# CK-Transformer：用于指代表达式理解的常识知识增强型 Transformer

**Zhi Zhang**阿姆斯特丹大学 ILLC z.zhang@uva.nl

**Helen Yannakoudakis**信息学系，伦敦国王学院 helen.yannakoudakis@kcl.ac.uk

**Xiantong Zhen**
联影
zhenxt@gmail.com

**Ekaterina Shutova**
阿姆斯特丹大学 ILLC,
e.shutova@uva.nl

arXiv:2302.09027v1 [cs.CV] 17 Feb 2023

## 摘要

多模态指代表达理解（REC）旨在定位由自然语言表达式描述的图像区域，近年来在研究界日益受到关注。本文特别关注结合常识知识的指代表达理解（KB-Ref），这项任务通常需要超越空间、视觉或语义信息进行推理。我们提出了一种新型的常识知识增强变换器（CK-Transformer）框架，能够有效地将常识知识融入图像中物体的表示，从而帮助识别表达式所指的目标物体。我们在多个 KB-Ref 任务基准数据集上进行了广泛实验。结果表明，所提出的 CK-Transformer 达到了新的最先进水平，其准确率比现有最先进方法绝对提高了 3.14%。 [1]。

## 1 引言

指代表达理解（REC）旨在根据输入的自然语言表达，在图像中定位目标对象/区域。这项任务的性质要求进行多模态推理，并联合理解视觉与语言。过去几年中，研究者提出了若干 REC 任务和数据集，例如 RefCOCO （Yu 等人，2016）、RefCOCOg（Mao 等人，2016）和 RefCOCO+ （Yu 等人，2016）（RefCOCOs）。这些“传统”REC 任务通常侧重于根据物体的视觉或空间信息来识别物体，例如其颜色、形状、位置等；因此，主要评估模型基于视觉属性和空间关系进行推理的能力。

然而在实际生活中，人们常常使用非视觉或非空间信息来描述物体——例如，考虑这句话（表达式）：“给我拿些柔软但富含淀粉的东西吃”（Wang 等人，2020）。这类实例要求进行超越空间和视觉属性的推理，并需要结合表达式中蕴含的常识知识（事实）进行解读，例如有关图像中哪些物体可食用、柔软且富含淀粉的知识。最近，Wang 等人 （2020）提出了一个新数据集 KB-Ref，用于评估模型不仅基于视觉和空间特征，还基于常识知识进行推理的能力。该数据集经过专门设计，要求至少使用一条来自知识库（KB）的事实，才能识别目标对象（由某个表达式指代）。

因此，从知识库中搜索合适的事实也是 KB-Ref 的关键环节。与唯一已有的相关工作（Wang 等人，2020）相比，该工作针对每个物体候选项，保留事实的平均 Word2Vec（Mikolov 等人，2013）嵌入与给定表达式之间余弦相似度最高的前 K 个事实；我们的框架则侧重于多模态嵌入，并同时对表达式和图像进行推理，以识别前 K 个事实。多模态特征编码了更丰富的信息，有助于在不同的（语义）上下文中改进推理并识别相关事实；例如，上述表达式可以通过图像中的物体“香蕉”来回答（或者等效地，通过另一张图像中的物体“土豆泥”来回答）。

本文提出了一种用于KB-Ref的新型多模态框架——常识知识增强变换器（Commonsense Knowledge Enhanced Transformers，简称CK-Transformer或CK-T）——将（前K条）事实整合到图像中的所有候选对象中，以更好地识别目标对象。具体而言，我们的贡献有四点：1）我们提出了CK-T（见

[1]代码将发布在 https://github.com/FightingFighting/CK-Transformer

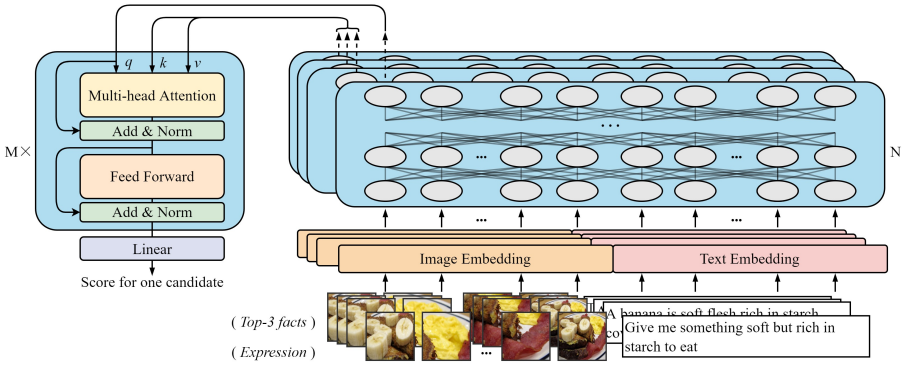



图 1：CK-Transformer。对于每个候选项（图中的第一个候选项），给定一个表达式、一组视觉区域候选项和前 K 个事实（图中 K=3 个），模型首先将表达式和所有前 K 个事实编码为相应的多模态特征，然后融合这些特征，并将其映射为该候选项的匹配分数。

图1），该方法能够有效整合来自不同模态的多样化输入：视觉、指代表达和事实；2）据我们所知，我们的方法首次将视觉信息引入（前K条）相关事实的识别过程；3）我们的方法仅使用每个（候选）对象的前3条事实，便达到了新的最先进水平；与现有工作使用多达前50条事实相比，该方法的效率也显著更高；4）我们将事实引入“传统”的指代表达理解（REC）任务，从而提升了性能。

