A Survey on Compositional Understanding

Yiwei Sun

2023.10.10

智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab

- 任务介绍
- 主流数据集
- 主流方法
- 破旧立新
- 总结反思

任务介绍

什么是视觉语言模型的组合理解能力?

视觉与语言存在一种共同的、基础的特性，即组合性。具体而言，图由目标与关系构成，文本由词组及其搭配构成。我们可以用一对多元函数表示组合性：

$$y_{image} = f(O; R)$$
$$y_{text} = g(W; R)$$

用(O, R)和(W, R)分别表示图像和文本

组合理解能力指的是视觉语言模型对实体关系的表达能力。对于语义相同的O和W，共享的R，有以下等式：

$$y_{image} = y_{text}$$

事实上，文本的表达的全面程度肯定不如图像，因为往往输入的是W和R的子集，因此有：

$$y_{image} \simeq y_{text}$$

任务介绍

视觉语言模型缺乏组合理解能力！

an old person kisses a young person

a young person kisses an old person

目前以CLIP为主的视觉语言模型无法对齐语义相似的图文对。

任务介绍

任务范式（如何度量模型的组合理解能力？准确率）

1. 单图-多文：

单图-多文数据集：

Score_1 = Sim( , `an old person kisses a young person`)

Score_2 = Sim( , `a young person kisses an old person`)

If Score_1 > Score_2 , then num_correct++

任务介绍

任务范式（如何度量模型的组合理解能力？准确率）

2. 单文-多图 ：

单文-多图数据集：

Score_1 = Sim( , an old person kisses a young person)

Score_2 = Sim( , an old person kisses a young person)

If Score_1 > Score_2 , then num_correct++

任务介绍

任务范式（如何度量模型的组合理解能力？）

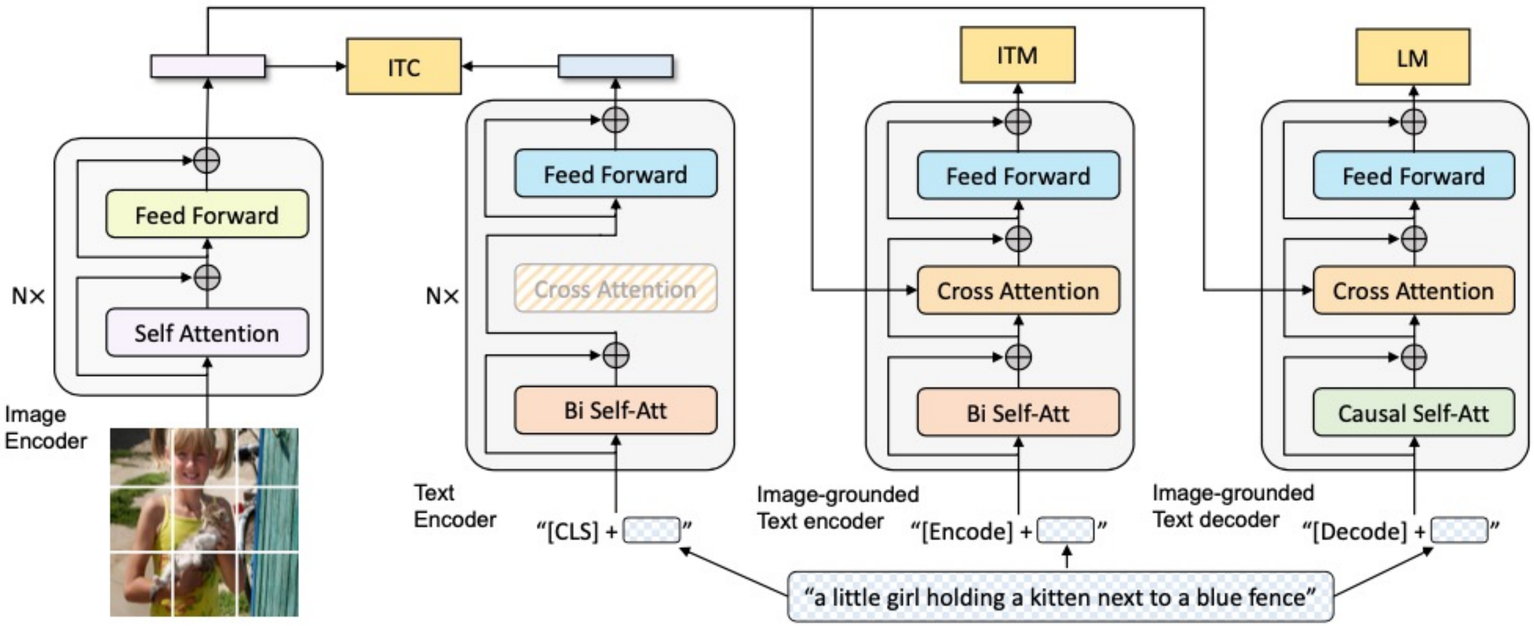
3. Image score & Text score & Group Score　　多图多文数据集

