13.0	33.0	30.0

表 3：IDC 对 Spot-the-Diff 测试分割的结果。

Model	Input	PT	B	M	C	R
Rel-Att (2019)	ResNet	–	6.7	12.8	26.4	37.4
DUDA (2019)	ResNet	–	6.5	12.4	22.8	37.3
BiDiff (2022)	ResNet	–	6.9	**14.6**	27.7	38.5
CLIP4IDC	Raw	✓	**8.2**	**14.6**	**32.2**	**40.4**

表 4：图像编辑请求测试分割的结果。

			CLEVR-Change				Spot-the-Diff			
Model	$\mathcal{L}$	Params	B	M	C	R	B	M	C	R
CLIP-FT	–	135.57M	49.9	34.8	133.9	70.8	11.0	12.8	43.3	33.5
CLIP4IDC	–	135.65M	54.2	37.9	147.5	75.4	11.0	12.9	43.0	33.4
CLIP4IDC	✓	135.65M	**56.9**	**38.4**	**150.7**	**76.4**	**11.6**	**14.2**	**47.4**	**35.0**

表5：IDC在两个数据集上的消融结果。

我们还通过 CLEVR-Change 上的不同类型的变化对模型进行了评估，如表 2 所示。CLIP4IDC 在颜色、纹理、移动和拖放类型上的表现优于 IDC-PCL。

Spot-the-Diff 和 Image-Editing-Request 的结果。表 3 和表 4 显示，CLIP4IDC 在两个真实数据集的所有指标上都实现了比基线更高的准确度。

消融。我们对不同的 CLIP 架构和适应策略进行了消融研究。表 5 显示，没有适应阶段的 CLIP4IDC（没有公式 8 中的 $\mathcal{L}$）在 CLEVR-Change 上的表现优于直接 CLIP 微调（“CLIP-FT”）。在更具挑战性的真实世界数据集 Spot-the-Diff 上，我们观察到了相同的趋势。因此，具有 $\mathcal{L}$ 的适应阶段进一步提高了性能。这证实了在适应阶段学习捕捉更细粒度的视觉差异是有益的。

	CLEVR-Change						Spot-the-Diff						Editing-Request					
	Image Pair ⇔ Text			Text ⇔ Image Pair			Image Pair ⇔ Text			Text ⇔ Image Pair			Image Pair ⇔ Text			Text ⇔ Image Pair		
Model	R@1	R@5	R@10	R@1	R@5	R@10	R@10	R@20	R@50	R@10	R@20	R@50	R@1	R@5	R@10	R@1	R@5	R@10
CLIP4IDC	46.4	83.0	86.6	26.8	58.7	70.0	3.7	7.3	16.8	6.2	10.5	20.0	17.1	28.4	33.8	17.3	33.7	41.9

Table 6: Results of IP-T and T-IP retrieval on the three datasets.

			Image Pair ⇒ Text					Text ⇒ Image Pair					Captioning			
Model	N_{intra}	N_{inter}	R@1	R@5	R@10	MdR↓	MnR↓	R@1	R@5	R@10	MdR↓	MnR↓	B	M	C	R
	6	6	46.1	79.8	83.9	2.0	49.6	26.4	57.1	68.4	4.0	29.4	54.0	37.4	146.5	75.2
	7	5	46.1	80.8	84.5	2.0	45.5	27.0	57.8	69.0	4.0	28.2	54.5	37.5	148.4	75.5
	8	4	**47.2**	80.7	84.4	2.0	46.3	**27.7**	**58.7**	69.7	4.0	29.9	54.1	37.4	147.3	75.4
CLIP4IDC	9	3	46.4	**83.0**	**86.6**	**2.0**	**39.2**	26.8	58.6	**70.0**	**4.0**	**25.6**	**54.8**	**37.8**	**148.6**	**75.8**
	10	2	37.5	68.5	73.9	2.0	88.8	22.9	52.3	63.9	5.0	54.4	51.5	35.4	134.6	71.5
	11	1	24.7	47.2	53.3	7.0	143.6	17.8	40.2	50.9	10.0	84.8	45.0	32.7	122.8	67.9
	12	0	2.3	7.0	11.8	182.0	459.9	1.1	3.9	5.9	419.0	716.5	38.8	29.5	90.9	60.6

