



I can't believe there's no images!
Learning Visual Tasks Using Only Language Supervision

Sophia Gu* Christopher Clark* Aniruddha Kembhavi
Allen Institute for Artificial Intelligence
{sophiag, chrisc, anik}@allenai.org

VLM, modality gap, text as image

曹耘宁
2023/9/12

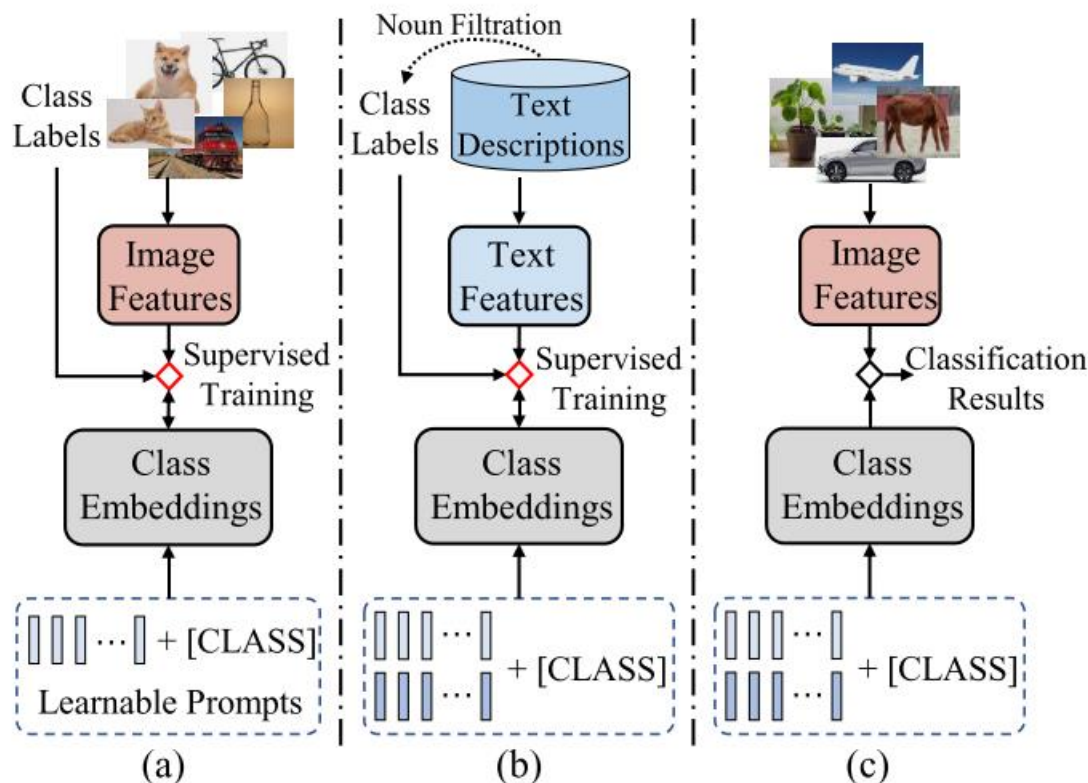


- 研究背景
- 研究方法
- 实验效果
- 总结

Text as image

3

- 在视觉语言模型中，用文字代替图像进行训练
 - ⊙ 收集训练集的成本低
 - ⊙ Cross modality transfer





语料生成

4

- 成对：