## 2 相关工作

**结合常识知识的指代表达理解**与传统的指代表达理解任务不同（详情参见附录A ），KB-Ref 侧重于根据需要运用常识知识进行推理的表达式查询物体。作者针对事实、表达式和图像的整合，对基线模型 ECIFA 进行了基准测试，并通过比较图像特征与图像中所有候选对象对应的前 K 个事实特征之间的匹配分数，选出目标对象（Wang 等人，2020）。在我们的框架中，我们通过比较事实与表达嵌入之间的余弦相似度，为每个候选项选出前 K 个事实；这些嵌入由多模态编码器生成，而非由 ECIFA 模型所用的文本编码器生成。

**预训练视觉语言编码器** 已有多种预训练多模态编码器（Su 等人，2019；Li 等人，2019；Chen 等人，2020；Tan 和 Bansal，2019）已被提出，并在视觉语言任务上取得了最先进的结果。目前，UNITER（Chen 等人，2020）作为性能强大的预训练编码器在REC任务（RefCOCOs）上取得了最佳性能。本文对UNITER进行了改造，使其能够作为事实检索中的多模态编码器，并成为CK-T的一部分。

## 3 方法

我们将 KB-Ref 建模为一个分类问题，基于图像 $I$ 中的一组候选项（图像区域） $I = \{c_j\}_{j=1}^{n}$，这些候选项 $I = \{c_j\}_{j=1}^{n}$ 来自真实标签或预训练目标检测器的预测。具体而言，给定一个表达式 $e$、一幅图像 $I$ 和一个知识库，我们首先为每个候选项 $c_i$从知识库中检索前 K 个事实 $F_i^K = \{f_j\}_{j=1}^{k}$，然后将 $e$、 $I$和$F_I^K$（关于 $I$所选取的事实）同时输入 CK-T，以便在所有候选项中预测目标对象。

### 3.1 基于图像的事实检索

对于每个候选项 $c_i$，我们会从知识库中检索给定图像的所有事实（关于我们框架中使用的知识库的详细信息，请参见附录D），检索依据是其类别（例如，候选物体可能属于“汽车”类别）。然后，我们计算这些事实与给定表达式之间的余弦相似度，相似度由一个相似度提取器得出，该提取器是通过改造 UNITER 训练得到的。具体而言，将图像–表达式对和图像–事实对作为输入，分别从 UNITER 跨模态输出的位置提取表达式特征和事实特征（对应于 [CLS] 词元的输入，详情请参见附录B ），然后计算二者之间的余弦相似度。在训练过程中，受Devlin 等人(2018)的启发，我们将50% 的真实事实替换为知识库中的随机事实（相似度为0），

以帮助模型更好地区分有用事实与无用事实。最后，我们为每个候选项保留与表达式相似度较高的前 K 个事实 $F_i^K$ $c_i$。

### 3.2 常识知识增强型 Transformer

CK-T 由一个双模态编码器（见3.2.1）和一个事实感知分类器（见3.2.2）组成。

#### 3.2.1 双模态编码器

双模态编码器以 UNITER-base 和 N=12 层（Chen 等人，2020）初始化，用于整合图像和文本两种模态（$e$ 或 $f_i$）。具体而言，生成由图像嵌入和文本嵌入组成的输入嵌入 $\boldsymbol{E}^{Inp}$（与 UNITER 相同，详见附录 C）后，对于每个候选项 $c_i$，我们分别从同一编码器中与 $\boldsymbol{f}_i^e$ 和 $\boldsymbol{f}_i^f$ 对应的视觉输出位置提取表达式感知和事实感知的对象特征；提取过程以所有候选项 $I$ 以及 $e$ 或 $f_i$为输入。

#### 3.2.2 事实感知分类器

事实感知分类器由多头注意力层和全连接层组成。对于每个候选项 $c_i$，$\boldsymbol{f}_i^e$ 和 $\boldsymbol{F}_i^f$（$c_i$ 的全部 K 个事实感知对象特征）同时输入整合器（键和 值 来自 $\boldsymbol{F}_i^f$，查询来自 $\boldsymbol{f}_i^e$），并融合为一个三源对象特征$\boldsymbol{f}_i^t$（图像、表达式和前 K 个事实）。

最后，通过一个线性层将$\boldsymbol{f}_i^t$ 映射为 $c_i$的匹配分数 $s_i$，优化目标是最小化所有候选项 $I$ 对应的所有分数$\{s_j\}_{j=1}^n$ 上的交叉熵损失。

## 4 结果

我们将 CK-T 与现有方法进行比较，分别考察 KB-Ref 任务中不使用和使用事实的情况。接着，我们探讨引入视觉信息对事实检索的重要性。此外，我们将事实引入传统的 RefCOCOs 数据集。该数据集采集自 MSCOCO（Lin 等，2014年），但表达式类型和物体候选项设置有所不同。我们使用现成的检测器 Faster R-CNN（搭载 ResNet-101）提取图像区域特征（Ren 等，2015），依据边界框（bbxes）（真实标签或检测器的预测结果）进行提取。有关这些数据集和实验设置的详情，请参见附录 D 和E 。[2]。