`an old person kisses a young person` `a young person kisses an old person`

Group Score：2对Image2Text和Text2Image的检索结果都正确。

任务介绍

任务范式（如何度量模型的组合理解能力？）　　　主要评估三类模块



ITC分支：计算图文特征的余弦相似度　　ITM分支：二分类向量第一维的取值

LM分支：计算输出“输入文本”的似然概率

任务介绍

目前研究组合理解的主流机构/高校（共计20篇，2022-2023）

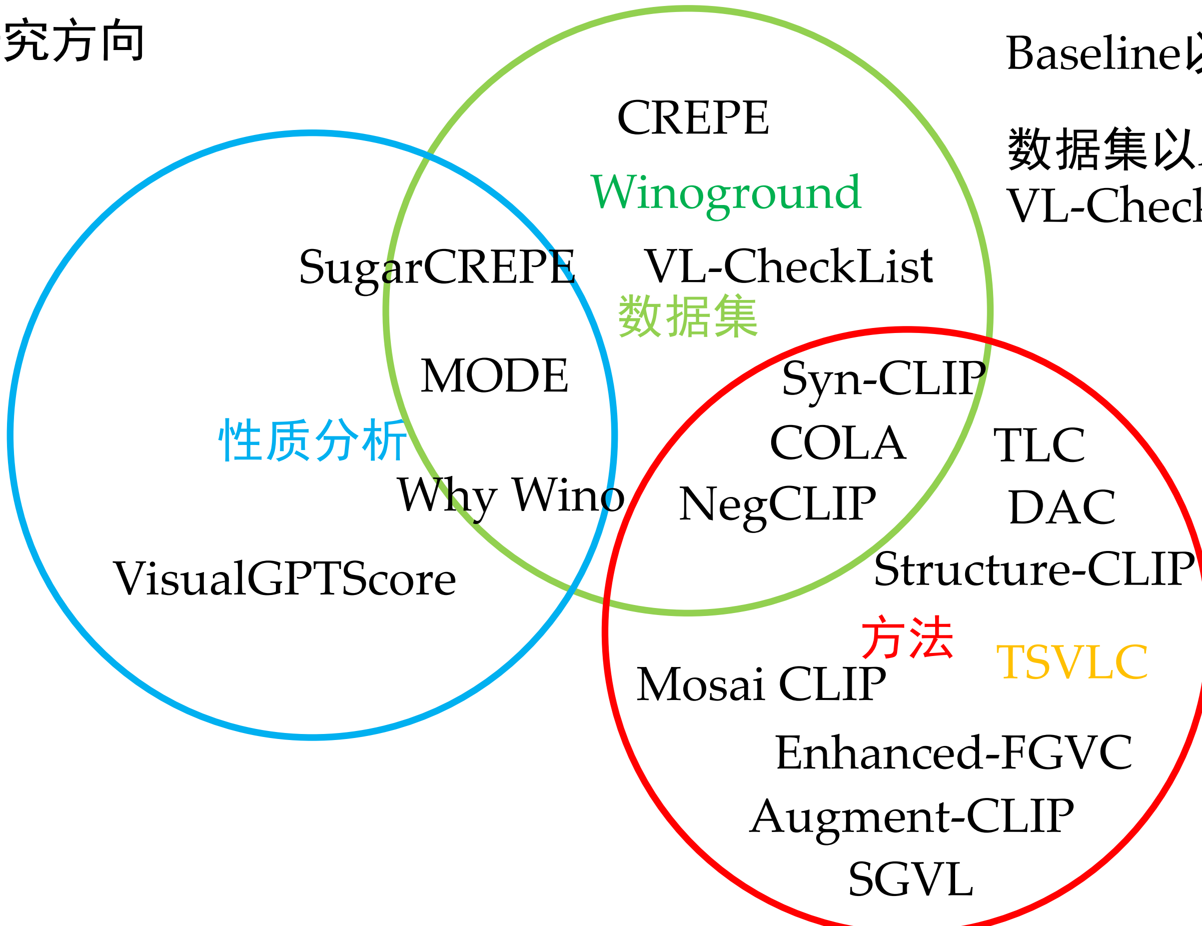
机构/高校	论文
Meta AI	Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality
	Cola: How to adapt vision-language models to Compose Objects Localized with Attributes?
	Simple Token-Level Confidence Improves Caption Correctness
	Coarse-to-Fine Contrastive Learning in Image-Text-Graph Space for Improved Vision-Language Compositionality
IBM Research	Dense and Aligned Captions (DAC) Promote Compositional Reasoning in VL Models
	Teaching Structured Vision & Language Concepts to Vision & Language Models
	Going Beyond Nouns With Vision & Language Models Using Synthetic Data
	Incorporating Structured Representations into Pretrained Vision & Language Models Using Scene Graphs
Google Research	What You See is What You Read? Improving Text-Image Alignment Evaluation
Mila	Contrasting Intra-Modal and Ranking Cross-Modal Hard Negatives to Enhance Visio-Linguistic Fine-grained Understanding

任务介绍

机构/高校	论文
卡耐基梅隆大学	VisualGPTScore: Visio-Linguistic Reasoning with Multimodal Generative Pre-Training Scores
	Cross-modal Attention Congruence Regularization for Vision-Language Relation Alignment
斯坦福大学	WHEN AND WHY VISION-LANGUAGE MODELS BE-HAVE LIKE BAGS-OF-WORDS, AND WHAT TO DOABOUT IT?
	CREPE: Can Vision-Language Foundation Models Reason Compositionally?
浙江大学	VL-CheckList: Evaluating Pre-trained Vision-Language Models with Objects, Attributes and Relations
华盛顿大学	SugarCrepe- Fixing Hackable Benchmarks for Vision-Language Compositionality
马里兰大学	Augmenting CLIP with Improved Visio-Linguistic Reasoning
华中科技大学	Structure-CLIP: Enhance Multi-modal Language Representations with Structure Knowledge
德克萨斯大学	Why is Winoground Hard? Investigating Failures in Visuolinguistic Compositionality
香港科技大学	An Examination of the Compositionality of Large Generative Vision-Language Models