表 7：在 CLEVR-Change 测试分割的 IP-T、T-IP 检索和 IDC 任务中，在 CLIP4IDC 中设置不同层数的结果。

3.3 适配结果

我们在表 6 中报告了用于适应的检索任务中三个数据集的测试分割的结果。这些来自图像对和文本检索任务的结果只是为了证明该模型能够捕捉图像对中的细节。下面评估了检索任务对字幕准确性的影响。

4 IDC适配评估

我们研究了在 CLEVR-Change 测试集上，CLIP4IDC 中不同的架构选项对检索准确率的影响。表 7 展示了在 *intra* 和 *inter* 模块中设置不同层数的影响。可以看出，改进是通过为 *intra* 模块分配大量层来实现的。但这并不意味着不需要 *inter* 层，正如在减少中间层数时准确率下降所显示的那样。此外，当删除 *inter* 层，即 $N_{inter} =$ 为 0 时，该架构与 Luo et al. (2021) 的架构类似，其准确率大大降低。这归因于两个单独的图像嵌入所表示的全局信息无法定位它们之间的变化。

为了进一步研究基于检索的适应性与字幕制作准确率之间的关系，我们使用冻结图像编码器从适应阶段开始对字幕制作任务的模型进行微调。从表 7 可以看出，一般来说，检索任务的适应性越好，召回率越高，字幕制作效果就越好。观察结果表明，引入的检索任务和用于检索的指标是 IDC 性能的有力指标。

5 结论和未来工作

在本研究中，我们研究了如何微调 CLIP 以进行图像差异字幕制作。引入基于检索的自适应来改善字幕制作的视觉表示，并缩小 CLIP 预训练和 IDC 的目的和数据域之间的差距。实验结果证明了 CLIP4IDC 模型和应用域自适应的有效性。

在未来的工作中，我们将进一步探索增强视觉和语言领域之间的关系。具体来说，CLIP4IDC 采用了 CLIP，它不像其他预先训练的 VL 模型（Lu et al., 2019; Su et al., 2019; Li et al., 2019）那样早地涉及跨模态交互，这些模型允许从头开始进行交互。将其他 VL 模型改编为 IDC 自然是一个有趣的未来方向。此外，探索除我们的对比方法之外的其他方法（例如域混淆（Tzeng et al., 2014））来连接视觉和语言领域是另一个可行的方向。

致谢

这项工作得到了芬兰科学院 317388、329268 和 345791 项目的支持。我们也感谢阿尔托科学 IT 项目和 CSC - 芬兰科学 IT 中心提供的计算资源。

参考

Kenan Emir Ak、Ying Sun 和 Joo Hwee Lim。2022 年。通过想象学习：基于文本的图像处理和更改字幕的联合框架。*IEEE Transactions on Multimedia*。

Satanjeev Banerjee 和 Alon Lavie。2005 年。Meteor：一种自动 mt 评估指标，与人类判断的相关性得到改善。在 *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*，第 65-72 页。

Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、翟晓华、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。 2020 年。一张图像相当于 16x16 个单词：用于大规模图像识别的 Transformers。在*International Conference on Learning Representations*中。

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recog- nition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770 – 778.

Mehrdad Hosseinzadeh 和 Yang Wang。2021 年。通过从辅助任务中学习来更改图像字幕。在 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第 2725-2734 页。

Qingbao Huang, Yu Liang, Jielong Wei, Cai Yi, Hanyu Liang, Ho-fung Leung, and Qing Li. 2021. Image difference captioning with instance-level fi ne-grained feature representation. *IEEE Transactions on Multimedia*.