[2]包括关于我们模型效率的讨论

| 模型 | 准确率 (%) Val | Test |
|---|---|---|
| CMN（Hu 等人，2017年） | 41.28 | 40.03 |
| SLR（Yu 等人，2017年） | 44.03 | 42.92 |
| VC（Niu 等人，2019） | 44.63 | 43.59 |
| LGARNs（Wang 等人，2019） | 45.11 | 44.27 |
| MAttNet（Yu 等人，2018） | 46.86 | 46.03 |
| ECIFA-nf（Wang 等人，2020） | 37.95 | 35.16 |
| CK-T-nf（我们的方法） | **58.02** | **57.53** |
| ECIFA (Wang 等人，2020) | 59.45 | 58.97 |
| MAtt+E (Wang 等人，2020) | 64.08 | 63.57 |
| CK-T-Word2Vec | 60.40 | 61.39 |
| CK-T-Uw/oImage | 64.44 | 64.78 |
| CK-T（我们的方法） | **65.62** | **66.71** |
| 人类 | – | 90.31 |

表1：使用真实边界框和物体类别时，KB-Ref 数据集在不使用和使用事实情况下的准确率（分别为上半部分和下半部分）。

通过对 K 和 M 进行参数搜索（见附录F 中的图3 和图4），我们将 M 设为 = 2 个事实感知分类器模块，并为每个候选项选取前3条事实。

**真实边界框和类别**遵循 Wang 等人 (2020) 的做法，我们报告了在 KB-Ref 上不使用和使用事实时的结果。如表 1 （上半部分）所示，CK-T-nf 是不使用事实的 CK-T 版本，[3], 在测试集上的准确率达到 57。53%，比不使用事实的现有方法高约 11% – 22%。从表格下半部分可以看到，我们的事实增强型 CK-T 模型在测试集上取得了最高准确率（66。71%），比 ECIFA（由 Wang 等人提出的基线模型 Wang 等人 (2020)）高 7。74%，并比 MAtt+E[4]高 *3*。14%。+值得注意的是，ECIFA 和 MAtt+E 每个候选项都纳入了前50 条事实，这远多于我们的 CK-T 所使用的前3 条事实。我们推测，这是因为我们的事实检索方法利用了多模态事实与表达嵌入。

**预测边界框和类别** 为了便于与 ECIFA-d（Wang 等人，2020）进行公平比较，我们也为每张图像使用最多 10 个检测到的边界框（CK-T-m10）。如表 2所示，CK-T-m10 的准确率

[3]事实句子中的所有词元都被替换为单个 [MASK]词元。[4]Wang 等人 (2020)将其事实融合模块——情景记忆模块 (E)——引入 MAttNet 模型 (Matt) 中，该模型Yu 等人,2018)广泛用于传统的指代表达理解。

| 模型 | 准确率（%） | |
|---|---|---|
| | Val | Test |
| ECIFA-d（Wang 等人，2020） | 24.11 | 23.82 |
| CK-T-m10（我们的方法） | 28.33 | 28.71 |
| CK-T-m100（我们的方法） | **35.66** | **35.96** |

表 2：使用预测的边界框和物体类别在 KB-Ref 上的准确率。

比 ECIFA-d 在测试集上的准确率高出 ≈ 5%。CK-T-m100 是一种最多使用 100 个检测到的边界框的变体，与 CK-T-m10 相比，准确率大幅提升了 ≈ 7%。这种差异主要归因于正确检测到的边界框数量以及预测类别数量的增加。具体而言，我们发现，使用排名前 100 的边界框后，包含目标边界框的样本数从 18，901 增至 31，653；在这些目标边界框中，类别预测正确的数量从 11，324 增至 15，928，而 KB-Ref 数据集的样本总数为 43，284。这也可以解释 CK-T 与 CK-T-m10 之间准确率的大幅下降。

**将图像特征融入事实检索**我们尝试了多种事实检索方法，并在 KB-Ref 上评估其效果（表 1）。首先，我们使用（Wang 等人，2020）检索到的前 k 个事实；他们使用预训练的 Word2Vec（Skip-Gram）（Mikolov 等人，2013）进行事实检索（CK-T-Word2Vec）。随后，我们还从仅以文本为输入的相似度预测器中选取事实（CK-T-Uw/oImage）[5]，而非像 CK-T 那样使用图文对。如表1所示，与 CK-T-Word2Vec 相比，CK-T-Uw/oImage 和 CK-T 在测试集上都取得了更高的准确率。与 CK-T-Uw/oImage 相比，CK-T 的准确率高出约 2% 。这主要归因于事实检索过程中使用了额外的视觉信息（这些事实检索方法选出的事实示例见附录 I）。

**在传统指代表达理解任务中引入事实** 我们利用 CK-T，将 KB-Ref 所用知识库中的事实融入 RefCOCOs 任务。表 3 展示了基于真实边界框和类别的结果对比（关于检测结果的讨论见附录 G）。与 $U_{REC}$[6]相比，引入事实的模型取得了更好的

[5]受 Frank 等人 （2021）的启发，我们将所有物体候选项特征替换为所有图像区域特征的平均值。[6]Chen 等人 （2020）通过微调 UNITER，在 Re-fCOCOs 上取得了最先进的结果。（我们重新微调该模型

| Task | | 准确率（%） | |
|---|---|---|---|
| | | $U_{REC}$ | 简介事实 |
| Ref-COCO | Val | 90.98 | **91.43** |
| | 测试 A | 91.50 | **92.09** |
| | 测试 B | 90.89 | **90.95** |
| Ref-COCO+ | Val | 83.23 | **83.45** |
| | 测试 A | 85.09 | **85.49** |
| | 测试 B | **79.08** | **79.08** |
| Ref-COCOg | Val | 86.23 | **87.21** |
| | Test | 85.79 | **87.59** |