任务介绍

目前的研究方向

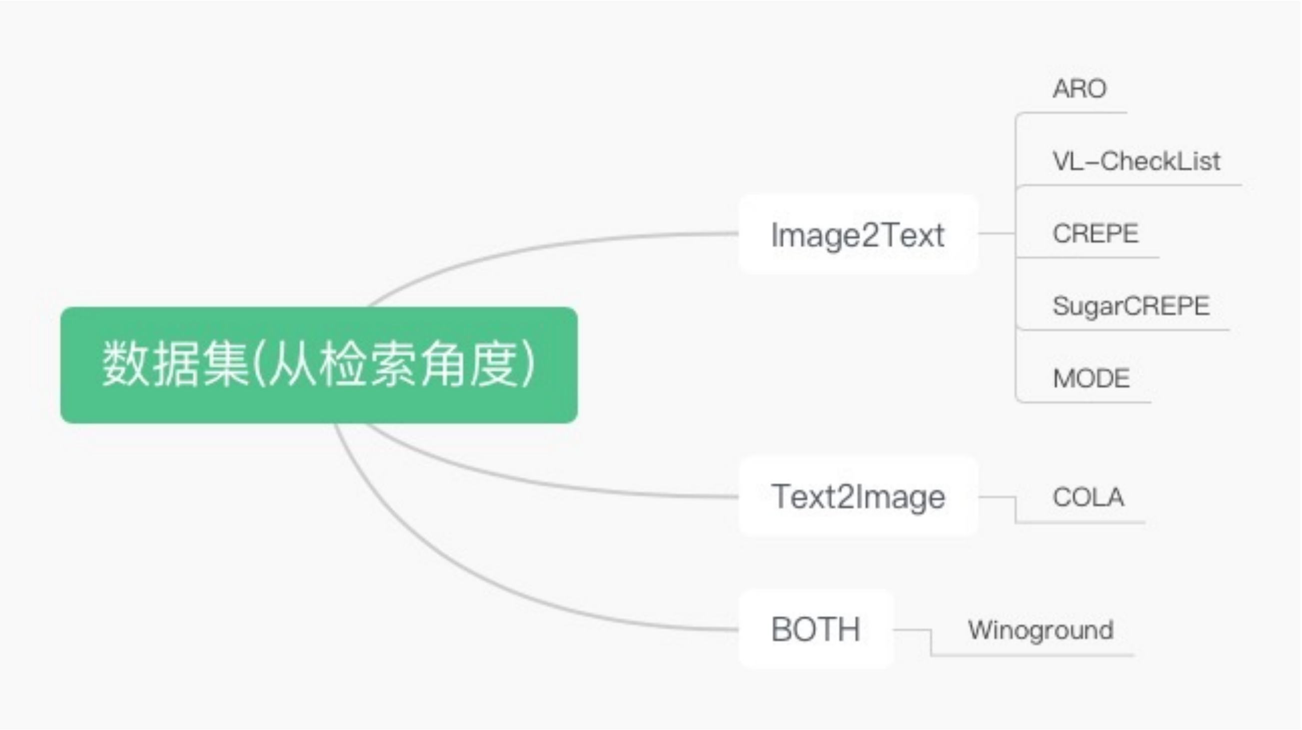


Baseline以CLIP为主

数据集以ARO、Winoground、VL-CheckList为主

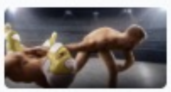
- 任务介绍
- 主流数据集
- 主流方法
- 破旧立新
- 总结反思

主流数据集



数据集(从检索角度)
Image2Text
ARO
VL-CheckList
CREPE
SugarCREPE
MODE
Text2Image
COLA
BOTH
Winoground

主流数据集-Winoground

		an old person kisses a young person	a young person kisses an old person
		the taller person hugs the shorter person	the shorter person hugs the taller person
		the masked wrestler hits the unmasked wrestler	the unmasked wrestler hits the masked wrestler
		a person watches an animal	an animal watches a person
		the person without earrings pays the person with…	the person with earrings pays the person without…
		a bird eats a snake	a snake eats a bird

包含两对图文对。文本包含的单词完全一致但语义有所差别。

最早提出了组合理解的概念，数据集质量高但规模小，仅包含400个样本。

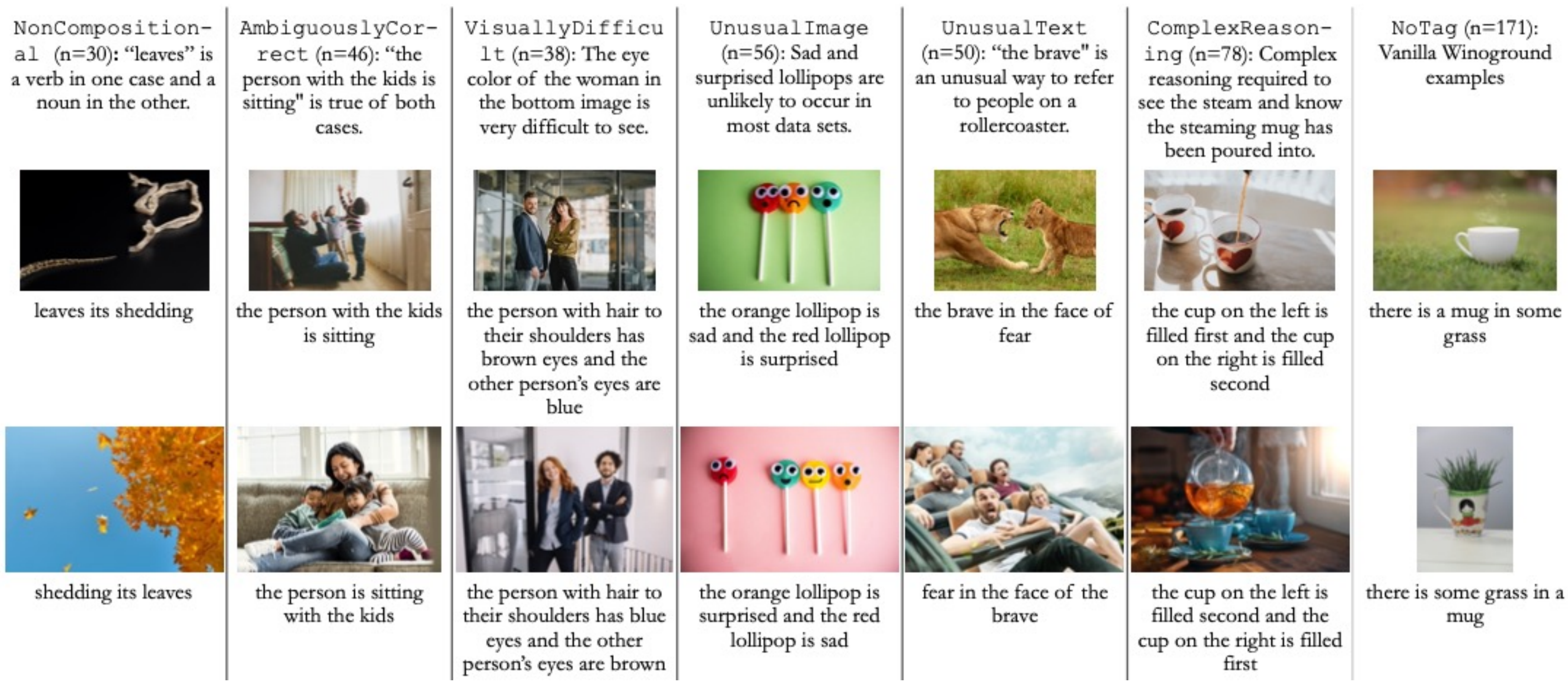
链接：https://huggingface.co/datasets/facebook/winoground/viewer/default/test