Harsh Jhamtani 和 Taylor Berg-Kirkpatrick。2018 年。学习描述相似图像对之间的差异。在 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*，第 4024 – 4034 页。

Hoeseong Kim、Jongseok Kim、Hyungseok Lee、Hyun- sung Park 和 Gunhee Kim。2021 年。具有周期一致性的不可知变化字幕。在 *Proceedings of the IEEE/CVF International Conference on Computer Vision*，第 2095-2104 页。

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331 – 7341.

Junnan Li、Ramprasaath Selvaraju、Akhilesh Gotmare、Shafiq Joty、Caiming Xiong 和 Steven Chu Hong Hoi。2021 年。融合前对齐：基于动量蒸馏的视觉和语言表征学习。*Advances in neural information processing systems*，34：9694 – 9705。

Liunian Harold Li、Mark Yatskar、Da Yin、Cho-Jui Hsieh 和 Kai-Wei Chang。2019 年。Visualbert：用于视觉和语言的简单且高效的基线。*arXiv preprint arXiv:1908.03557*。

Chin-Yew Lin。2004 年。ROUGE：自动评估摘要的软件包。在 *Text Summariza- tion Branches Out*，第 74-81 页，西班牙巴塞罗那。计算语言学协会。

Jiasen Lu、Dhruv Batra、Devi Parikh 和 Stefan Lee。2019 年。Vilbert：针对视觉和语言任务的预训练任务无关的视觉语言学表征。*Advances in neural information processing systems*，32。

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*.

Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。2002 年。Bleu：一种自动评估机器翻译的方法。在 *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*，第 311-318 页。

Dong Huk Park、Trevor Darrell 和 Anna Rohrbach。2019 年。强大的变更字幕。在 *Proceedings of the IEEE/CVF International Conference on Computer Vision*，第 4624 – 4633 页。

Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sas-try、Amanda Askell、Pamela Mishkin、Jack Clark 等人。2021 年。从自然语言监督中学习可迁移的视觉模型。在 *International Conference on Machine Learning*，第 8748-8763 页。PMLR。

Shaoqing Ren、Kaiming He、Ross Girshick 和 Jian Sun。2015 年。Faster r-cnn：通过区域提议网络实现实时对象检测。*Advances in neural information processing systems*, 28。

Xiangxi Shi、Xu Yang、Jiuxiang Gu、Shafiq Joty 和 Jianfei Cai。2020 年。从另一角度寻找：用于变化字幕的视点自适应匹配编码器。在 *European Conference on Computer Vi- sion*，第 574 – 590 页。Springer。

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Yaoqi Sun, Liang Li, Tingting Yao, Tongyv Lu, Bolun Zheng, Chenggang Yan, Hua Zhang, Yongjun Bao, Guiguang Ding, and Gregory Slabaugh. 2022. Bidi- rectional difference locating and semantic consis- tency reasoning for change captioning. *International Journal of Intelligent Systems*, 37(5):2969 – 2987.

Hao Tan、Franck Dernoncourt、Zhe Lin、Trung Bui 和 Mohit Bansal。2019 年。通过语言表达视觉关系。在 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*，第 1873-1883 页。

Mingkang Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li. 2021. Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM*

International Conference on Multimedia，第 4858 至 4862 页。

Eric Tzeng、Judy Hoffman、Ning Zhang、Kate Saenko 和 Trevor Darrell。2014 年。深度领域混淆：最大化领域不变性。*arXiv preprint arXiv:1412.3474*。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Ł ukasz Kaiser 和 Illia Polosukhin。 2017年，你所需要的就是关注。 *Advances in neural information processing systems*, 30。

Ramakrishna Vedantam、C Lawrence Zitnick 和 Devi Parikh。2015 年。Cider：基于共识的图像描述评估。在 *Proceedings of the IEEE conference on computer vision and pattern recognition*，第 4566 – 4575 页。

Oriol Vinyals、Alexander Toshev、Samy Bengio 和 Dumitru Erhan。2015 年。展示和讲述：神经图像标题生成器。在 *Proceedings of the IEEE conference on computer vision and pattern recognition*，第 3156-3164 页。