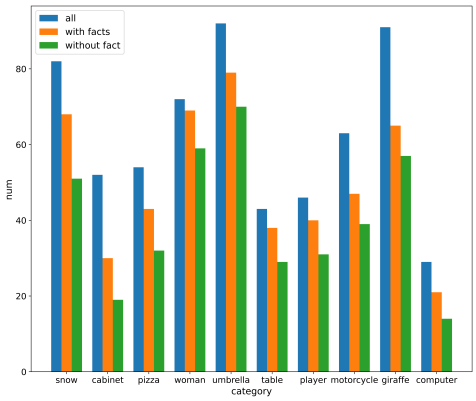
表3：将事实引入 RefCOCO、RefCOCO+和 RefCOCOg。RefCOCO 和 RefCOCO+有两个不同的测试集，即测试 A 和测试 B，分别包含图像中的多个人和多个物体。

准确率相当或更高，适用于所有 RefCOCOs 任务；其中 RefCOCOg 的提升幅度高于 RefCOCO 和 RefCOCO+。这是因为与 RefCOCO 和 RefCOCO+ 相比，RefCOCOg 图像中的同类别对象候选项较少，平均每张图像分别为 1.63 个和 3.9 个（Yu 等人，2016），因此整合到不同候选项中的检索事实更加多样化（我们首先根据类别检索事实），有助于区分候选项。McNemar检验也证实了这一差异：与引入事实之前相比，引入事实后 RefCOCOg 的错误比例变化具有统计显著性（$p$值 $= 1。19e{-}08 < \alpha = 0.05$），而 RefCOCO 和 RefCOCO+ 中的错误比例则相近（McNemar检验的详细信息见附录 H ）。然而，常识知识对传统指代表达理解的总体影响并不显著。这主要是因为 RefCOCOs 候选项中的类别数量（78）远少于 KB-Ref 中的 1805 个（Wang 等人，2020）。这限制了所选事实的多样性，进而影响常识知识的实用程度。

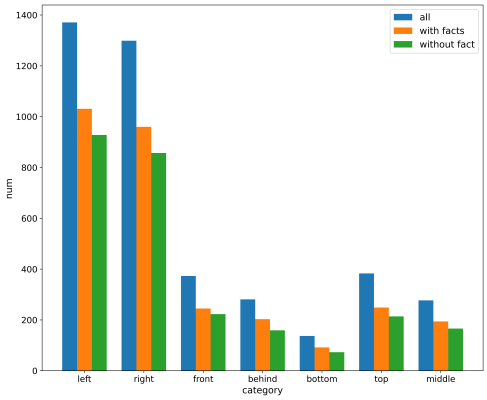
## 5 分析

为了探究常识知识在哪些情况下能发挥作用，我们对 KB-Ref 测试集上的模型表现进行了细粒度分析。具体而言，我们从三个方面比较了使用事实和不使用事实的模型（CK-T 和 CK-T-nf）所预测的样本：物体类别、空间关系和边界框大小。

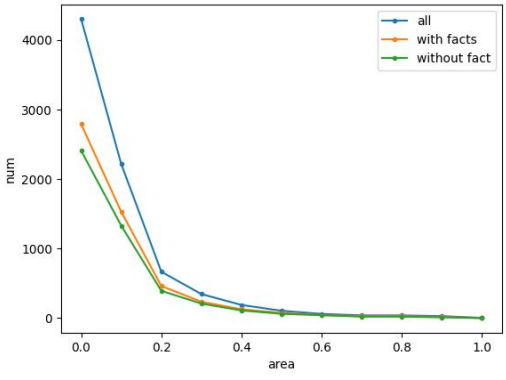
以进行公平比较并执行 McNemar检验）




（a）引入事实后提升最显著的前10个类别。




（b）空间关系分析。




（c）不同边界框尺寸分析。

图2：细粒度分析。全部：测试集中的样本总数；使用事实：CK-T正确预测的测试样本数；不使用事实：CK-T-nf正确预测的测试样本数。

**物体类别** 测试集包含 1502 个类别，CK-T 在其中 1347 个类别上的表现优于 CK-T-nf。提升最明显的前 10 个类别见图2(a)。对于表现未见提升的 155 个类别，我们发现每个类别的平均样本数为 6.68，因此结果的可靠性较低。

**空间关系** 接下来，我们研究在有事实和没有事实的情况下解决指代表达理解任务在多大程度上依赖空间推理，以及对于物体之间的哪些特定空间关系，使用事实最为关键。与（Kazemzadeh等人，2014年；Johnson等人，2017年）的研究类似，我们关注以下空间关系：左、右、前、后、底部、顶部、中间。如图2(b)所示，使用事实的模型（CK-T）在所有空间关系上的表现均优于不使用事实的模型（CK-T-nf）。

**边界框的大小** 随后，我们以边界框的大小作为代理指标，研究事实在识别不同大小物体时所发挥的作用。我们使用边界框的归一化面积作为边界框大小的度量。如图 2(c)所示，事实能够提升所有边界框大小下的模型性能。

## 6 结论

本文提出了 CK-Transformer，该模型将常识知识和表达式有效地整合到相应视觉物体的表示中，以便在 KB-Ref 上进行多模态推理。仅使用最相关的前 3 个事实，我们的 CK-Transformer 便在 KB-Ref 上取得了新的最先进的性能。我们还证明，视觉信息有助于事实检索。最后，我们表明，常识知识能够在三个不同的数据集上提升传统REC任务的性能。