论文：Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality

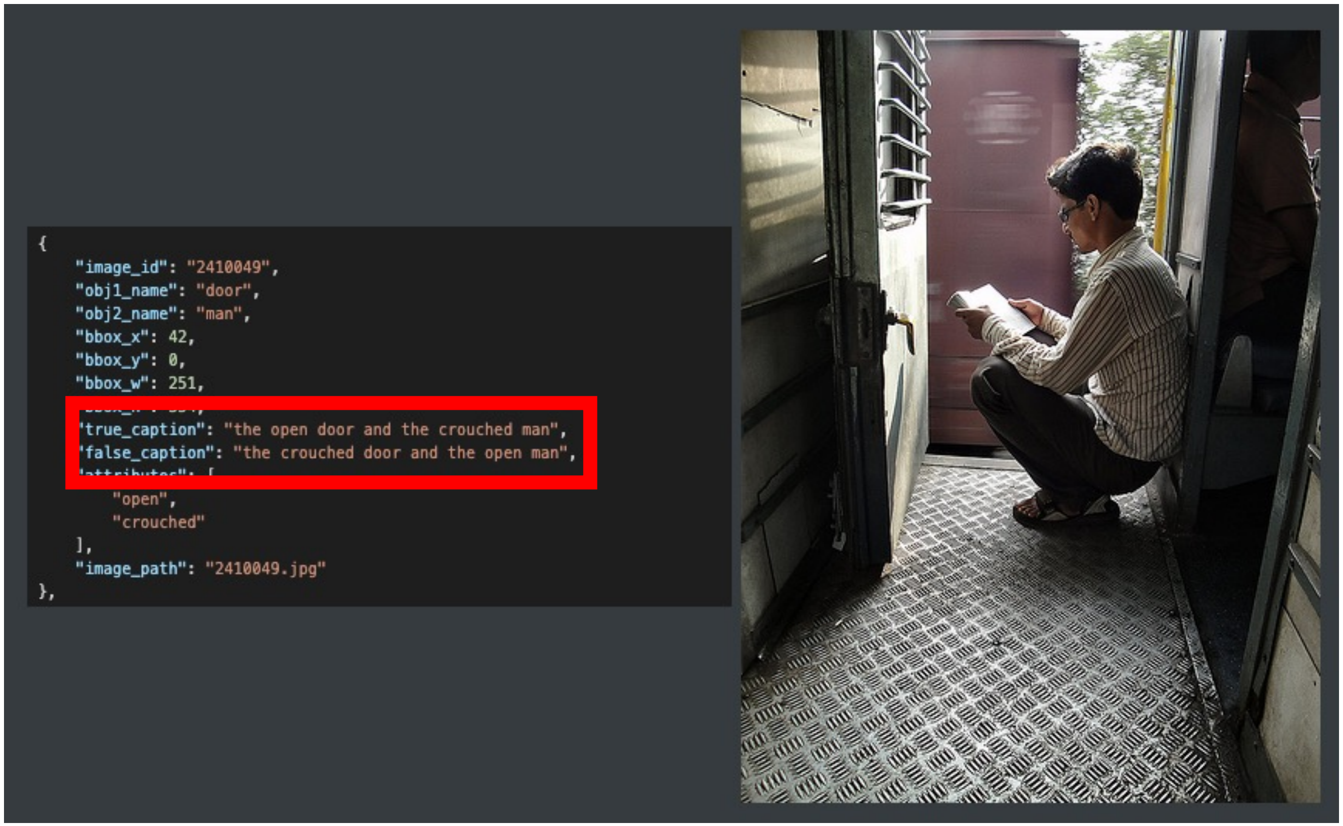
主流数据集-Winoground

“要在Winoground上取得好的performance不仅需要组合理解能力，还需要其他能力。”