 - ⊙ 现有caption数据集
 - ⊙ Label+LLM扩写

- 扩展语料库（不一定成对）
 - ⊙ 网络数据
 - ⊙ 书籍
 - ⊙ LLM生成

Modality Gap

5

- 两种模态间存在gap
 - ⊙ 模态间共有的信息 → 对齐
 - ⊙ 模态内独特的信息 → 不能强迫对齐

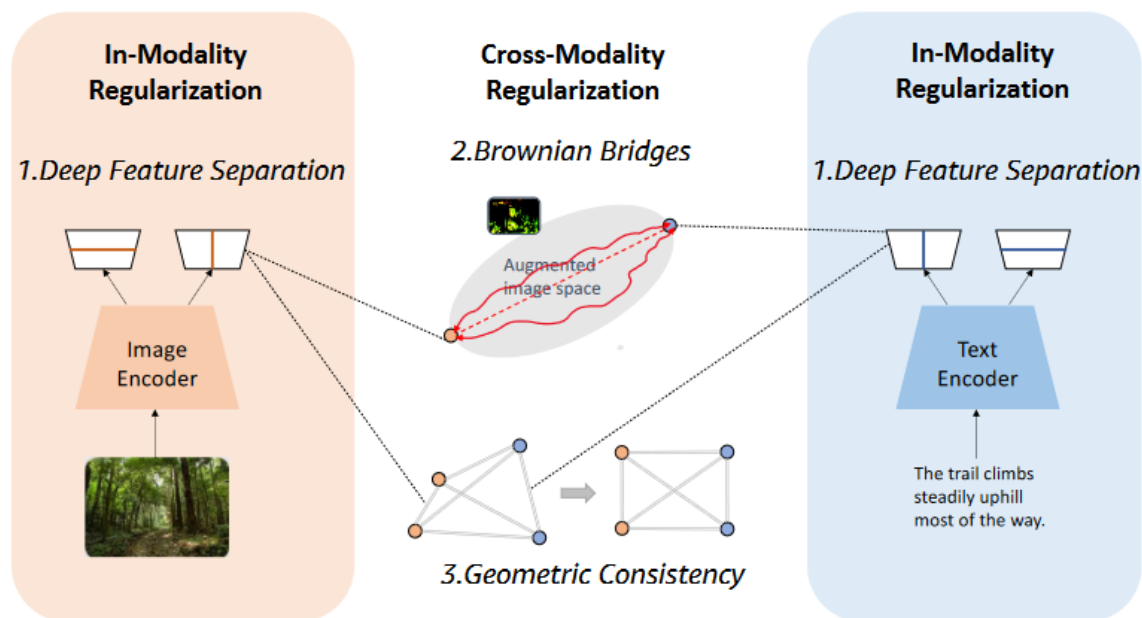


Figure 1. Constructing latent modality structures to improve multi-modal representation learning.



- 研究背景
- 研究方法
- 实验效果
- 总结

Method

7

总体框架:

- 训练时采用text encoder，测试时替换成image encoder
- 可以用于各种多模态任务：VQA, caption, visual entailment, visual news