## 7 局限性

我们模型的计算需求受事实数量的影响。具体而言，我们使用一块 Titan RTX 图形处理器，以 64 的批大小训练 CK-Transformer 10000 步；事实数量分别为 3、5、10、20 时，训练耗时依次为 2.5、3、3.5、7 天。CK-Transformer 在训练时的平均处理速度分别为每秒 3.8、2.8、2.1、0.7 个样本，在测试时则分别为每秒 8.3、7.3、6.6、1.1 个样本，事实数量为

以上所述的这些数值。由此可见，我们模型的计算需求相当可观，未来的工作应考虑提高计算效率，从而减少环境影响。

## 参考文献


Peter Anderson、Xiaodong He、Chris Buehler、Damien Teney、Mark Johnson、Stephen Gould 和 Lei Zhang。2018 年。用于图像描述和视觉问答的自底向上和自顶向下注意力。载于*IEEE* 计算机视觉与模式识别会议论文集，第 6077–6086 页。

Yen-Chun Chen、Linjie Li、Licheng Yu、Ahmed El Kholy、Faisal Ahmed、Zhe Gan、Yu Cheng 和 Jingjing Liu。2020 年。UNITER：通用图像文本表征学习。载于欧洲计算机视觉会议，第 104–120 页。Springer。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018 年。BERT：用于语言理解的深度双向 Transformer预训练。*arXiv* 预印本 *arXiv:1810.04805*。

Stella Frank、Emanuele Bugliarello 和 Desmond Elliott。2021 年。视觉与语言还是视觉服务于语言？多模态 Transformer中的跨模态影响。*arXiv* 预印本 *arXiv:2109.04448*。

Ronghang Hu、Marcus Rohrbach、Jacob Andreas、Trevor Darrell 和 Kate Saenko。2017 年。使用组合模块化网络对指代表达中的关系进行建模。载于*IEEE*计算机视觉与模式识别会议论文集，第1115–1124页。

Ronghang Hu、Huazhe Xu、Marcus Rohrbach、Jiashi Feng、Kate Saenko 和 Trevor Darrell。2016年。自然语言物体检索。载于*IEEE*计算机视觉与模式识别会议论文集，第4555–4564页。

Justin Johnson、Bharath Hariharan、Laurens Van Der Maaten、Li Fei-Fei、C Lawrence Zitnick 和 Ross Girshick。2017年。CLEVR：用于组合语言和初等视觉推理的诊断数据集。载于*IEEE* 计算机视觉与模式识别会议论文集，第2901–2910页。

Sahar Kazemzadeh、Vicente Ordonez、Mark Matten 和 Tamara Berg。2014年。ReferItGame：在自然场景照片中指称物体。载于*2014*年自然语言处理实证方法会议（*EMNLP*）论文集，第787–798页。

Ranjay Krishna、Yuke Zhu、Oliver Groth、Justin John- son、Kenji Hata、Joshua Kravitz、Stephanie Chen、Yannis Kalantidis、Li-Jia Li、David A Shamma 等。2016。Visual genome：使用众包密集图像标注连接语言与视觉。*arXiv* 预印本 *arXiv:1602.07332*。

Liunian Harold Li、Mark Yatskar、Da Yin、Cho-Jui Hsieh 和 Kai-Wei Chang。2019。Visualbert：一种简单且高效的视觉与语言基线。*arXiv* 预印本 *arXiv:1908.03557*。

Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。2014年。微软 COCO：上下文中的常见物体。载于欧洲计算机视觉会议，第 740–755 页。Springer。

Ilya Loshchilov 和 Frank Hutter。2017年。解耦权重衰减正则化。*arXiv* 预印本 *arXiv:1711.05101*。

Junhua Mao、Jonathan Huang、Alexander Toshev、Oana Camburu、Alan L Yuille 和 Kevin Murphy。2016年。无歧义物体描述的生成与理解。载于*IEEE* 计算机视觉与模式识别会议论文集，第 11–20 页。

Tomas Mikolov、Ilya Sutskever、Kai Chen、Greg S Corrado 和 Jeff Dean。2013年。词语与短语的分布式表征及其组合性。载于神经信息处理系统进展，第 3111–3119 页。

Yulei Niu、Hanwang Zhang、Zhiwu Lu 和 Shih-Fu Chang。2019。变分上下文：利用视觉和文本上下文对指代表达进行定位。*IEEE*模式分析与机器智能汇刊，43(1):347–359。

Vicente Ordonez、Girish Kulkarni 和 Tamara Berg。2011。Im2Text：使用 100 万张带字幕照片描述图像。神经信息处理系统进展，24:1143–1151。

Shaoqing Ren、Kaiming He、Ross Girshick 和 Ji an Sun。2015。Faster R-CNN：利用区域提议网络实现实时目标检测。神经信息处理系统进展，28:91–99。

Piyush Sharma、Nan Ding、Sebastian Goodman 和 Radu Soricut。2018。Conceptual Captions：一个经过清洗、添加上位词的图像替代文本数据集，用于自动图像描述。载于计算语言学协会第 *56* 届年会论文集（第 *1* 卷：长篇论文），第 2556–2565 页。

Robyn Speer、JoshuaChin 和 CatherineHavasi。2017年。Conceptnet 5.5：通用知识的开放式多语言图谱。载于第三十一届*AAAI*人工智能会议。

Weijie Su、Xizhou Zhu、Yue Cao、Bin Li、Lewei Lu、Furu Wei 和 Jifeng Dai。2019。VL-BERT：通用视觉语言表示的预训练。*arXiv*预印本 *arXiv:1908.08530*。

Hao Tan 和 Mohit Bansal。2019。Lxmert：从 Transformer 中学习跨模态编码器表征。*arXiv* 预印本 *arXiv:1908.07490*。