-- 《Why is Winoground hard》



主流数据集-VGA&VGR



探究$(attribute, object)$的绑定关系

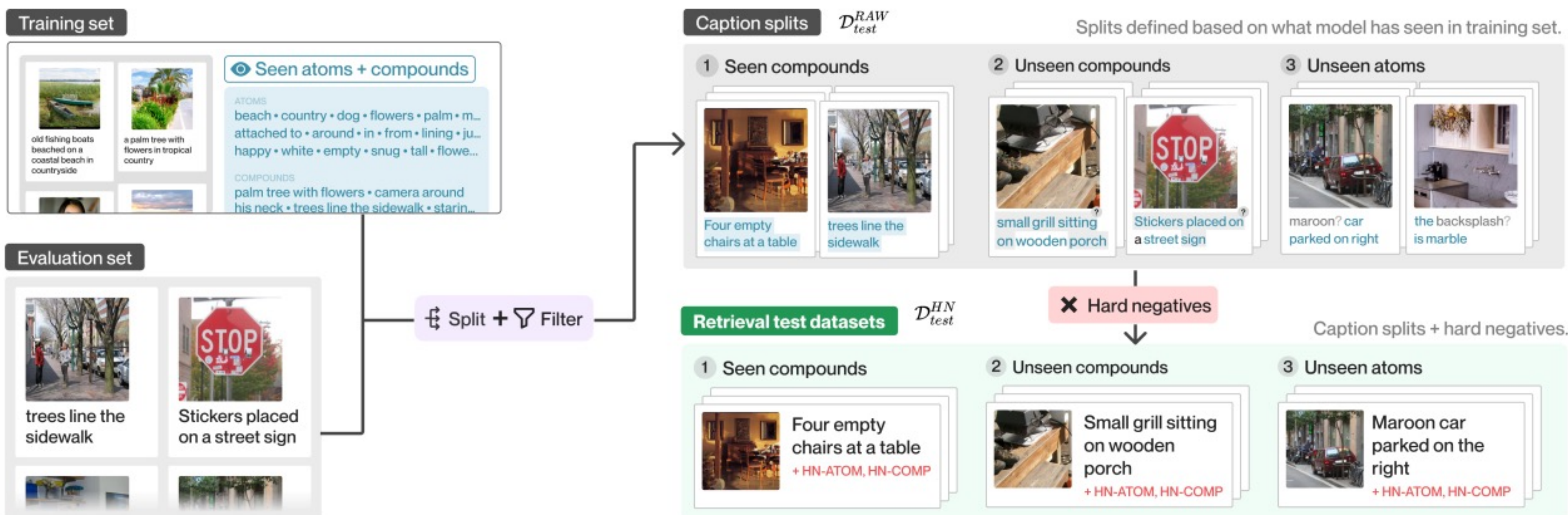
主流数据集-VGA&VGR

探究($object,\ relationship,\ object$)的绑定关系

主流数据集-CREPE

衡量角度：
Systematicity：对于“原子和化合物之间的组合关系”的理解程度
Productivity：对于复杂表述的理解程度。

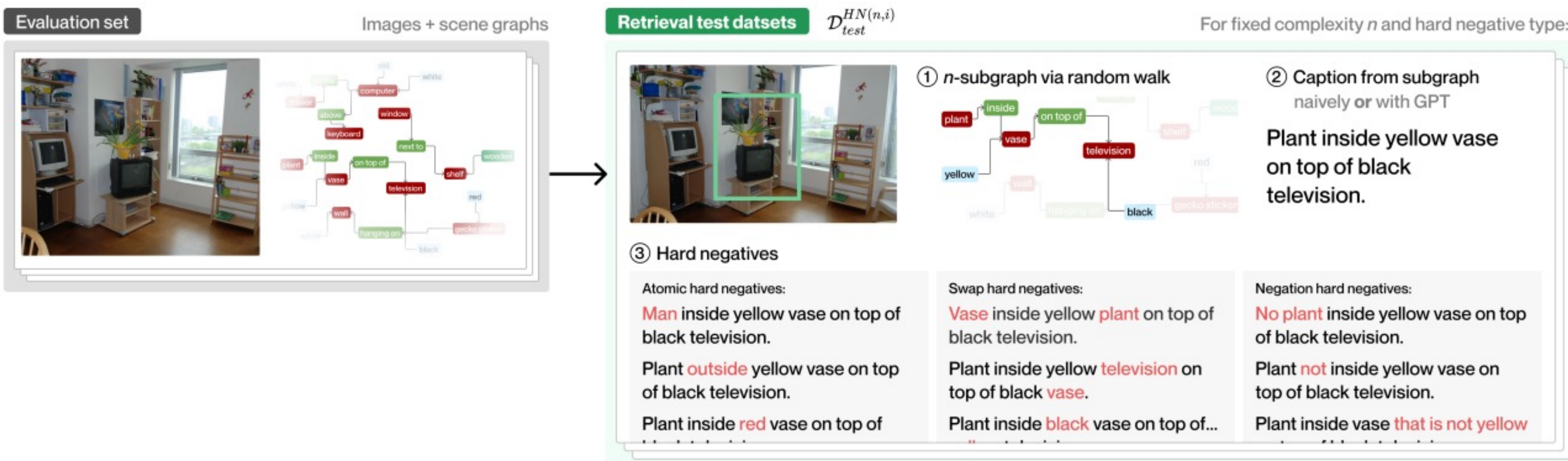
(O,R,O) (A,O)



SC：化合物都见过　UC：原子见过，化合物没见过　UA：存在没见过的原子

HN-COMP: a pink car → a blue car and a pink bird　度量对于化合物的理解程度

主流数据集-CREPE



在VG提供的场景图中执行随机游走，得到n个子图（Object），然后将相关的对象、属性、关系进行简单连接或者用GPT生成文本。

1. 随机替换：对应模型会忽略单个原子内容
2. 交换同类型原子：对应模型会将caption视为词袋；
3. 随机否定：对应模型是否理解否定的含义，将属性与目标或者目标与目标的绑定关系解除。

- 任务介绍
- 主流数据集
- 主流方法
- 破旧立新
- 总结反思

主流方法

- 组合理解优化方法
  - 负样本增强
    - WHEN AND WHY VISION–LANGUAGE MODELS BE–HAVE LIKE BAGS–OF–WORDS, AND WHAT TO DOABOUT IT?
    - Contrasting Intra–Modal and Ranking Cross–Modal Hard Negatives to Enhance Visio–Linguistic Fine–grained Understanding
    - Teaching Structured Vision & Language Concepts to Vision & Language Models
  - 数据合成
    - Dense and Aligned Captions (DAC) Promote Compositional Reasoning in VL Models
    - Going Beyond Nouns With Vision & Language Models Using Synthetic Data
  - 场景图辅助
    - Incorporating Structured Representations into Pretrained Vision & Language Models Using Scene Graphs
    - Coarse–to–Fine Contrastive Learning in Image–Text–Graph Space for Improved Vision–Language Compositionality
    - Structure–CLIP: Enhance Multi–modal Language Representations with Structure Knowledge
  - 生成式模型
    - VisualGPTScore: Visio–Linguistic Reasoning with Multimodal Generative Pre–Training Scores
    - Simple Token–Level Confidence Improves Caption Correctness
  - 其他
    - 知识蒸馏: Augmenting CLIP with Improved Visio–Linguistic Reasoning
    - 问答任务辅助: What You See is What You Read? Improving Text–Image Alignment Evaluation
    - 跨模态融合: Cola: How to adapt vision–language models to Compose Objects Localized with Attributes?

主流方法-负样本增强

方法：通过各种方式操纵文本，生成负样本。提出针对负样本的损失函数。

初衷：CLIP之所以将caption当做词袋的主要原因是ITC Loss与低质量负样本的耦合带来的捷径效应。

Apple is red. Bird eats snake. ……

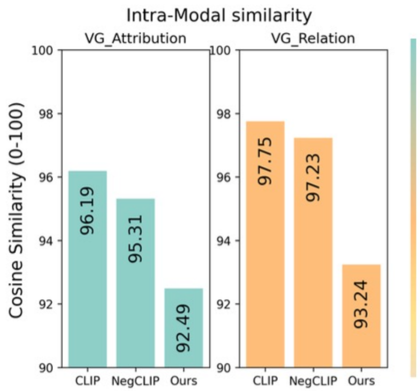
ITC Loss只需比较相似度且正负样本之间几乎不存在相似的原子，甚至化合物。因此，仅关注显著原子即可。

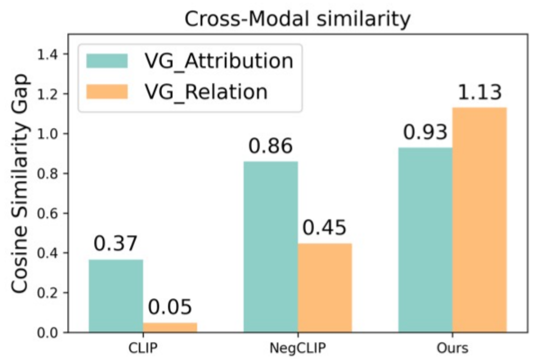
1. TSVLC：LLM随机生成单词进行替换（动词、名词、形容词）；
2. NegCLIP：按照规则扰动（交换属性、交换目标、调整词序等）；

上述方法的损失函数对负样本的利用程度不高。 I2T Loss的分母；T2I Loss不使用。

主流方法-负样本增强

Contrasting Intra-Modal and Ranking Cross-Modal Hard Negatives to Enhance Visio-Linguistic Fine-grained Understanding