Kelvin Xu、Jimmy Ba、Ryan Kiros、Kyunghyun Cho、Aaron Courville、Ruslan Salakhudinov、Rich Zemel 和 Yoshua Bengio。2015 年。展示、注意和讲述：利用视觉注意力生成神经图像标题。在 *International conference on machine learn- ing*，第 2048-2057 页。PMLR。

Linli Yao、Weiying Wang 和 Qin Jin。2022 年。使用预训练和对比学习进行图像差异字幕。在 *Proceedings of the AAAI Conference on* 和 *Artificial Intelligence* 中，第 3108 – 3116 页。

Luowei Zhou、Hamid Palangi、Lei Zhang、Houdong Hu、Jason Corso 和 Jianfeng Gao。2020 年。统一视觉语言预训练，用于图像字幕和 vqa。在 *Proceedings of the AAAI Conference on Artificial Intelligence*，第 34 卷，第 13041-13049 页。

数据集

CLEVR-Change (Park et al., 2019) 是由 CLEVR 引擎生成的合成数据集。图像中对象之间的几何差异已注释。它分为训练、验证和测试部分，分别有 67,660、3,976 和 7,970 个图像对。Spot-the-Diff (Jhamtani and Berg-Kirkpatrick, 2018) 描述了从 VIRAT 地面视频数据集中采样的 13,192 个真实图像对中的多个场景变化，并带有人工注释的字幕。平均而言，每幅图像对有 1.86 个句子来描述差异。字幕设置了两种解码策略，包括单句解码和多句解码。按照 Jhamtani 和 Berg-Kirkpatrick (2018) 的做法，我们通过设置

真实描述作为多个参考标题。Image-Editing-Request（Tan et al.，2019）是一个由相机镜头、绘画和动画组成的数据集，大多数图像都是真实的。它包含 3,939 个图像对，并带有由人类注释者编写的说明。

B 实施细节

IDC 自适应设置。视觉和语言编码器使用 CLIP ViT-B/32（Dosovitskiy 等，2020）初始化。句子长度为 32，语言编码器中的层数为 $N_G = 12$。文本嵌入的维度为 $d_T = 512$。图像的大小为 224×224，每幅图像由内核大小为 32、步幅为 32 和通道数为 768 的 2D 卷积网络处理。图像补丁的数量为 $n = 49$，图像补丁的维度为 $d_I = 768$。Transformer 内部和之间的层数分别为 $N_{intra} = 9$ 和 $N_{inter} = 3$。应用 Adam 优化器，初始学习率为 10^{-7}。通过将所有随机种子固定为 42，在两个 NVIDIA Tesla V100 GPU 上对模型进行了 12 个时期的训练。

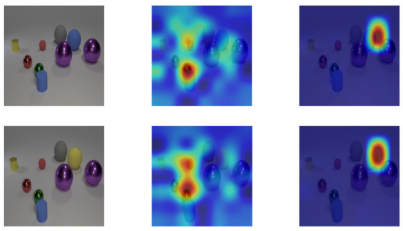
IDC 微调设置。我们使用 IDC 自适应模型初始化视觉编码器，并将词向量 $d_T =$ 的维数设置为 512。字幕模型是从头开始学习的。所有数据集上的字幕编码器和解码器中的 Transformer 层数均为 3。Transformer 中的注意层有 8 个头和 10% 的 dropout 概率，其隐藏层大小为 512。

对于直接CLIP微调，其视觉编码器的参数使用CLIP ViT-B/32初始化。其字幕模型的设置与CLIP4IDC中的设置相同。

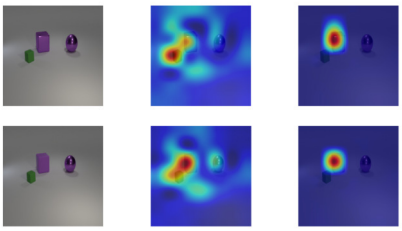
Adam 用于视觉编码器，初始学习率为 10^{-7}，用于字幕模型，初始学习率为 10^{-4}。该模型最多训练 50 个 epoch，批处理大小为 16。在推理中采用贪婪解码，最大 32 步，用于生成句子。实验在 NVIDIA Tesla V100 GPU 上进行。