Niket Tandon、Gerard De Melo 和 Gerhard Weikum。2017年。WebChild 2.0：细粒度常识知识蒸馏。载于 *ACL 2017* 论文集：系统演示，第 115–120 页。

Peng Wang、Dongyang Liu、Hui Li 和 Qi Wu。2020。《给我一些吃的：利用常识知识进行指代表达式理解》。载于第*28*届 *ACM* 国际多媒体会议论文集，第28–36页。

Peng Wang、Qi Wu、Jiewei Cao、Chunhua Shen、Lianli Gao 和 Anton van den Hengel。2019。《邻里守望：通过语言引导的图注意力网络进行指代表达式理解》。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第1960–1968页。

Yonghui Wu、Mike Schuster、Zhifeng Chen、Quoc V Le、Mohammad Norouzi、Wolfgang Macherey、Maxim Krikun、Yuan Cao、Qin Gao、Klaus Macherey 等。2016。Google神经机器翻译系统：弥合人工翻译与机器翻译之间的差距。*arXiv* 预印本 *arXiv:1609.08144*。

Licheng Yu、Zhe Lin、Xiaohui Shen、Jimei Yang、Xin Lu、Mohit Bansal 和 Tamara L Berg。2018年。Mattnet：用于指代表达理解的模块化注意力网络。载于*IEEE*计算机视觉与模式识别会议论文集，第 1307–1315 页。

Licheng Yu、Patrick Poirson、Shan Yang、Alexander C Berg 和 Tamara L Berg。2016年。指代表达中的上下文建模。载于欧洲计算机视觉会议，第 69–85 页。Springer。

Licheng Yu、Hao Tan、Mohit Bansal 和 Tamara L Berg。2017年。用于指代表达的联合说话者-听者-强化者模型。载于*IEEE*计算机视觉与模式识别会议论文集，第7282–7290页。

Hanwang Zhang、Yulei Niu 和 Shih-Fu Chang。2018 年。通过变分上下文在图像中实现指代表达式落地。载于 *IEEE* 计算机视觉与模式识别会议论文集，第 4158–4166 页。

## A 指代表达式理解

指代表达理解的早期方法通过结合卷积神经网络（CNNs）和循环神经网络（RNNs）对图像与语言进行联合嵌入，并预测给定输入表达式和图像时概率最大的目标物体（Mao et al.，2016年；Huet al.，2016年；Zhang et al.，2018年）。为了对输入表达式中编码的不同类型信息（主体外观、位置以及与其他物体的关系）进行建模，后续工作采用模块化（注意力）网络，将输入与图像中的相应区域进行“匹配”，并将匹配得分最高的区域预测为目标（Hu etal.，2017年；Yuet al.，2018年）。

## B UNITER

UNITER 在四个大规模图文数据集上，使用四项预训练任务进行训练：掩码语言建模（MLM）、掩码区域建模（MRM）、图文匹配（ITM）和词区域对齐（WRA）。这四个数据集分别是 COCO（Lin 等人，2014年）、Visual Genome（Krishna 等人，2016）、Conceptual Captions（Sharma 等人，2018）和 SBU Captions（Ordonez 等人，2011）。这使 UNITER 能够捕捉图像与语言之间的细粒度对齐。UNITER 的架构与 BERT（Devlin 等人，2018）类似，不同之处在于输入和输出。具体而言，输入由一幅图像（一组视觉区域候选项）、一个句子和 [CLS]标记组成，它们分别在 UNITER 顶层产生不同的输出，即视觉输出、语言输出和跨模态输出。

## C 输入嵌入

与 UNITER 相同，我们提取输入嵌入 $\boldsymbol{E}^{Inp}$，其由分别对应于候选物体 $I$和文本（一个表达式 $e$ 或一个事实 $f_i$）的图像嵌入和文本嵌入组成。

**图像嵌入**图像嵌入 $\boldsymbol{E}_I$ 通过将三种嵌入相加计算得出：视觉特征嵌入、视觉几何嵌入和模态分段嵌入。我们首先使用 Faster R-CNN（池化 RoI 特征）提取所有候选项的视觉特征 $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_n\}$，并为所有候选项构建几何特征 $\boldsymbol{G} = \{\boldsymbol{g}_1, \boldsymbol{g}_2, ..., \boldsymbol{g}_n\}$，其中 $\boldsymbol{g}_i$ 是一个 7 维向量，包含与边界框对应的几何信息候选项 $c_i$，即归一化后的上、左、下、右坐标、宽度、高度和面积，记作 $\boldsymbol{g}_i = [x1, y1, x2, y2, w, h, w * h]$。视觉特征嵌入和视觉几何嵌入通过全连接层 $fc$将视觉特征和几何特征映射到同一向量空间中生成：

$$\boldsymbol{E}_I = LN(fc(\boldsymbol{V}) + fc(\boldsymbol{G}) + \boldsymbol{M}_I) \quad (1)$$

其中，$LN$ 是层归一化层，$\boldsymbol{M}_I$是图像输入的模态分段嵌入（类似于 BERT 模型中两个句子的片段嵌入）。

**文本嵌入** 类似地，文本嵌入$\boldsymbol{E}_T$ 基于三种不同类型的嵌入计算得到：词元嵌入、位置嵌入和模态嵌入（通常还存在第四种嵌入，即类似于 BERT 的句子片段嵌入，但在我们的任务中，表达式和事实都仅由一个句子组成，因此只使用第一个句子片段嵌入）。与 BERT（Devlin 等人，2018）类似，文本 $W = \{w_1, w_2, ..., w_u\}$ 首先通过 WordPieces（Wu 等人，2016）进行词元化，然后根据其在文本序列中的位置构建词元嵌入 $\boldsymbol{T} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, ...\boldsymbol{t}_v\}$ 和位置嵌入 $\boldsymbol{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, ... \boldsymbol{p}_v\}$。

$$\boldsymbol{E}_T = LN(\boldsymbol{T} + \boldsymbol{P} + \boldsymbol{M}_T) \quad (2)$$