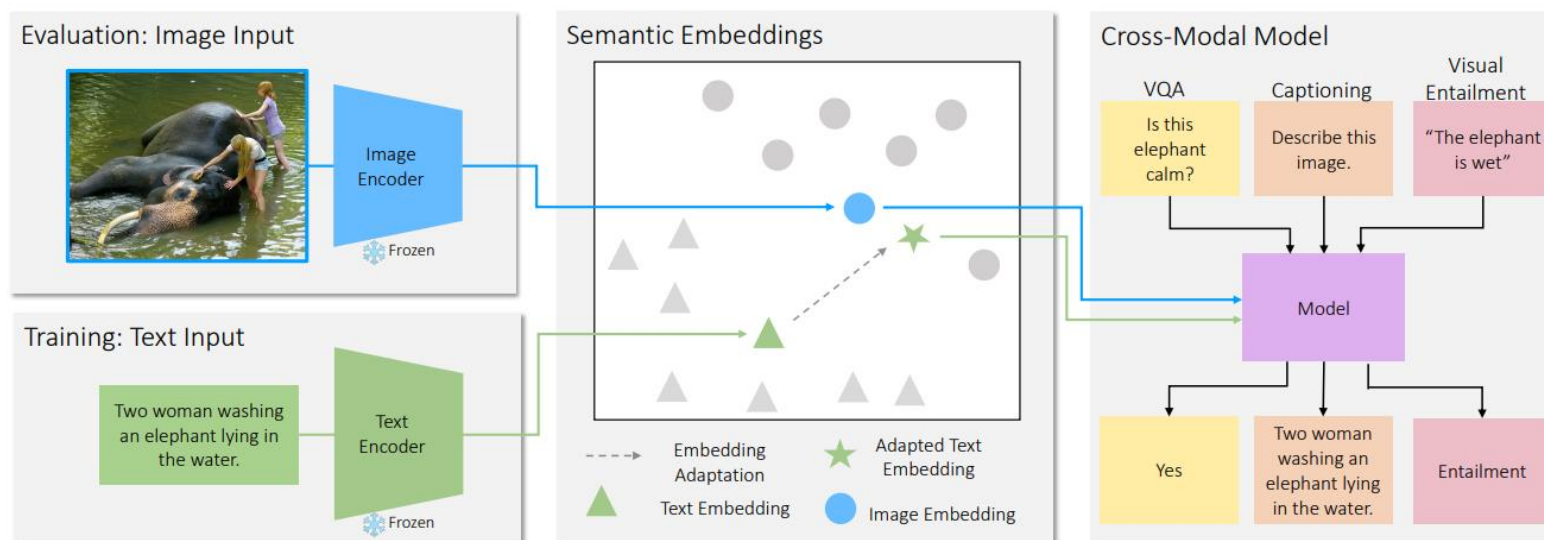


Figure 1: Overview of CLOSE. During training, input text is encoded into a vector with a text encoder and adapted with an adaptation method. A model learns to use the vector to perform a task such as VQA, captioning, or visual entailment. During testing, an input image is encoded with an image encoder instead to allow cross-modal transfer.

语料获取

8

- 不同风格的训练语料：
 - ⊙ 网络数据，评论
 - ⊙ 大语言模型，第一人称描述
 - ⊙ 书籍，人物



语料获取

9

- 大语言模型：
 - ⊙ Instruction tuning的方式获取高质量caption

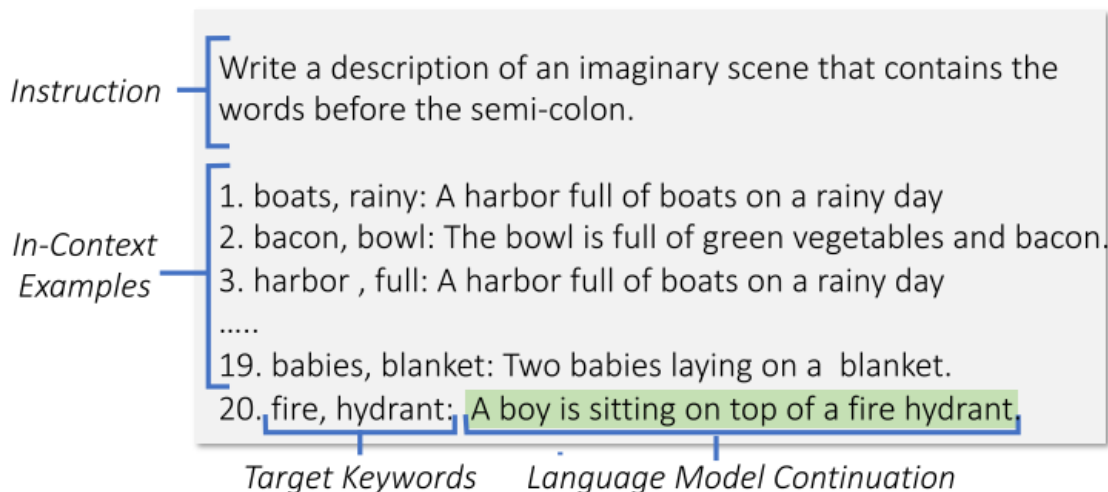


Figure 4: Prompt used to generate a synthetic caption from a language model. The language model's continuation (high-lighted text) is used as a synthetic caption.

Egocentric Captions



I saw a bird perched on a sand beach looking at the ocean.

We walked past a kitchen with a window looking out onto a street.

We are flying kites in a park.

My mom is making pancakes in the kitchen.

We visited an old building with a bicycle leaning against it, next to a brick wall.

We are playing a video game with controllers in our hands.

Uplifting Captions



A group of people are sitting around a table enjoying pizza and laughter.

A flock of birds fly overhead as the sun sets in the horizon.

Two girls sitting on the back of a boat contemplating life's mysteries.

A bunch of stuff animals on a train journey to reach home.