C 定性结果

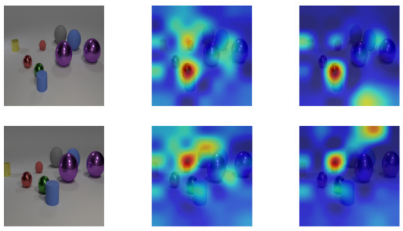
为了了解 IDC 适应的效果，图 3、4 和 5 分别可视化了 CLEVR-Change、Spot-the-Diff 和 Image-Editting-Request 数据集上的一些案例。



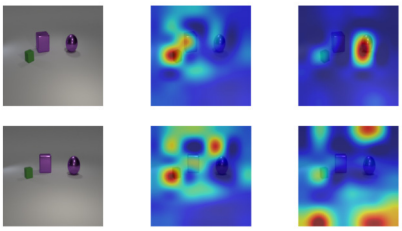
GT：蓝球变成黄色 CLIP4IDC：蓝球变成黄色



GT：绿色物体后面的大紫色金属块变成了橡胶 CLIP4IDC：大紫色金属球后面的大紫色金属块变成了橡胶

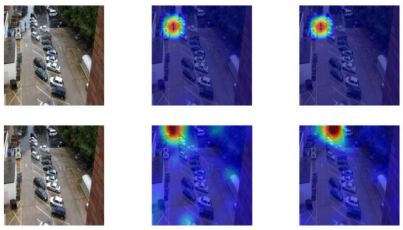


GT：没有区别 CLIP4IDC：没有变化

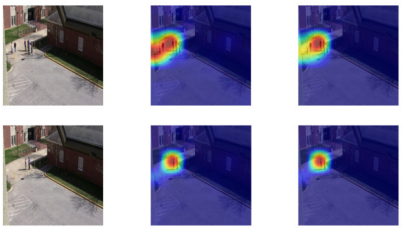


GT：没有变化 CLIP4IDC：没有变化

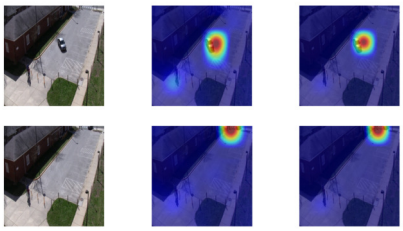
图 3：CLEVR-Change 上 CLIP4IDC 中视觉编码器输出的可视化。图片分为三列。第一列显示第一张和第二张原始图像。第二列显示编码器内部输出中的注意力图。最后一列显示编码器间输出中的注意力图。



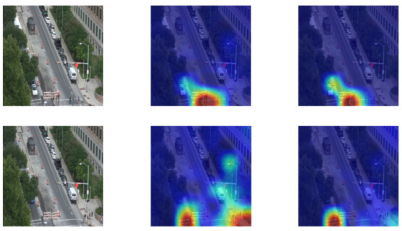
GT：行走的人已经不存在了 CLIP4IDC：在停车场行走的人已经不见了



GT：停车场里有一群人，人数较少 CLIP4IDC：右图中有两个人

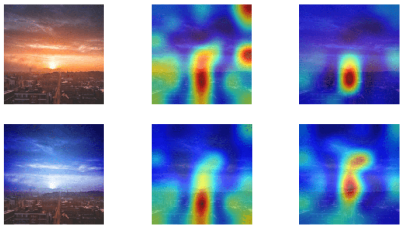


GT1：汽车已经开走了 GT2：有一辆车从图片右上方的入口进入 CLIP4IDC：汽车已经开走了

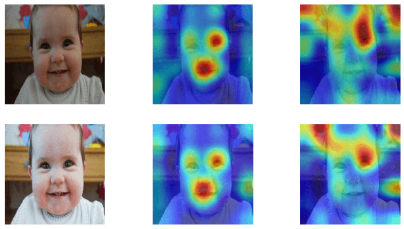


GT1：左边角落的白色汽车已经开走了 GT2：现在有人在等着过十字路口 CLIP4IDC：有人在人行道上行走

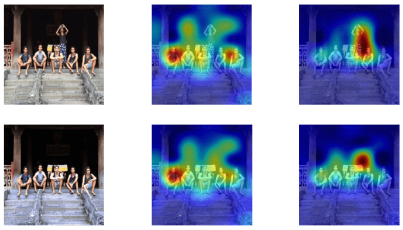
图 4：Spot-the-Diff 上 CLIP4IDC 中视觉编码器输出的可视化。



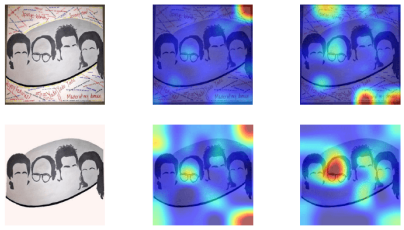
GT: color the sky blue
CLIP4IDC: make the image more blue



GT: brighten the entire photo
CLIP4IDC: brighten the photo



GT: remove girl in background
碌IP4IDC: remove the people from the background



GT: remove the background
CLIP4IDC: remove the background

图 5：CLIP4IDC 在 Image-Editting-Request 上对视觉编码器输出的可视化。

合成数据集 图 3 中的四个案例来自 CLEVR-Change。在每个案例的第二列中，我们可以看到 CLIP4IDC 的 *intra* 编码器关注的是更有可能需要信息来捕捉第二幅图像中的细粒度差异的区域。而在第三列中，*inter* 编码器过滤了与差异无关的信息，并关注第二幅图像中的变化。然而，对于下面两组图中显示的没有变化的案例，情况有所不同。*inter* 编码器似乎更均匀地关注各个区域以寻找任何变化，而不是固定在一个特定的区域上。

真实世界数据集图 4 和图 5 分别展示了 Spot-the-Diff 和 Image-Editing-Request 中的案例。可以看出，我们的 CLIP4IDC 捕捉到了真实世界和复杂案例中的细粒度差异。

D 基线方法的描述

最近的一些研究通过设计一个描述变化的语言模型，在 IDC 任务中取得了巨大进展，这些模型给出了 CNN 主干预先提取的视觉特征（He 等人，2016；Ren 等人，2015）。我们在实验中比较的基线如下：

- DUDA（2019）：双重注意模块
  提出一种区分干扰项和语义变化的方法，并定位变化。然后使用动态注意模块来描述变化。

- VAM (2020)：提出了一种新颖的视觉编码器，用于区分视点变化和语义变化。此外，它直接使用强化学习对模型进行微调，其中的奖励来自对生成的字幕的评估。 - IFDC (2021)：引入了一种语言生成器，该生成器由特征融合模块、基于相似性的差异查找模块和差异字幕模块组成。 - VACC (2021)：设计了一种差异编码器来编码视点信息并对差异进行建模。 - BiDiff (2022)：引入了一种变化字幕管道来定位图像对中的变化，并使用具有空间通道注意的解码器来生成描述。

这些方法通过细化或改进视觉特征来更好地捕捉图像对中的细粒度变化，从而不断提高模型精度。此外，受多任务学习成功的启发，还引入了以下训练方案。

• VACC (2021) 和 DUDA+Aux (2021)：这两项工作都提出了辅助模块，以将生成的字幕和前图像的复合特征与后图像特征进行匹配。 • IDC-PCL (2022)：提出了一种“ 预训练和微调 ”范式，包含以下三个预训练任务。给定视觉语言上下文，应用掩码语言建模 (MLM) 和掩码视觉对比学习 (MVCL) 任务分别将视觉上下文映射到语言并重建掩码图像特征。引入细粒度差异对齐 (FDA) 将字幕重写为硬样本，以最大化文本和图像对联合表示中的连接。 • CC-Full (2022)：这项工作提出共同训练基于文本的图像处理 (TIM) 和更改字幕 (CC) 模块。 CC 模块生成字幕，使用强化学习框架的 TIM 模块对其进行评估。TIM 模块生成图像，使用生成对抗网络的 CC 模块对其进行评估。