其中，$\boldsymbol{M}_T$ 是文本输入的模态分段嵌入。

**输入嵌入** 最终输入嵌入$\boldsymbol{E}^{Inp}$ 通过拼接图像嵌入$\boldsymbol{E}_I$ 和文本嵌入$\boldsymbol{E}_T$计算得出：

$$\boldsymbol{E}_{Inp} = [\boldsymbol{E}_I, \boldsymbol{E}_T] \quad (3)$$

## D 数据集

我们使用 KB-Ref 数据集（Wang 等人，2020），旨在评估基于常识知识的指代表达式理解任务。KB-Ref 包含 43,284 个表达式，涵盖 16,917 张图像中的 1,805 个物体类别；此外还包含一个知识库，其中的键值（类别–事实）对来自三个常识知识资源：维基百科、ConceptNet（Speer 等人，2017年）和 WebChild（Tandon 等人，2017年）。KB-Ref 划分为训练集（31,284 个表达式，对应 9,925 张图像）、验证集（4,000 个表达式，对应

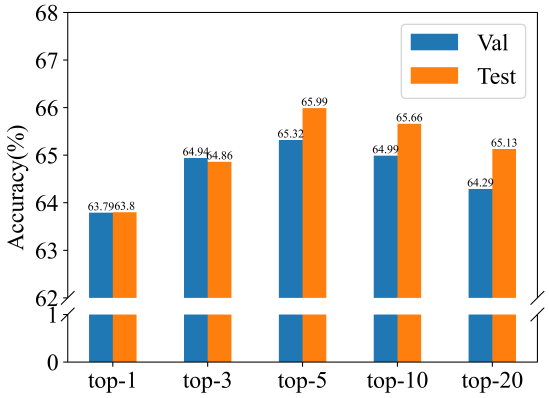



图3：不同事实数量（top-K）下的准确率。

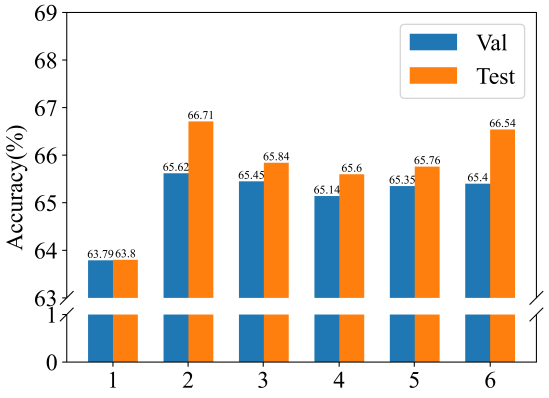



图4：不同事实感知分类器模块数量（M）下的准确率。

2,290 张图像）和测试集（8,000 个表达式，对应 4,702 张图像）。

我们还将常识知识引入指代表达式理解的传统任务/数据集，即 RefCOCO、Re-fCOCOg 和 RefCOCO+ [7]。这些数据集源自 MSCOCO图像数据集（Lin 等人，2014年），但在表达式类型和物体候选项设置方面有所不同。具体而言，RefCOCO+不允许在表达式中使用绝对位置词，并且大多数表达式侧重于物体的外观。Re- fCOCOg 中的表达式更长，包含更多描述性词语。RefCOCO 和 RefCOCO+ 在图像中包含更多属于同一类别的物体。

## E 实验设置

我们使用 Faster R-CNN 和 ResNet-101（Ren 等人，2015年）提取图像区域特征。该模型使用物体和属性标注在 Visual Genome（Krishna 等人，2016）上进行了预训练（Anderson 等人，2018）。对于边界框检测，我们保留置信度分数至少为 0.2 的边界框，该分数表示检测范围。在 CK-T 中，隐藏层维度为 768，多头注意力头的数量为 12。模型使用 Adamw（Loshchilov 和 Hutter，2017年）训练，学习率为 $6e^{-5}$，批大小为 64，使用 Titan RTX GPU。我们的 CK-Transformer 总共有 120M 个参数，其中事实感知分类器有 34M 个参数，双模态编码器有 86M 个参数。对于 UNITER 模型，我们采用与 *UNITER-base*相同的设置，但使用 Nvidia Apex[8] 用于加速训练。模型效率受事实数量的影响。具体而言，我们训练 CK-Transformer 10000 步，每步一个批次；事实数量分别为 3、5、10、20 时，训练耗时分别为 2.5、3、3.5、7 天。事实数量分别为 3、5、10、20 时，CK-Transformer 的平均训练速度分别为每秒 3.8、2.8、2.1、0.7 个样本，测试速度分别为每秒 8.3、7.3、6.6、1.1 个样本。

## F CK-T 结构的影响

我们探究了在开发集上，随着前 K 个事实的数量（K）和事实感知分类器模块（M）的变化，其对 KB-Ref 性能的影响。首先，我们保持事实感知分类器模块的数量不变，并将其设为 1，然后测试 K 从 1 到 20的不同取值。如图 3所示，随着 K 增大，性能开始提升，并在 K=5 时达到峰值，随后逐渐下降。