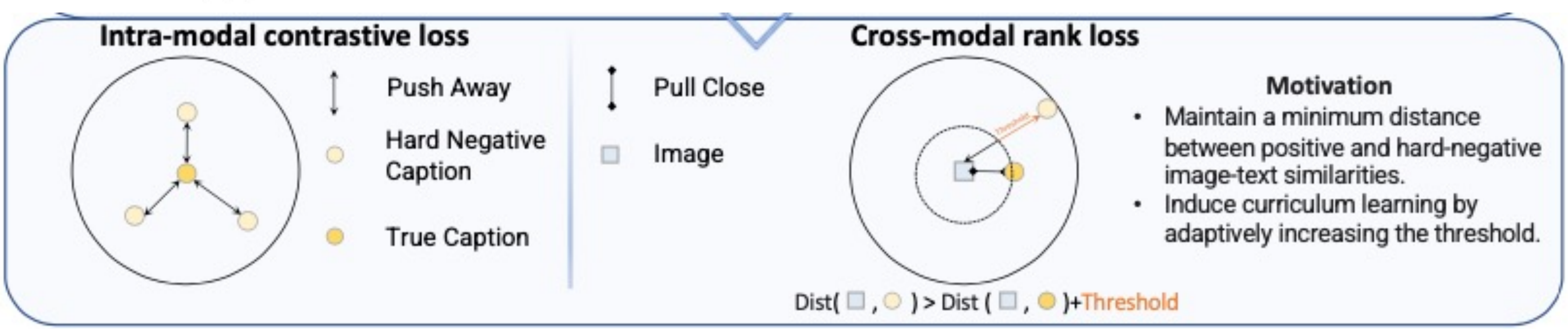


$$\mathcal{L}_{imc} = \sum_{(I,T)\in\mathcal{B}} -\log \frac{\exp^{S(I,T)}}{\boxed{\sum_{T_k\in\mathcal{T}_{hn}} \exp^{S(T,T_k)}}}$$

模态内的对比损失：降低true caption和false caption的相似度

$$\mathcal{L}_{cmr} = \sum_{(I,T)\in\mathcal{B}} \sum_{T_k\in\mathcal{T}_{hn}} max(0, S(I,T_k) - S(I,T) + Th_k)$$