A beautiful purple tulip flower pot is in bloom.

A man skiing down a snowy hill to conquer the high.

Harry Potter Captions



Harry Potter was so excited to start his first year at Hogwarts!

Gellert Grindelwald looked around at the assembled students and smiled.

Lucius Malfoy watched with satisfaction as the death eaters gathered around Harry Potter.

Delores Umbridge sat at her desk, a satisfied smile on her face.

Rubeus Hagrid roared with laughter as he saw the look of terror on Harry Potter's face.

Lord Voldemort laughed softly, a cold sound that made the hairs on the back of Harry Potter's neck shiver.

Reviews Captions



A perfect gift for a friend who has a flower garden. The roses are beautiful.

The leash is well made and easy to put on and take off. My dog is very happy with it.

This was a wedding cake for my husband and he loved it. He was very happy with the cake.

This is a great oven. It cooks evenly and is easy to clean. I would recommend it.

I bought this as a gift for a friend. She loves it. It is very soft and cuddly.

Fast delivery: I received my order in a timely manner and it was in good condition. I would order from them again.

Figure 5: Examples of stylistic captions produced by CLOSE trained with only text data, and then applied 0-shot to images.

Modality Gap

11

□ Modality Gap

- 在CoCo中，图文对的平均相似度为0.26，两个不相关caption的平均相似度为0.35
- 因为对比学习损失只要求匹配的图文对相似度相对于随机图文对的更高即可，不约束相似度绝对值的大小。

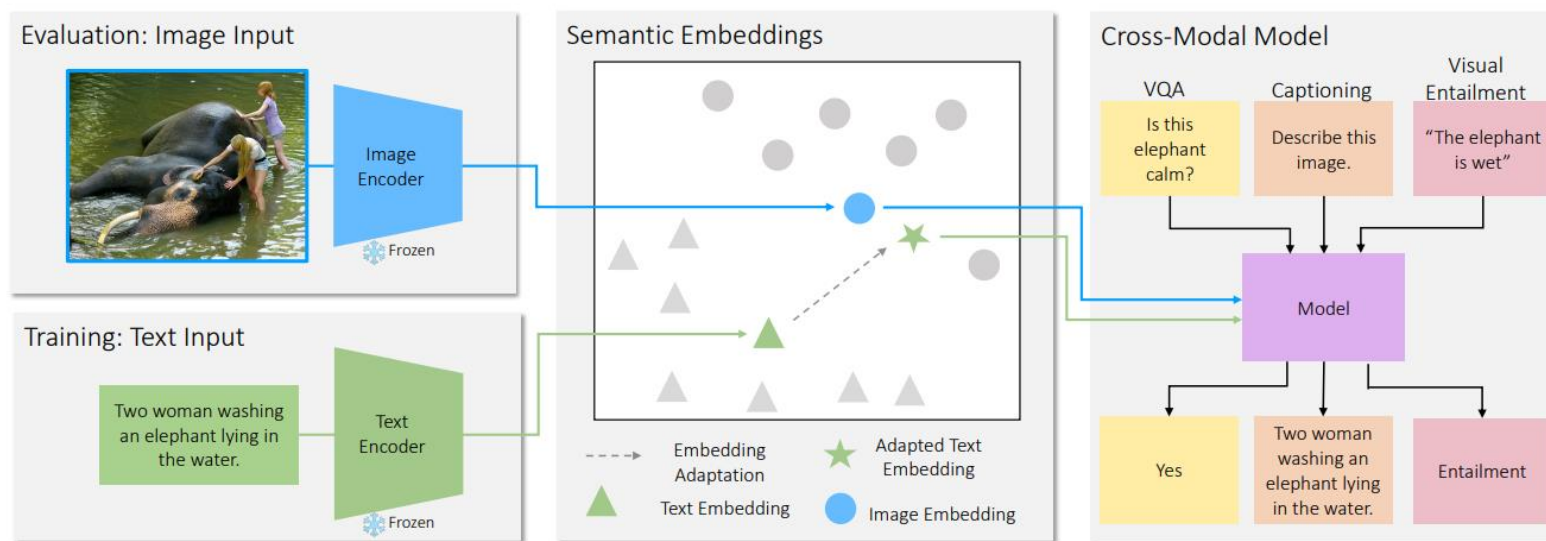


Figure 1: Overview of CLOSE. During training, input text is encoded into a vector with a text encoder and adapted with an adaptation method. A model learns to use the vector to perform a task such as VQA, captioning, or visual entailment. During testing, an input image is encoded with an image encoder instead to allow cross-modal transfer.