在第二个实验中，我们保持 K 不变，并将其设为 3，然后探究 M 取不同值的影响。我们发现，采用前3条事实和 2 个积分器层时，准确率最高，如图4所示。

### G 基于检测在传统指代表达理解任务中引入事实

在传统 REC任务中，基于检测到的边界框和类别引入事实后的结果如表 4所示。与基于真实边界框和类别的结果（表3）相比，基于检测的方法所取得的提升较小，甚至不如未引入事实的方法。

[7]遵循 Apache许可证2.0

[8]https://github.com/英伟达/apex

| Task | | 准确率 (%) | |
|---|---|---|---|
| | | U$_{REC}$ | 简介事实 |
| Ref-COCO | Val$^d$ | **81.15** | 81.06 |
| | 测试 A$^d$ | 86.85 | **86.87** |
| | 测试 B$^d$ | **74.48** | 73.97 |
| Ref-COCO+ | Val$^d$ | **74.74** | 74.68 |
| | 测试 A$^d$ | **81.05** | 80.70 |
| | 测试 B$^d$ | 65.88 | **66.07** |
| Ref-COCOg | Val$^d$ | 74.49 | **74.69** |
| | 测试集$^d$ | **75.24** | 74.86 |

表4：基于检测将事实引入 RefCOCO、RefCOCO+和 RefCOCOg（$d$）。

| Task | | McNemar检验（$p$ 值） |
|---|---|---|
| RefCOCO | 测试 A | 0.049 |
| | 测试 B | 0.905 |
| RefCOCO+ | 测试 A | 0.297 |
| | 测试 B | 0.966 |
| RefCOCOg | Test | $1.19e{-}08$ |

表 5：在 RefCOCOs 任务中引入事实前后模型之间的 McNemar检验。

## H McNemar检验

我们还报告了 RefCOCOs 任务上的准确率统计显著性（见表 3）。具体而言，我们分别在 RefCOCO、RefCOCO+和 Ref- COCOg 的测试集上，对引入事实前后的模型进行了 McNemar检验。如表5所示，对于 RefCOCOg 的测试集和 RefCOCO 的测试 A，$p$ 值 $= 1。19e{-}08$ 和 $p$值 $= 0.049$（$< 0.05$），这意味着引入事实后，错误比例相较于引入前存在统计学上的显著差异。然而，在其他任务（RefCOCO 的测试 B，以及 RefCOCO+的测试 A 和测试 B）中，引入事实后错误比例的变化在统计学上并不显著。这是合理的，因为检测产生的错误会影响事实检索（我们首先根据类别检索事实），从而将更多错误信息引入 CK-Transformer，导致性能下降。

## I 使用不同方法搜索事实的示例

如图5所示，这里展示了从三种不同的事实检索方法中选出的若干事实：CK-Transformer、CK-T-Uw/oImage 和 CK-T-Word2Vec。从表中可以看出，通常 CK-Transformer 模型（绿色）中的事实与指代表达式（蓝色）最相关，而 CK-T-Word2Vec 模型中的事实与表达式的相关性最差。

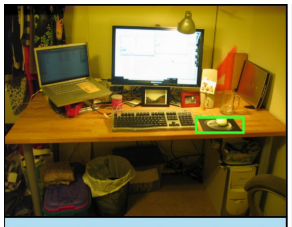
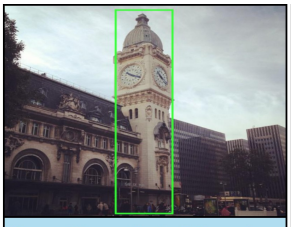
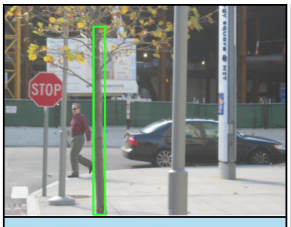
|  |  |  |  |
|---|---|---|---|
| **Exp:** The tool under the mouse can improve the usability of the mouse | **Exp:** The tall building has instr-ucments on the upper exterior walls that used as a reference to find out the time. | **Exp:** The most important part of the tree with branches and leaves | **Exp:** An access point as an underground public utility. |
| **Fact:** A mousepad enhances the usability of the mouse compared to using a mouse directly on a table. | **Fact:** Clock towers are a specific type of building which houses a turret clock and has one or more clock faces on the upper exterior walls. | **Fact:** The trunk is the most important part of the tree for timber production. | **Fact:** Manholes are often used as an access point for an underground public utility, allowing inspection, maintenance, and system upgrades. |
| **Fact (Uw/oImage):** A mousepad is a surface for placing and moving a computer mouse. | **Fact (Uw/oImage):** The tower has four clock faces, two of which are in diameter, at about high. | **Fact (Uw/oImage):** An automobile has a trunk. | **Fact (Uw/oImage):** A manhole is an opening to a confined space such as a shaft, utility vault, or large vessel. |
| **Fact (Word2Vec):** Mousepad on the mouse. | **Fact (Word2Vec):** Before the middle of the twentieth century, most people did not have watches, and prior to the 18th century even home clocks were rare. | **Fact (Word2Vec):** A trunk is an example of a box. | **Fact (Word2Vec):** These covers are traditionally made of metal, but may be constructed from precast concrete, glass reinforced plastic or other composite materials. |

图 5：示　例：不同搜索方法的事实搜索过程（使用排名第一的事实）：CK-T（绿色）、CK　　-T-Uw/oImage　（橙色）和 CK-T-Word2Vec（黄色）。