模态间的排序损失：true pair的相似度大于false pair的相似度



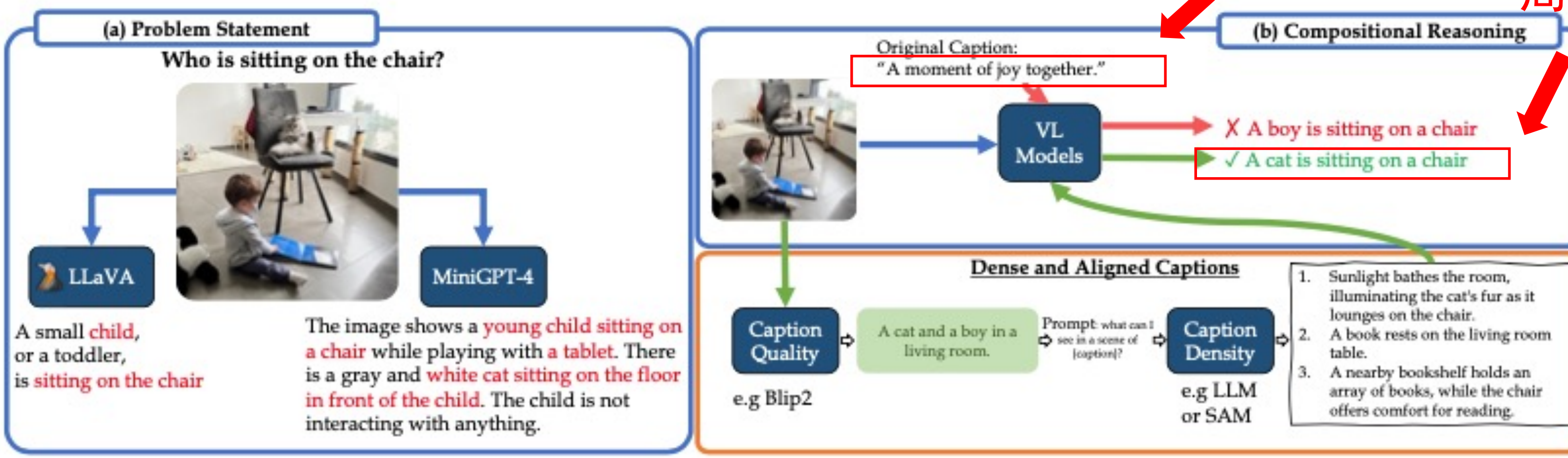
主流方法-合成数据

Dense and Aligned Captions (DAC) Promote Compositional Reasoning in VL Models

初衷：预训练数据集中，文本部分的质量和密度存在问题。

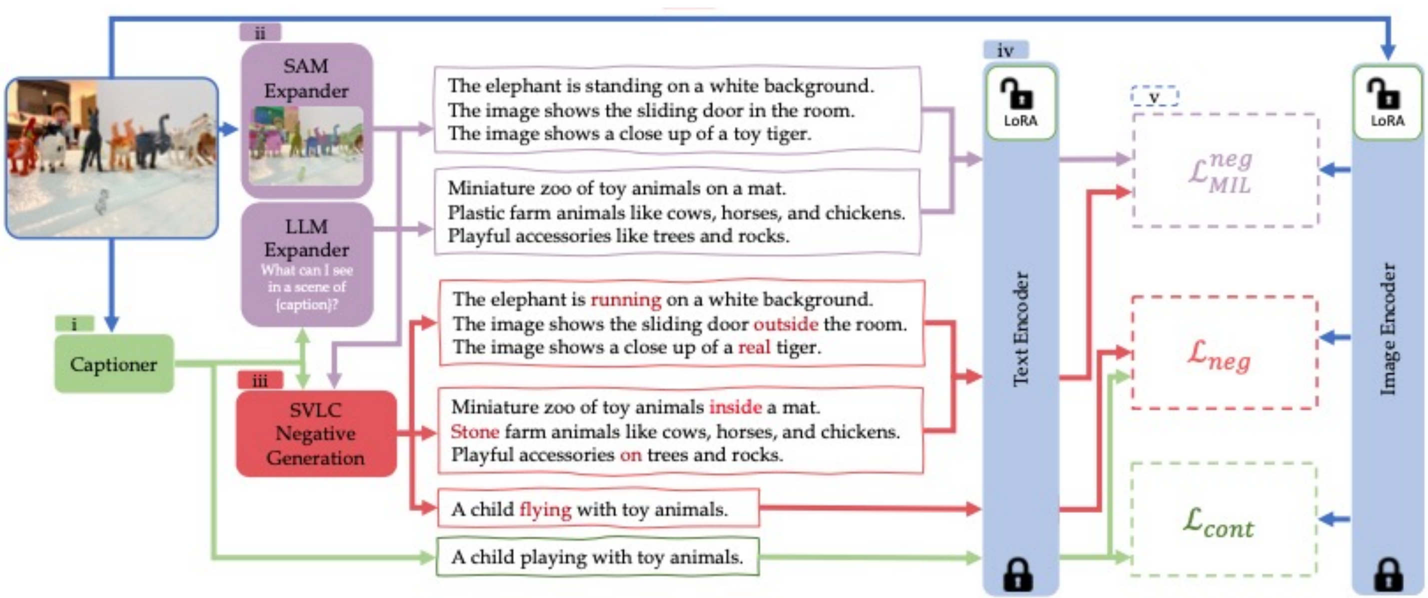
图文无关

局部相关



质量缺陷：错误匹配，负优化；
密度缺陷：抑制视觉端提取的特征。

主流方法-合成数据



Captioner：BLIP2 captioner 根据图片生成文本

Expander：
LLM：GPT，运用prompt：What can I see in a scene of [caption]？扩充文本；
SAM：得到许多图块，应用captioner生成文本。

亮点：只需要图像即可完成跨模态训练任务。

主流方法-合成数据

LLM有很强的推理能力，但是推理的结果会有较大的偏差。同理，SAM存在over-segmentation的情况，会带来噪声的分割结果。

本文采用了弱约束的损失函数来对抗噪声：

1. 将LLM/SAM生成的文本分割成M个部分，每一部分为$T_{i,m}$；
2. 采用Multiple Instance Learning的损失函数：

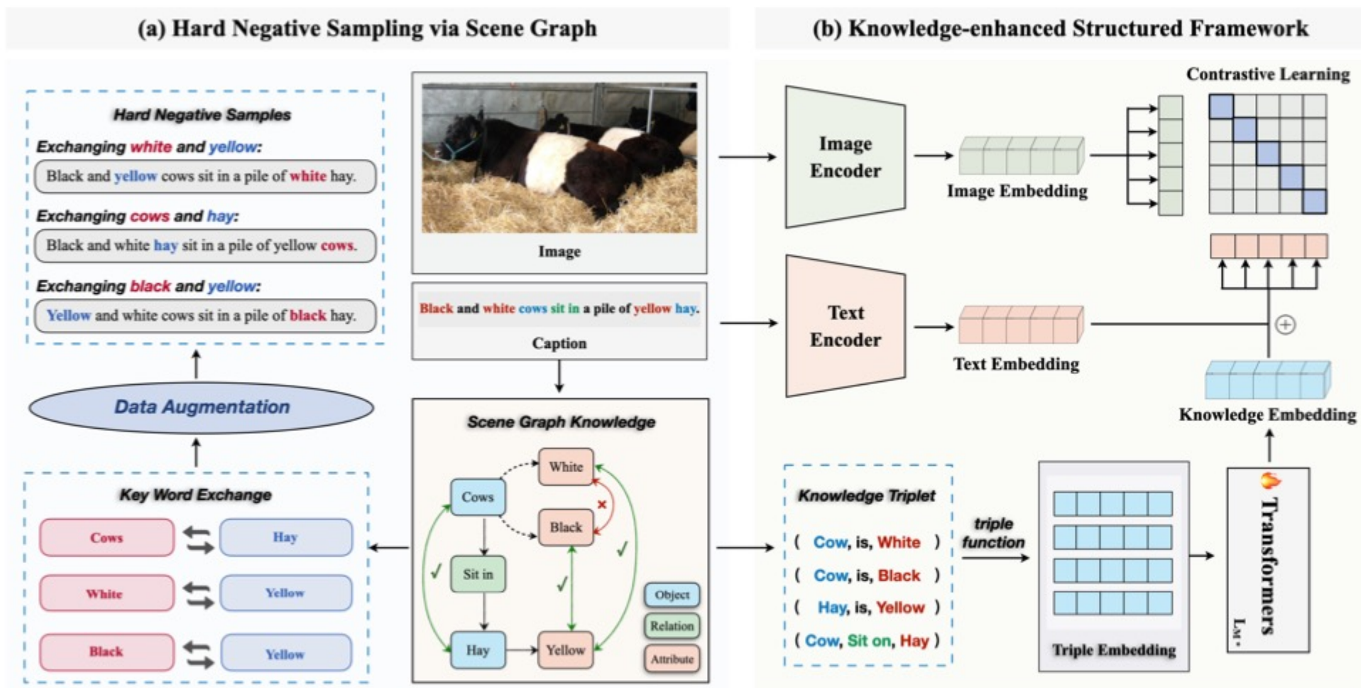
$$\mathcal{L}_{MIL}^{neg} = -\frac{1}{B}\sum_{i}^{B} \log \frac{\sum_m S(T_{i,m}, I_i)}{\left(\sum_m S(T_{i,m}^{neg}, I_i)\right) + \left(\sum_{j=1}^{B}\sum_m S(T_{j,m}, I_i)\right)}$$

多实例学习的样本单位为bag，即图中的钥匙串。
每个钥匙串包含多个钥匙，即实例。
任务的目标是预测哪个钥匙串包含能开门的钥匙？（至少含一个正例）

MIL的提出是为了在制药领域判断分子是否包含有效部分。

主流方法-场景图辅助

Structure-CLIP: Enhance Multi-modal Language Representations with Structure Knowledge