Modality Gap

12

- 缓解modality gap的方法：在文本向量上添加高斯噪声

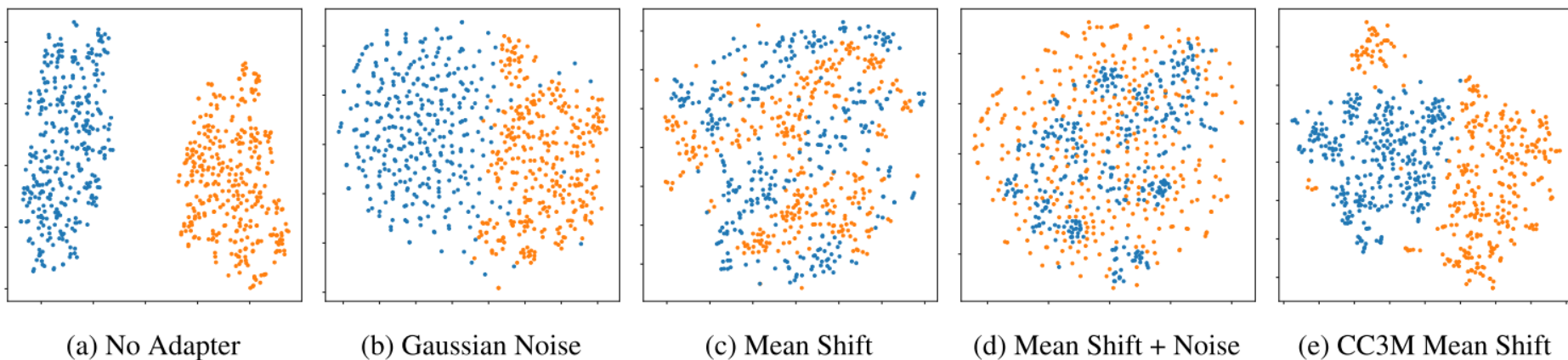


Figure 3: t-SNE [65] plots for various adapters on 350 randomly selected image vectors (blue) and paired caption vectors (orange) from COCO captions. The first two panels demonstrate CLOSE, and the remaining three show additional adapters we study in our analysis (Section 4).



Modality Gap分析

13

□ text vector的敏感性分析

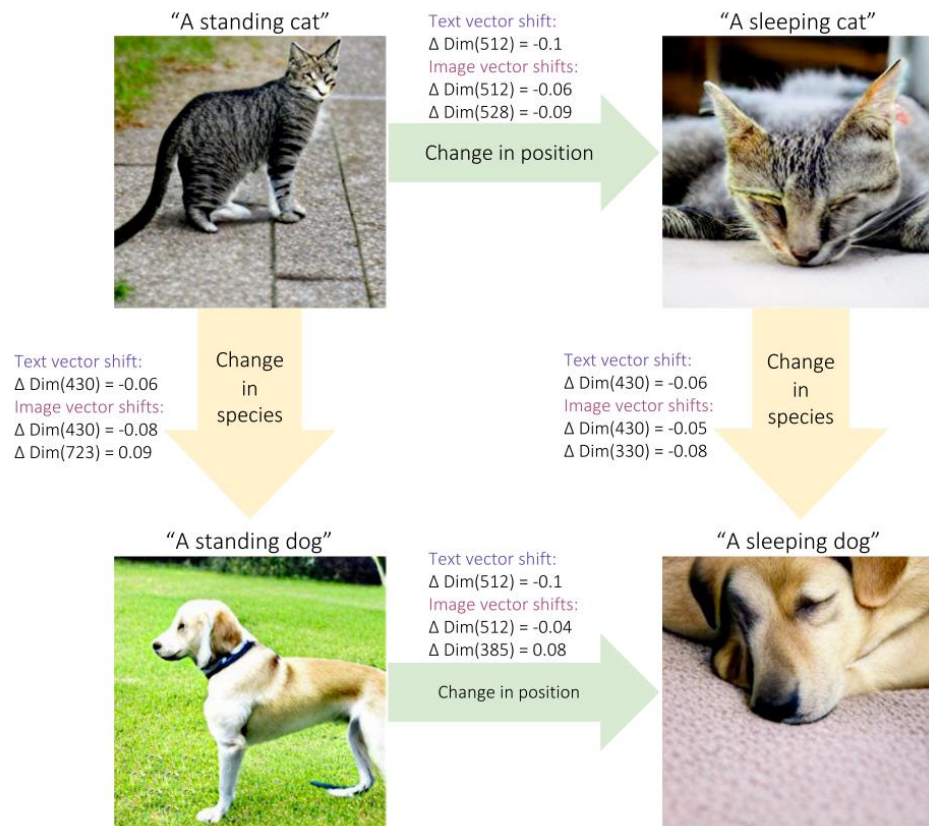
- ⊙ 平移量, modality gap, 与原向量的相似度

Bias	Mag.	MG	Δ	Cap.	VE	VQA
none	0.0	0.26	1.00	94.4	64.3	75.9
mean	0.8	0.62	0.69	92.8	64.7	75.4
-mean	0.8	-0.10	0.85	84.3	62.0	71.8
RNG	0.2	0.25	0.98	93.5	63.9	75.3
RNG	0.5	0.24	0.89	92.5	64.2	75.3
RNG	0.8	0.20	0.78	89.3	63.7	74.8
RNG	1.0	0.18	0.71	87.2	63.8	74.2
RNG	2.0	0.11	0.45	73.7	61.4	71.3

Modality Gap分析

14

- Text, image vector对偏移的敏感度不同
 - ⊙ 因为图像中细节信息更多，而文本语义不易改变





- 研究背景
- 研究方法
- 实验效果
- 后续工作
- 总结



实验效果

16

□ 在不同任务上的性能

Model	Text-Only	Cap. (Single)	Cap. (Mult.)	VE	VQA	E-VQA	VN
Prior Work	✓	-	ESPER Style [79] 78.2	CLIP Cls. [57] 66.6	TAP-C [57] 38.7	-	-
CLOSE w/o Noise	✓	16.4	68.7	68.2	60.2	59.8	32.1
CLOSE (Ours)	✓	80.5	95.3	75.9	59.6	62.9	80.8
CLOSE w/Tuned Noise		95.4	98.4	75.9	61.9	64.3	80.8
CLOSE w/Images		113.2	113.2	77.7	65.4	67.9	105.7

Table 1: Results on V&L tasks. Models in the last two rows require images and so are upper bounds for CLOSE. We report CIDEr [66] for captioning with single and multiple captions, visual entailment test accuracy, VQA 2.0 test-dev accuracy, E-VQA validation accuracy, visual news test CIDEr. See Appendix 2 for other metrics and more detailed results.



Ablation

17

□ 不同底座模型的消融实验

CLIP Model	T5 Model	Cap.	VE	VQA
ViT-L/14	small	94.4	74.9	59.9
ViT-L/14	base	95.4	76.1	64.3
ViT-L/14	large	93.9	75.1	65.2
ViT-B/32	base	91.1	75.3	61.4
RN101	base	90.0	75.4	59.8
RN50	base	90.2	75.3	60.4
RN50×4	base	92.0	75.3	61.5
RN50×16	base	93.4	74.4	62.5
RN50×64	base	96.1	75.8	64.2
OpenCLIP [24]	base	99.2	76.3	65.1
EVA-CLIP [13]	base	101.7	75.53	66.6



- 研究背景
- 研究方法
- 实验效果
- 总结

总结



19

- Text合成部分考虑了不同风格、连贯性，值得参考
- 模态间的gap/cross modality transfer是这类问题的关键，本文给出了详尽分析



Thanks!