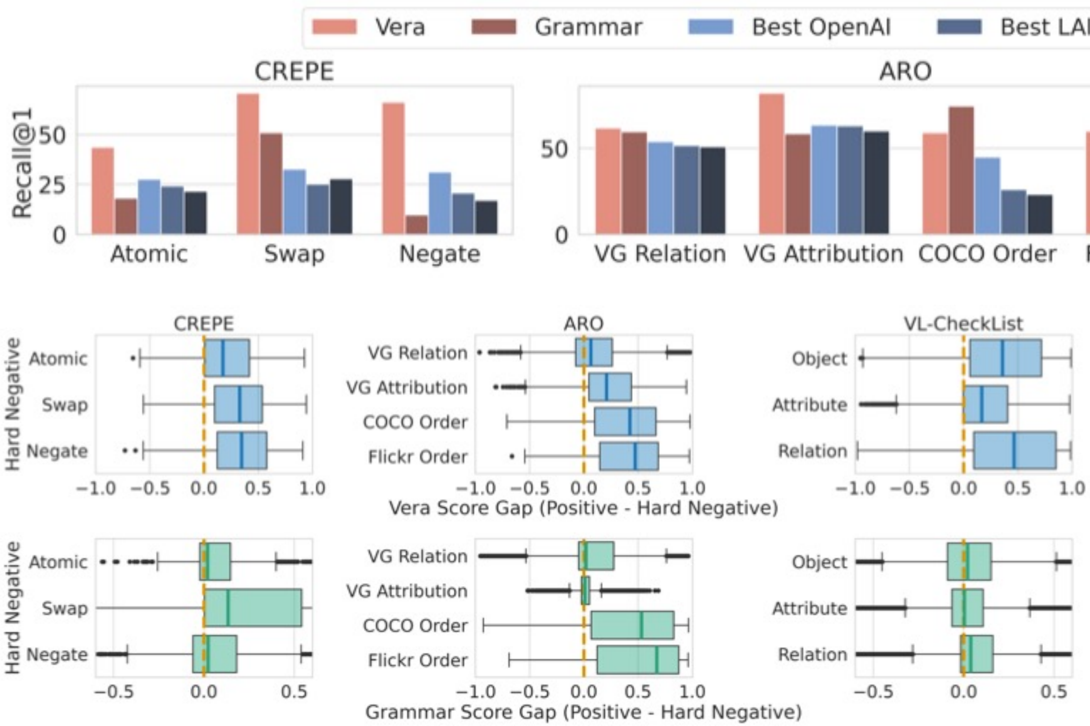
1. 用场景图规范负样本的生成：针对形容词的无效交换。
2. 用场景图提取结构信息，与文本端的输出融合。

Embedding fusion	VG-Attribution	VG-Relation
Concat	81.1	83.3
head + tail	81.3	83.1
head + relation + tail	81.9	83.3
head + relation - tail	**82.3**	**84.7**

- 任务介绍
- 主流数据集
- 主流方法
- 破旧立新
- 总结反思

破旧立新

SugarCrepe- Fixing Hackable Benchmarks for Vision-Language Compositionality



Vera Score：评估文本的合理程度，越合理分越高；Grammar Score：评估文本的流程程度，越流畅分越高。

显而易见，评估数据集中的负样本相较于正样本既不合理也不流畅。也就是说，现有的大部分数据集提供了取得高分的捷径。现有的数据驱动类的方法也存在缺陷。

破旧立新

Step-1：应用ChatGPT按照：单词定位、单词生成、句子生成的步骤生成合理负样本。

```
Given an input sentence describing a scene, your task
is to:
1. Locate the noun words in the sentence.
2. Randomly pick one noun word.
3. Replace the selected noun word with a new noun word
to make a new sentence.

The new sentence must meet the following three
requirements:
1. The new sentence must be describing a scene that is
as different as possible from the original scene.
2. The new sentence must be fluent and grammatically
correct.
3. The new sentence must make logical sense.

Here are some examples:

Original sentence: A man is in a kitchen making pizzas.
Nouns: ["man", "kitchen", "pizzas"]
Selected noun: man
New noun: woman
New sentence: A woman is in a kitchen making pizzas.

Original sentence: a woman seated on wall and birds
besides her
Nouns: ['woman', 'wall', 'birds']
Selected noun: wall
New noun: bench
New sentence: A woman seated on a bench and birds
besides her.
```

(a) REPLACE-OBJ.

```
Given an input sentence describing a scene, your task
is to first locate two swappable noun phrases in the
sentence, and then swap them to make a new sentence.
The new sentence must meet the following three
requirements:
1. The new sentence must be describing a different
scene from the input sentence.
2. The new sentence must be fluent and grammatically
correct.
3. The new sentence must make logical sense.

To complete the task, you should:
1. Answer the question of whether generating such a new
sentence is possible using Yes or No.
2. Output the swappable noun phrases.
3. Swap them to make a new sentence.

Here are some examples:

Input: A cat resting on a laptop next to a person.
Is it possible to swap noun phrases in the input
sentence to generate a new sentence that is different
from the input sentence and makes logical sense? Yes
Swappable noun phrases: laptop, person
Output: A cat resting on a person next to a laptop.

Input: A plate of donuts with a person in the
background.
Is it possible to swap noun phrases in the input
sentence to generate a new sentence that is different
from the input sentence and makes logical sense? Yes
Swappable noun phrases: a plate of donuts, a person
Output: A person with a plate of donuts in the
background.
```

(a) SWAP-OBJ.

```
Given an input sentence describing a scene, your task
is:
1. Find the objects in the sentence.
2. Randomly pick one object.
3. Generate a new object that's not in the sentence.
4. Add the new object next to the selected object to
make a new sentence.

The new sentence must meet the following three
requirements:
1. The new sentence must describe a clearly new and
different scene.
2. The new sentence must be fluent and grammatically
correct.
3. The new sentence must make logical sense.

Here are some examples:

Original sentence: An elephant standing under the shade
of a tree.
Objects: ["elephant", "shade of a tree"]
Selected object: elephant
New object: squirrel
New sentence: An elephant and a squirrel standing under
the shade of a tree.

Original sentence: A bench at the beach next to the sea
Objects: ['bench', 'beach', 'sea']
Selected object: bench
New object: umbrella
New sentence: An umbrella and a bench at the beach next
to the sea.
```

(a) ADD-OBJ.

Step-2：人工筛选与采样。

- 任务介绍
- 主流数据
- 主流方法
- 破旧立新
- 总结反思

总结反思

1. 优质数据能够使预训练模型带来蜕变，即数据驱动是微调的核心；但是！[1]中强调了损失函数 >> 优质数据。
2. 组合理解任务的最终归宿是指导跨模态预训练模型的训练；
3. LLM在构造数据方面得到了一定程度的探索和运用（数据集评估、数据集生成、负样本构造、正样本构造等）。
4. 我个人认为：组合理解任务的解决方案需要考虑两个要素：首先，文本编码器和视觉编码器都能全面地表征输入；其次，模态间需要得到很好地对齐。
5. 已经有文章（SugarCREPE、MODE）证明目前的数据驱动方案、Benchmark存在漏洞，该领域依然处于探索阶段。

[1] Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts