HOCHSCHULE
HANNOVER
UNIVERSITY OF
APPLIED SCIENCES
AND ARTS
–
Fakultät IV
Wirtschaft und
Informatik

Leveraging Pretrained Unimodal Models for Efficient Vision-Language Pretraining

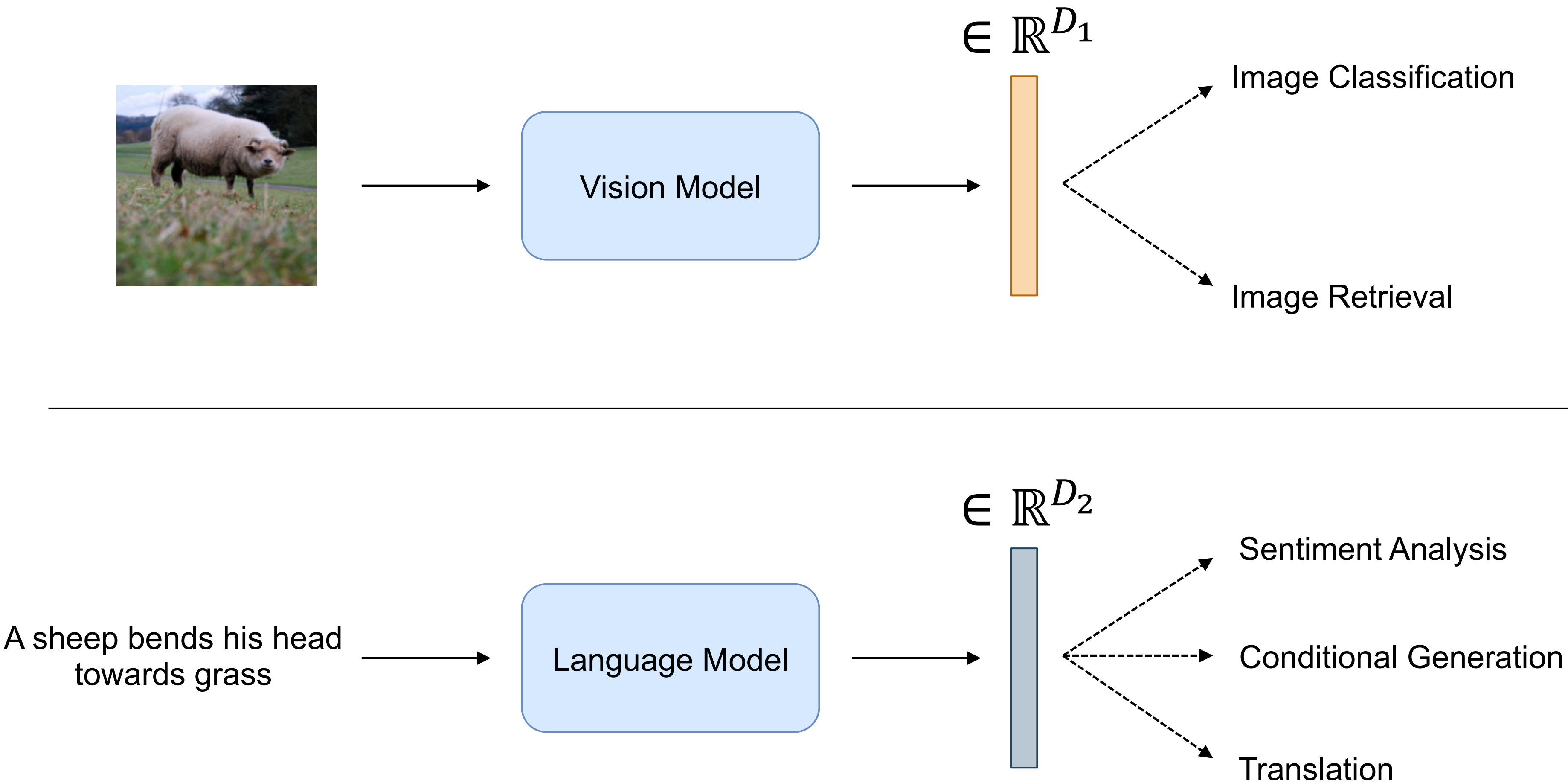
Master‘s thesis in Applied Computer Science

Tim Cares, 24.10.2024

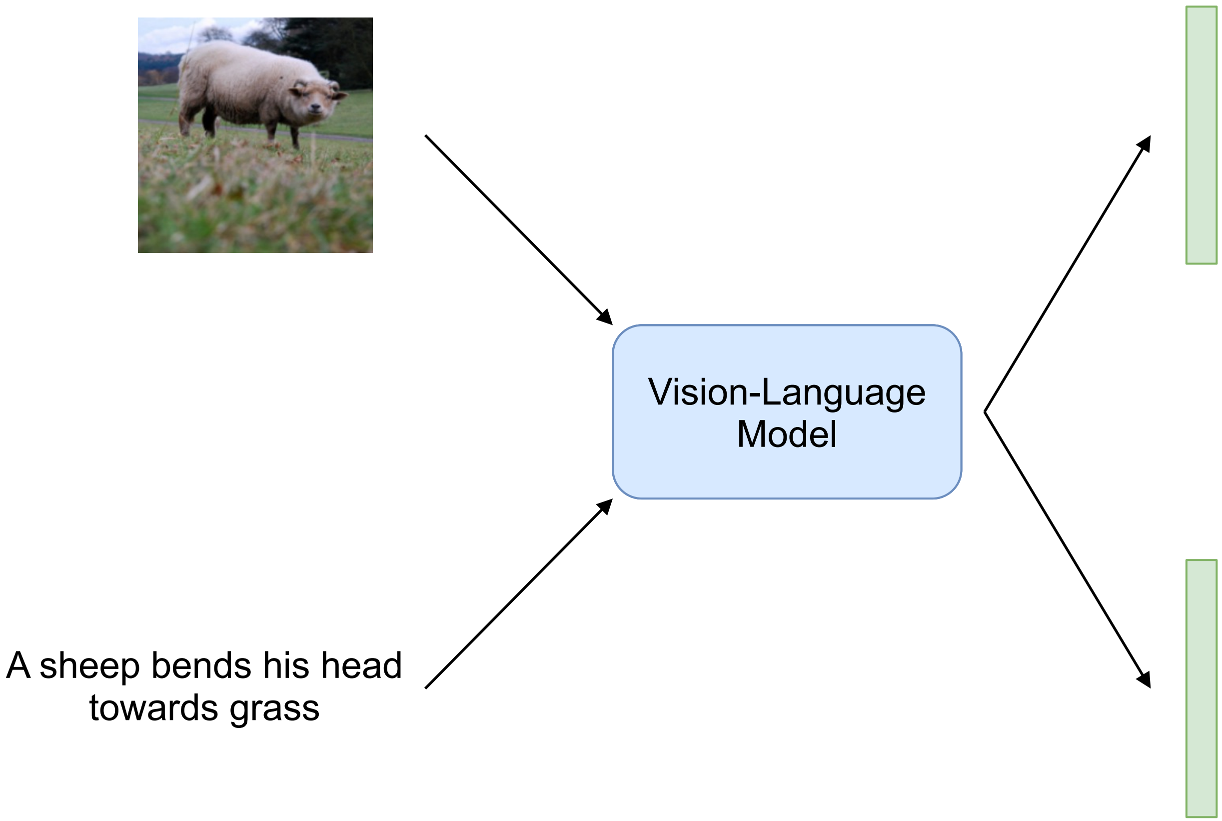
Content

Content

Motivation

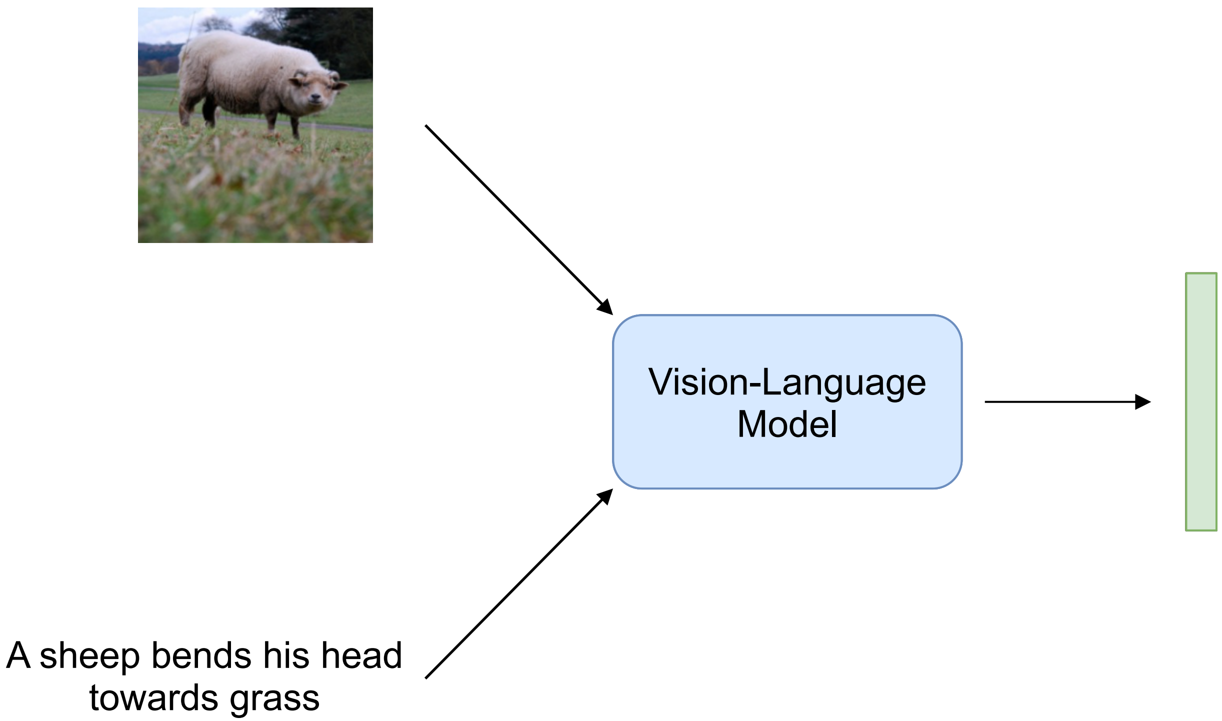

$\in \mathbb{R}^{D_1}$
Vision Model
Image Classification
Image Retrieval
$\in \mathbb{R}^{D_2}$
A sheep bends his head towards grass
Language Model
Sentiment Analysis
Conditional Generation
Translation

Motivation

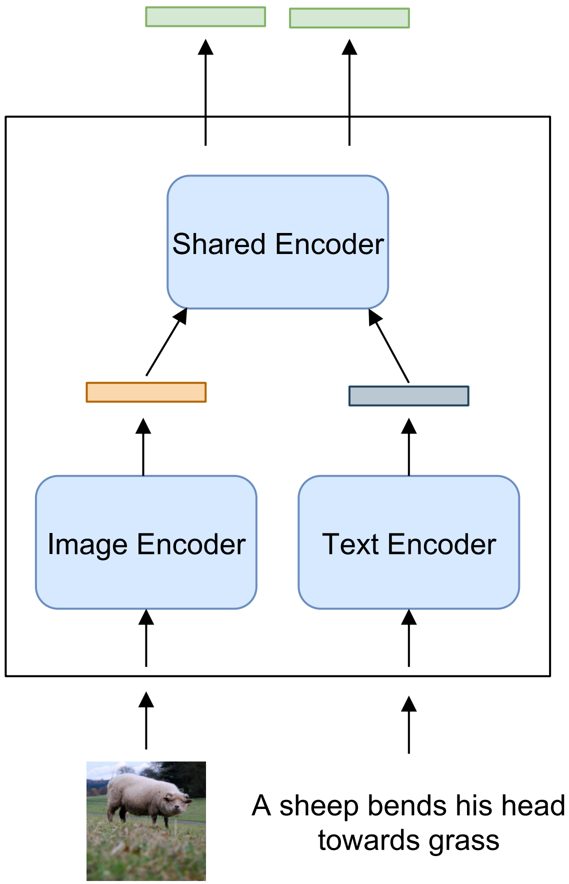


Motivation

Vision-Language Models



Motivation
Vision-Language Models



Motivation

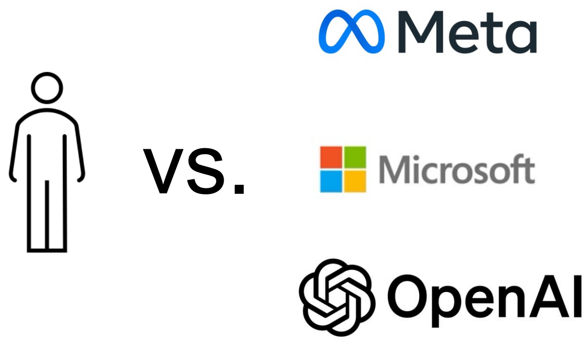
Existing Vision-Language Models

Approach	# Params	Training data (Image-Text pairs)	Estim. Costs ($)
CLIP	428M	400M	>77k
VLMo	562M	1B	>>10k
CoCa	2.1B	>3B	>350k

Content

Research Objective (Contributions)

- **Develop a method for (more) efficient Vision-Language Pretraining**
- It should be:
  - End-to-end self-supervised
  - Independent of pretrained multimodal components
  - Cheaper & smaller than existing VL models
  - Competitive in performance?

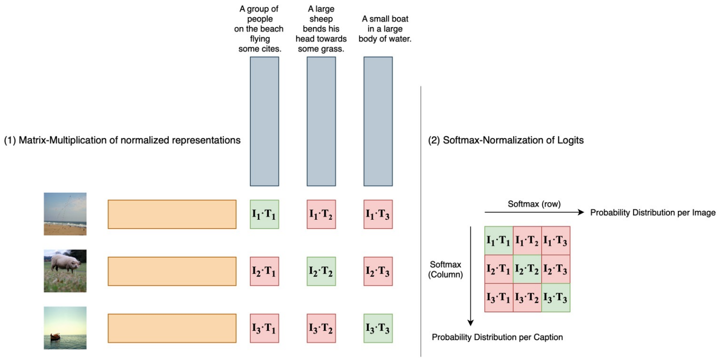


Content

Method

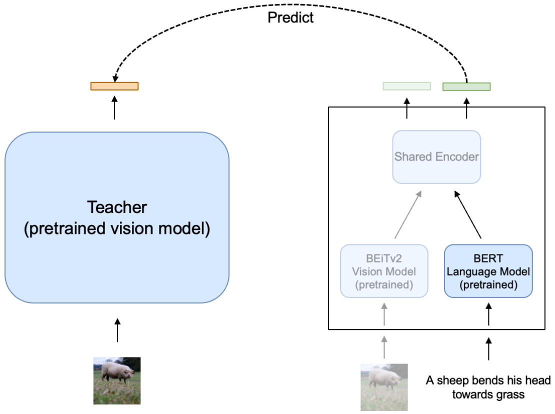
Overview

Contrastive Loss

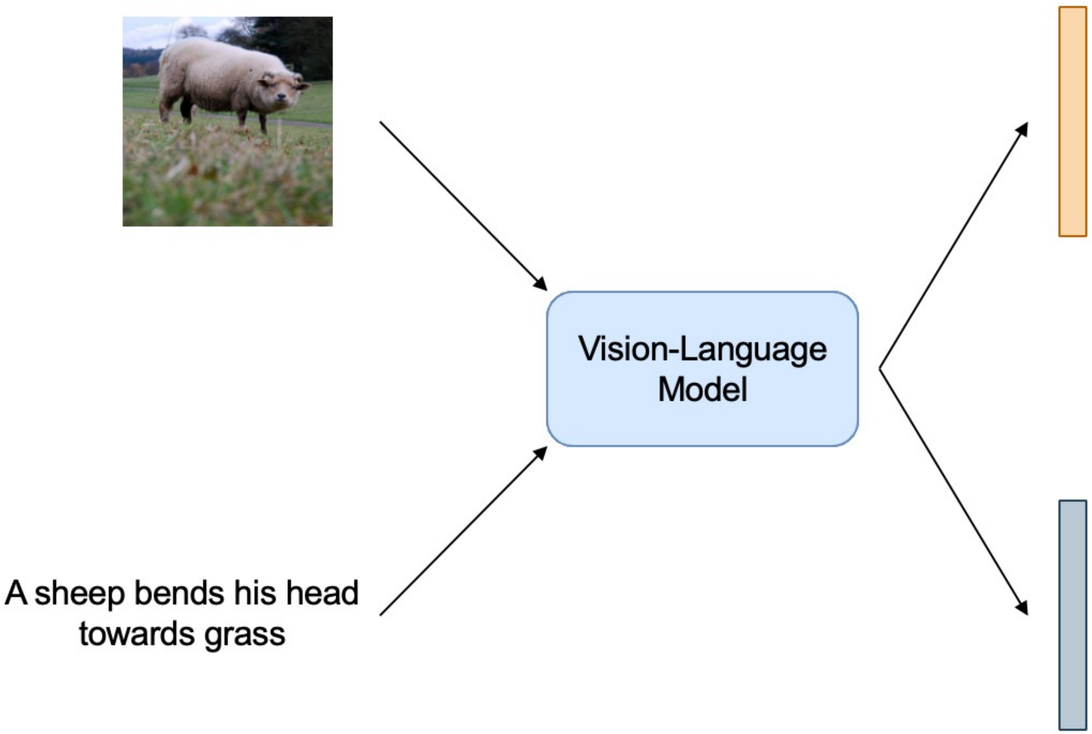


Pretrained Modules

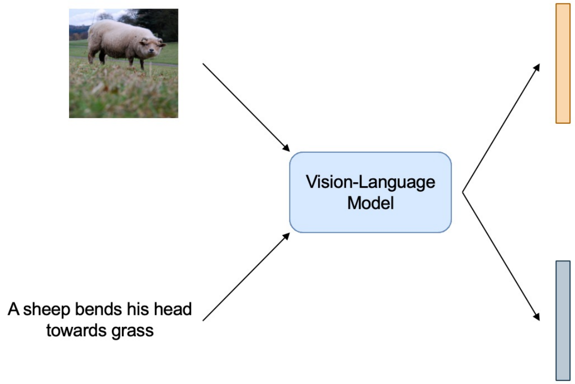
Knowledge Distillation (guidance)

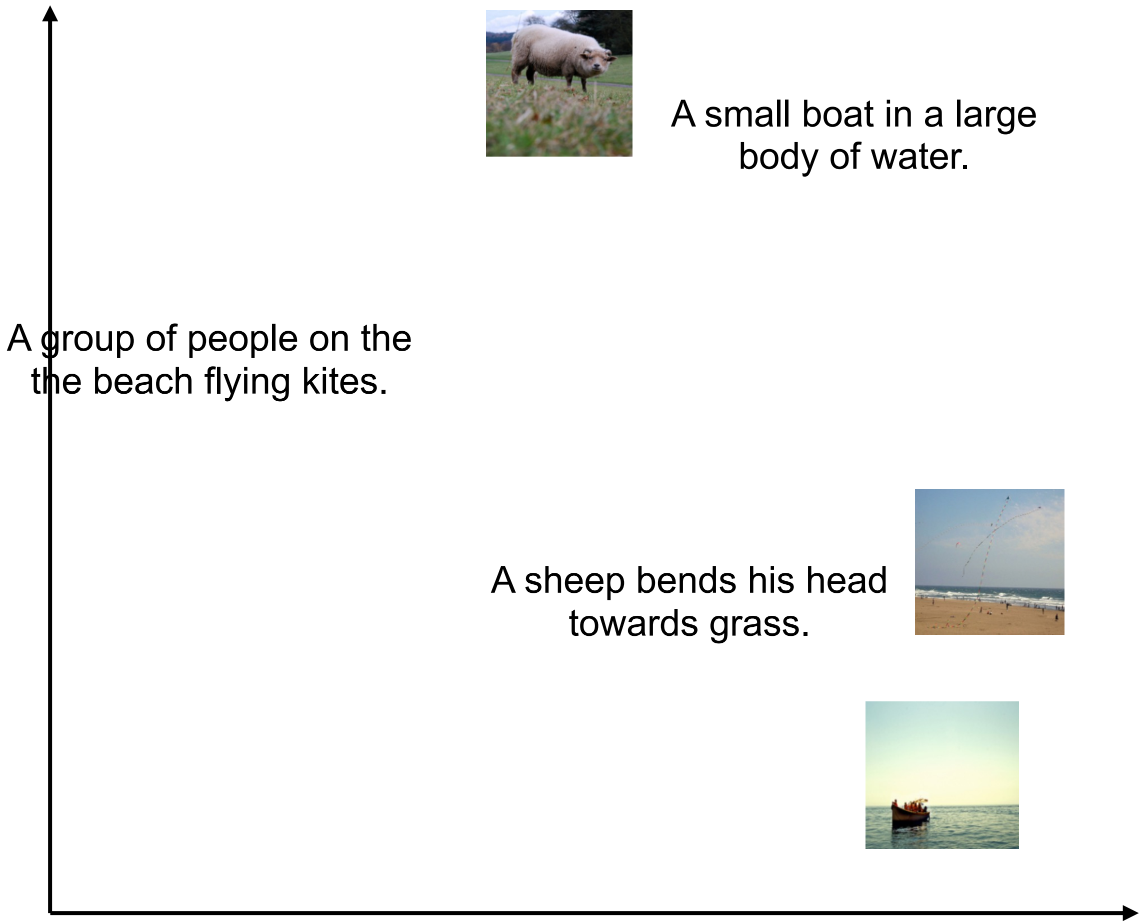


Method
Contrastive Loss



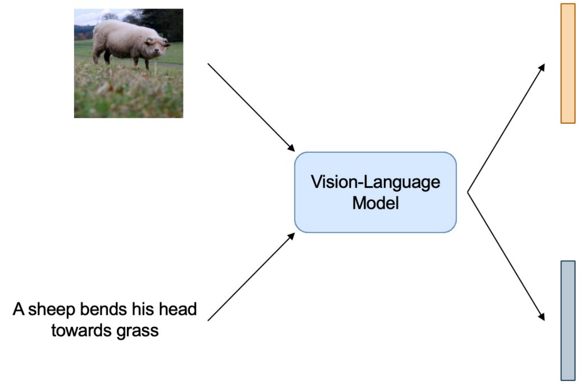
Method
Contrastive Loss

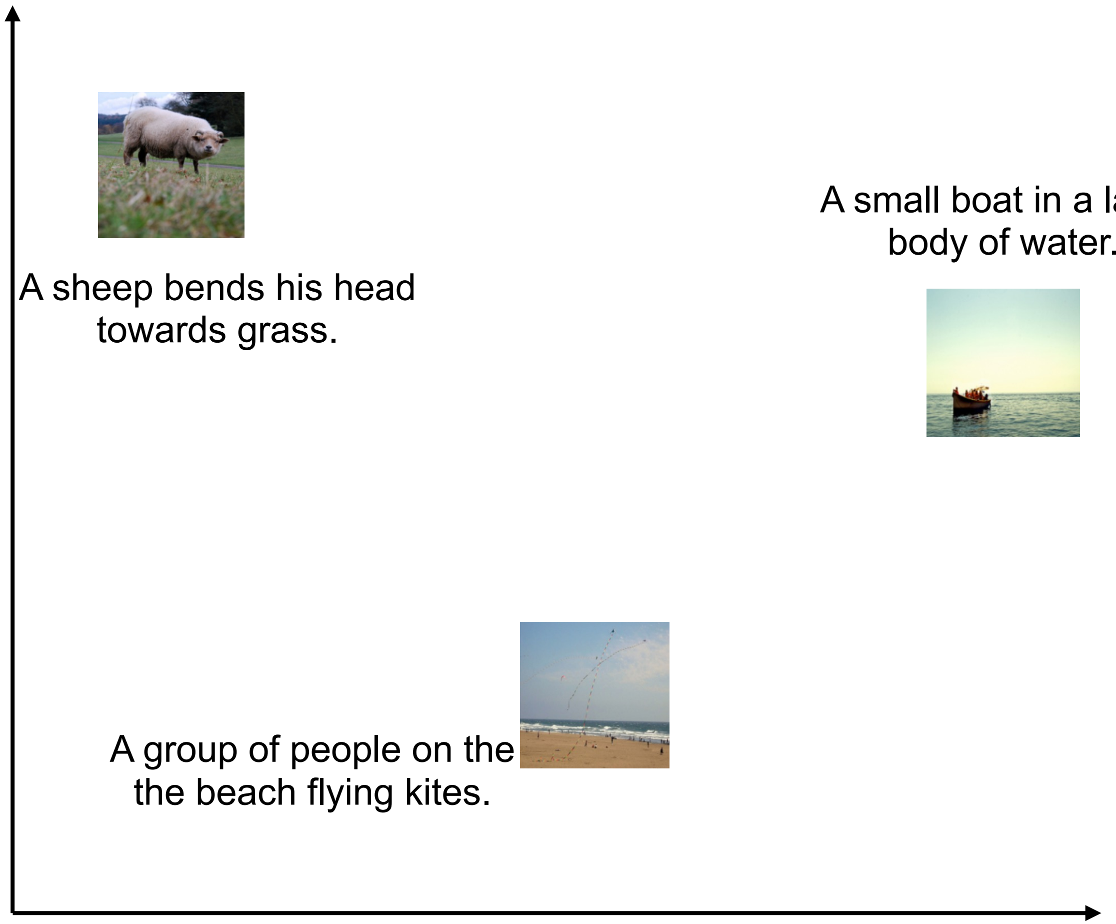




Method

Contrastive Loss

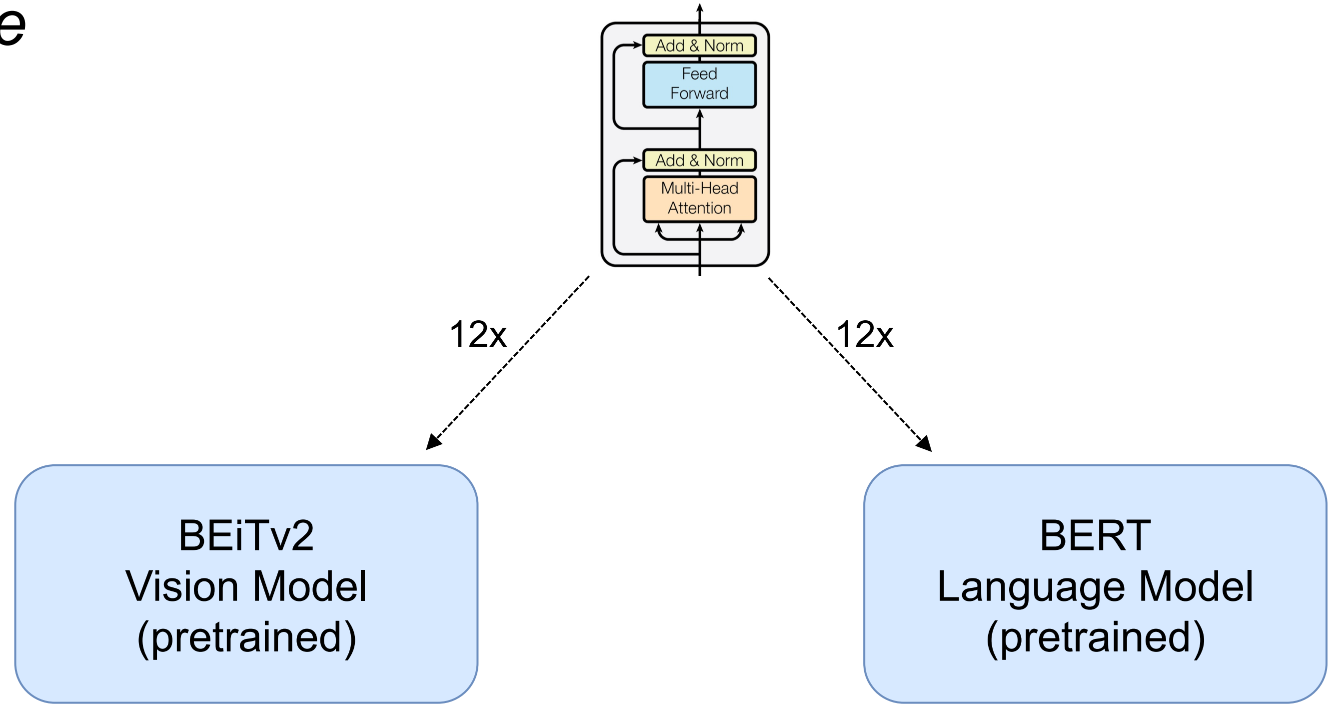




Goal: Find suitable image-text pair in a set of images and texts
=> Ensures *alignment*

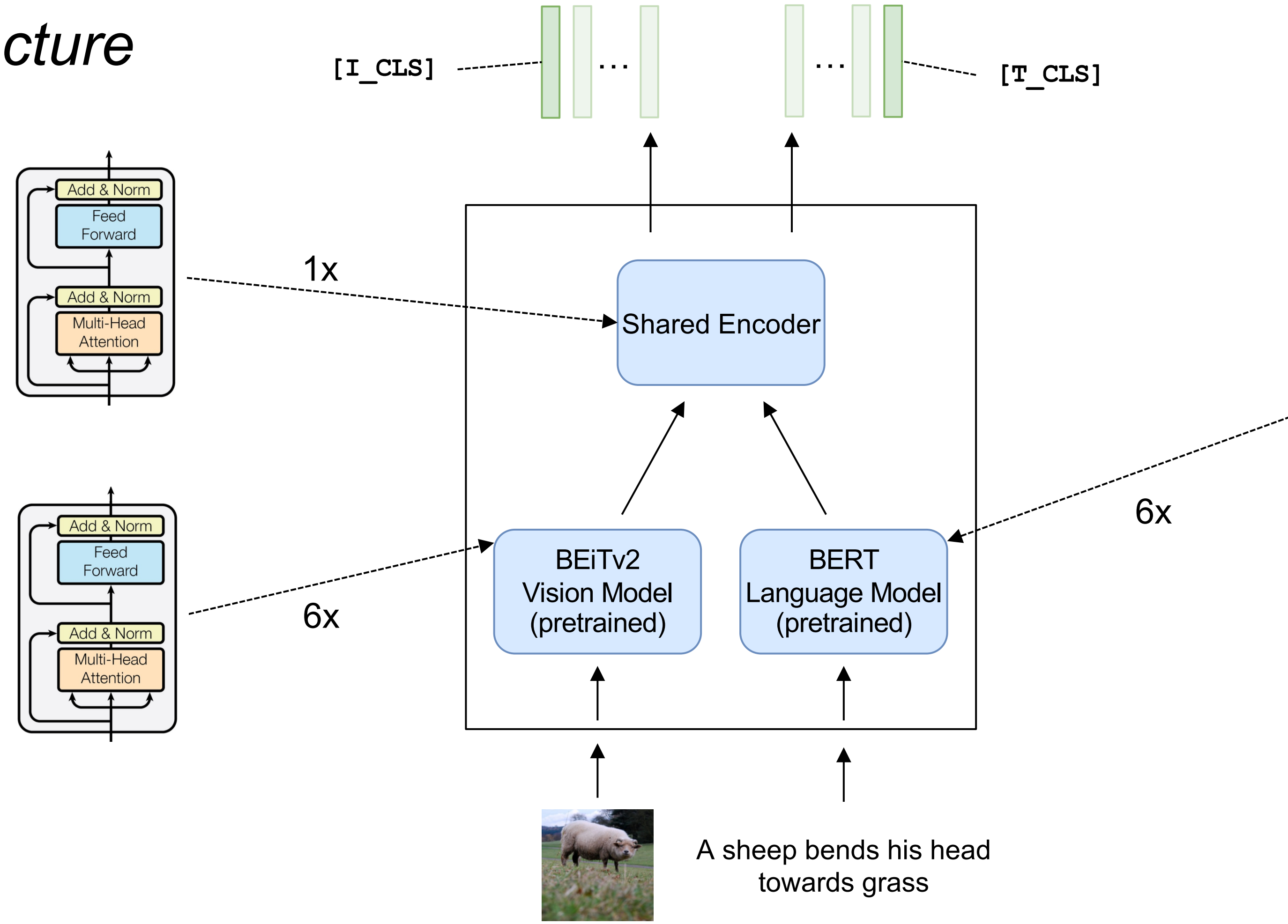
Method

Architecture



Method

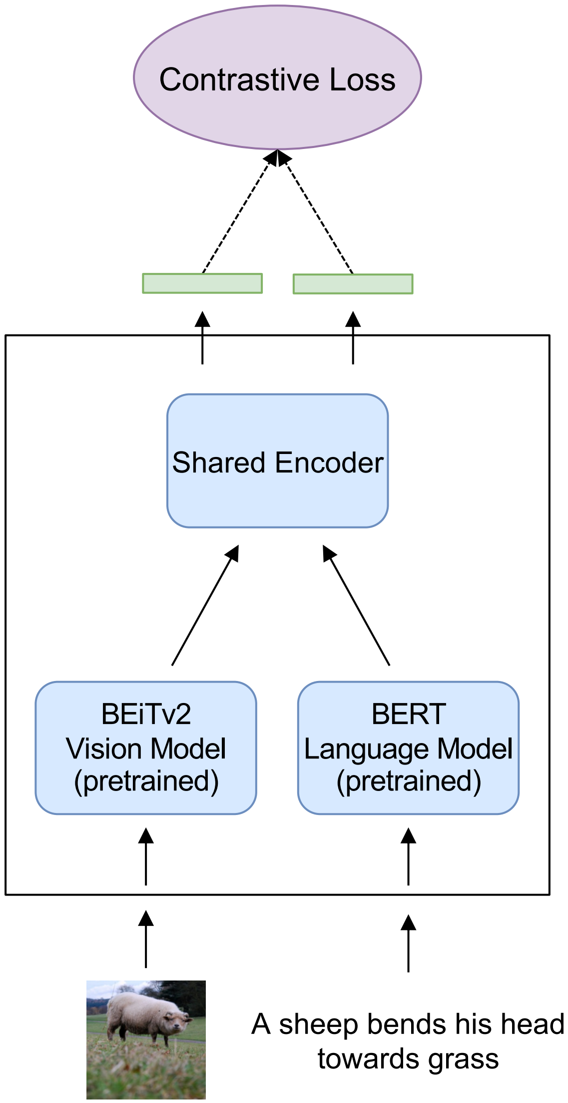
Architecture



Goal: Reduces data requirements and computational resources

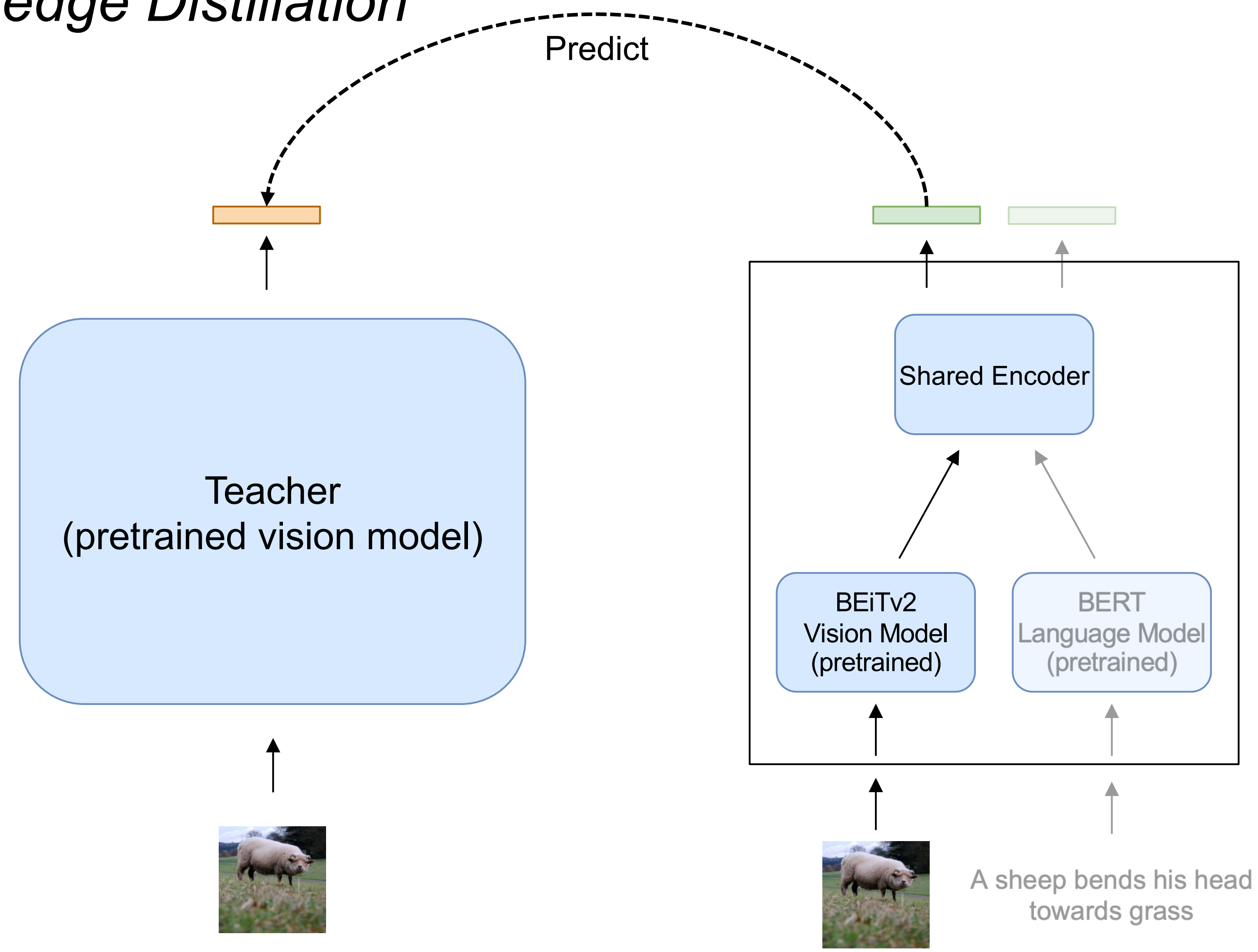
Method

Architecture



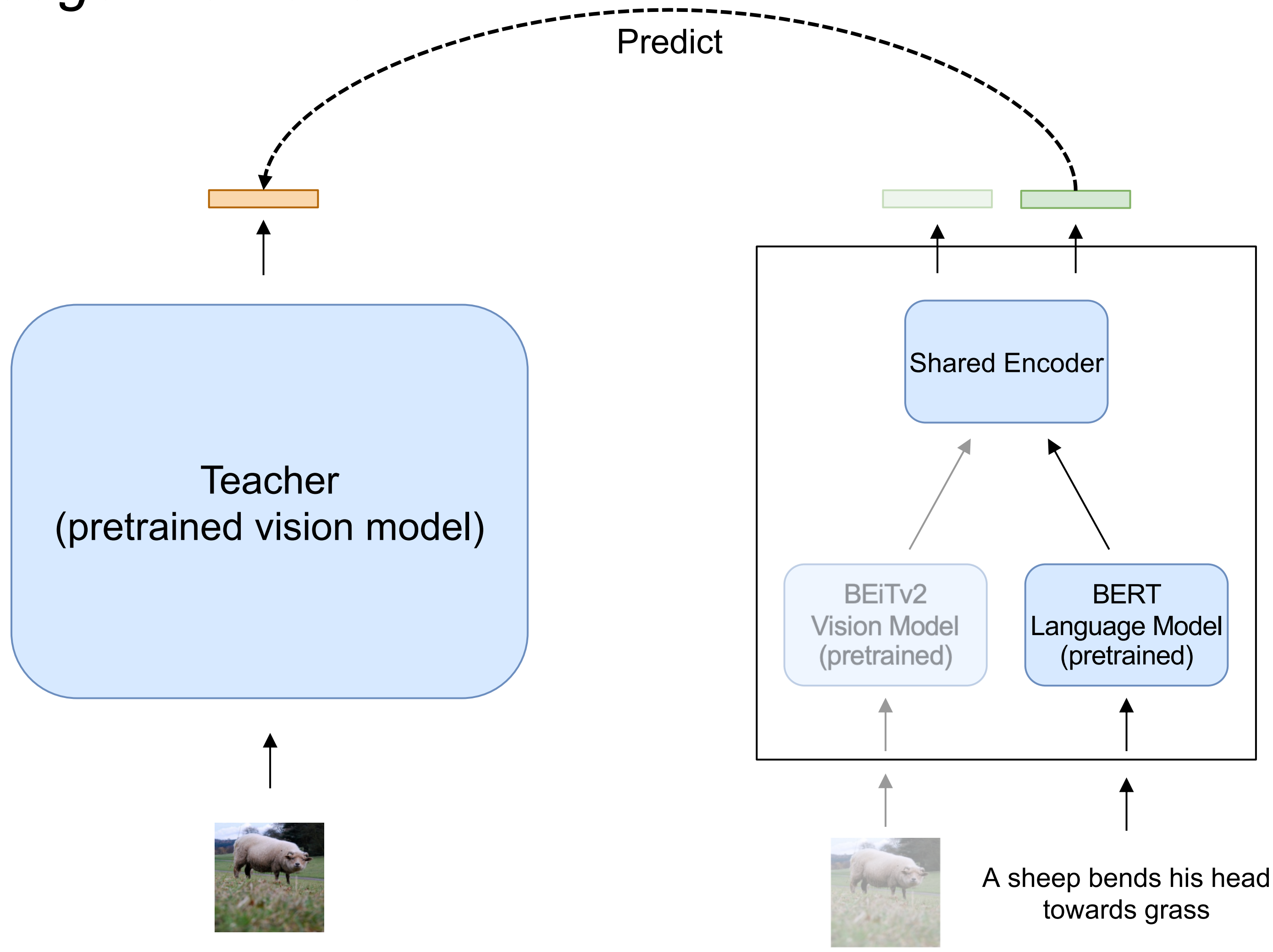
Method

Knowledge Distillation



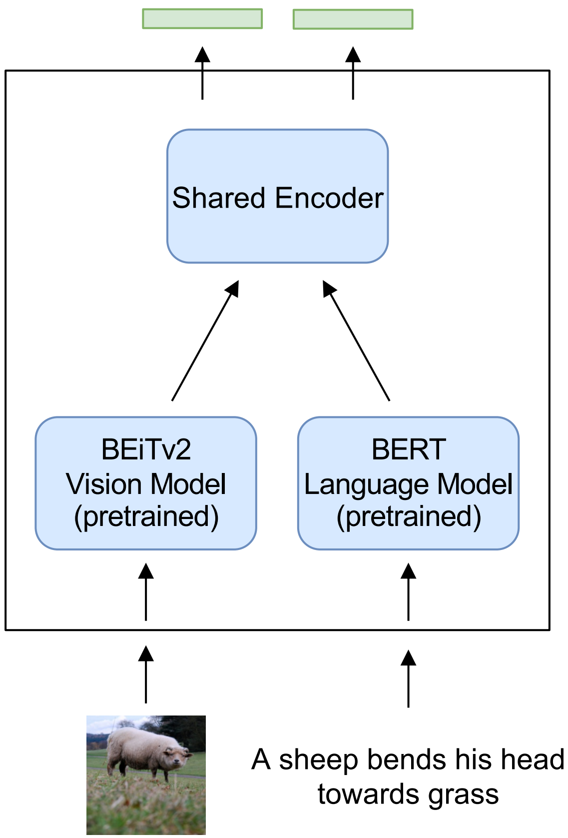
Method

Knowledge Distillation



Goal: Guides alignment => Improvement

Method
Knowledge Distillation



Goal: Guides alignment => Improvement

Content

Results

Image-Text Retrieval

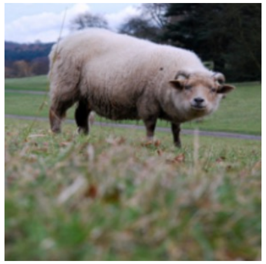


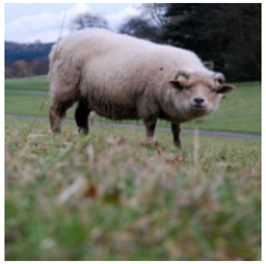
Image-to-Text

A group of people on the
the beach flying kites.

A sheep bends his head
towards grass.

A small boat in a large
body of water.

Results
Image-Text Retrieval



A group of people on the
the beach flying kites.

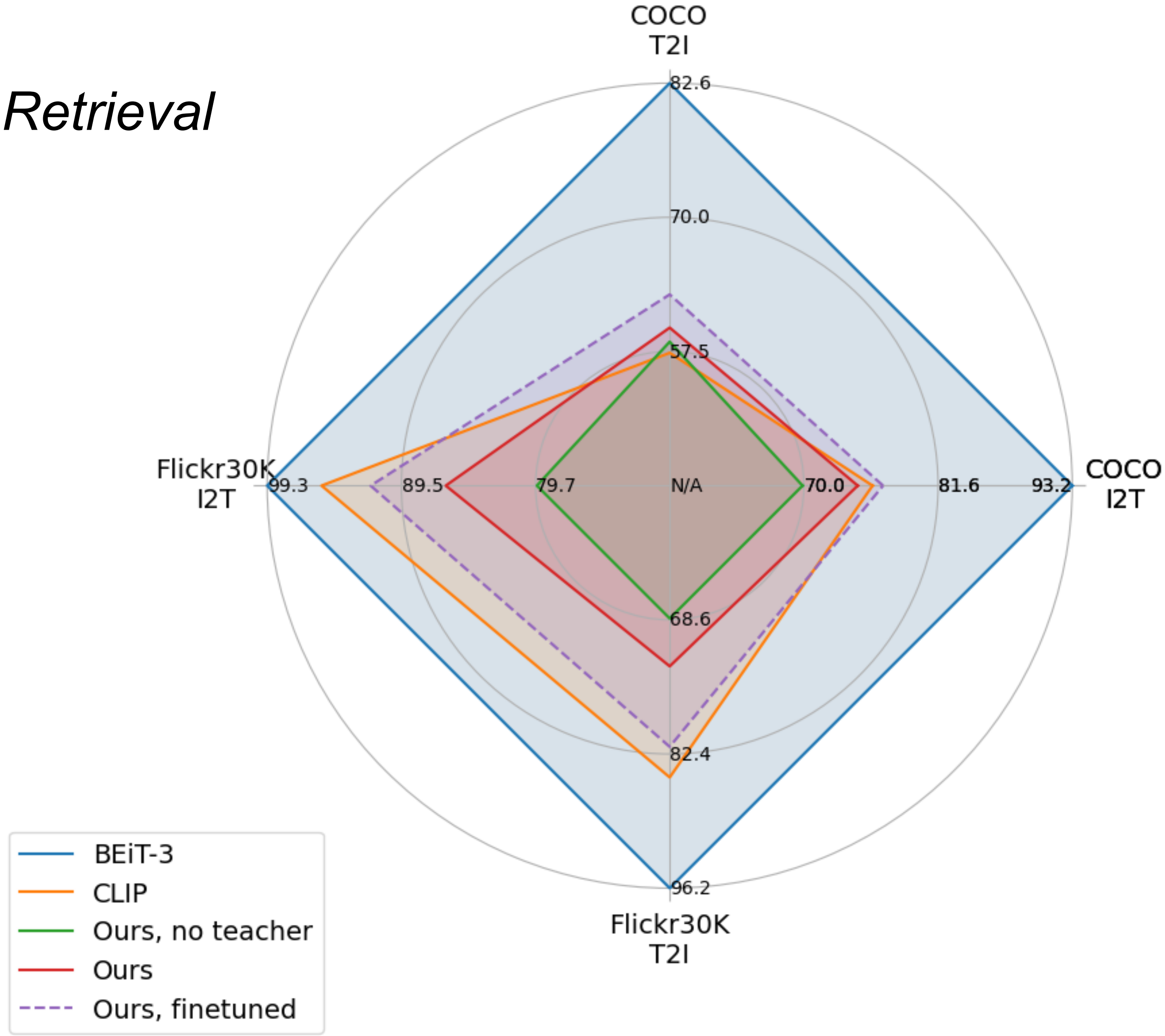
Text-To-Image

A sheep bends his head
towards grass.

A small boat in a large
body of water.

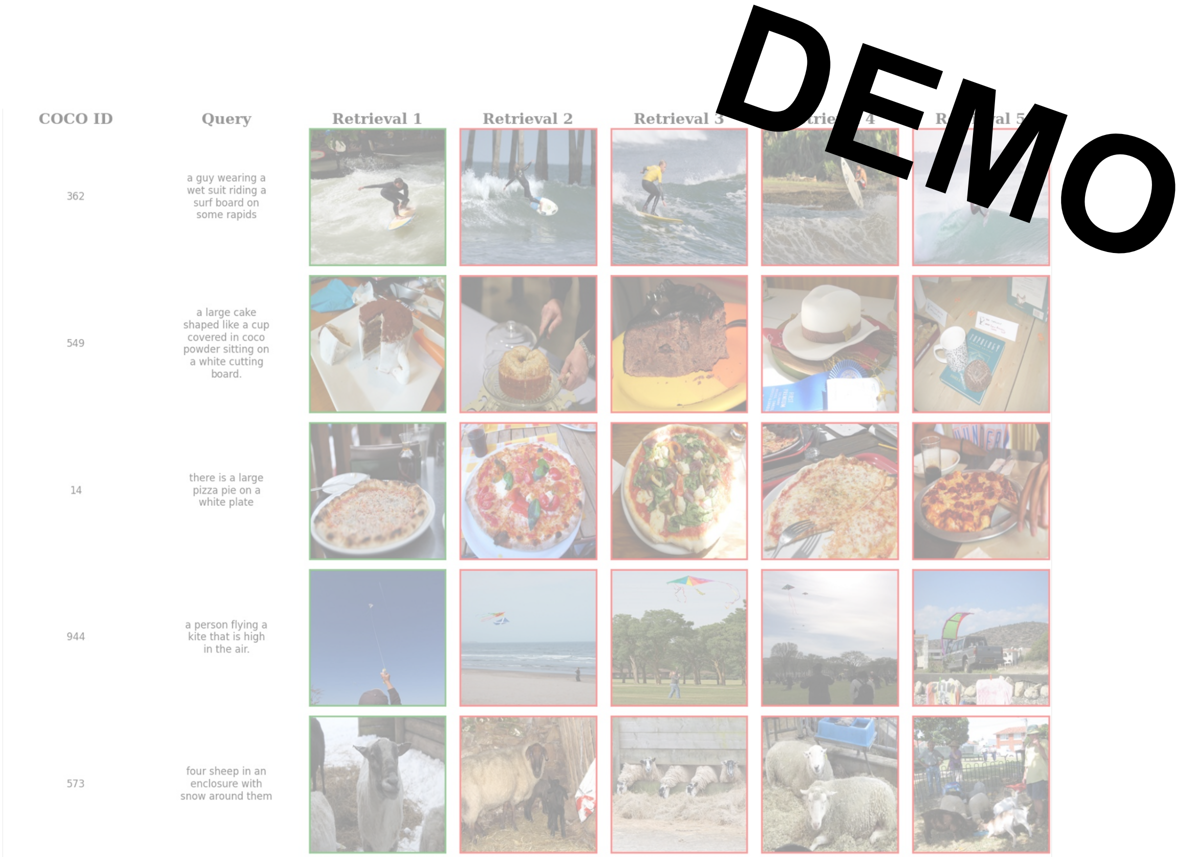
Results

Image-Text Retrieval



Results

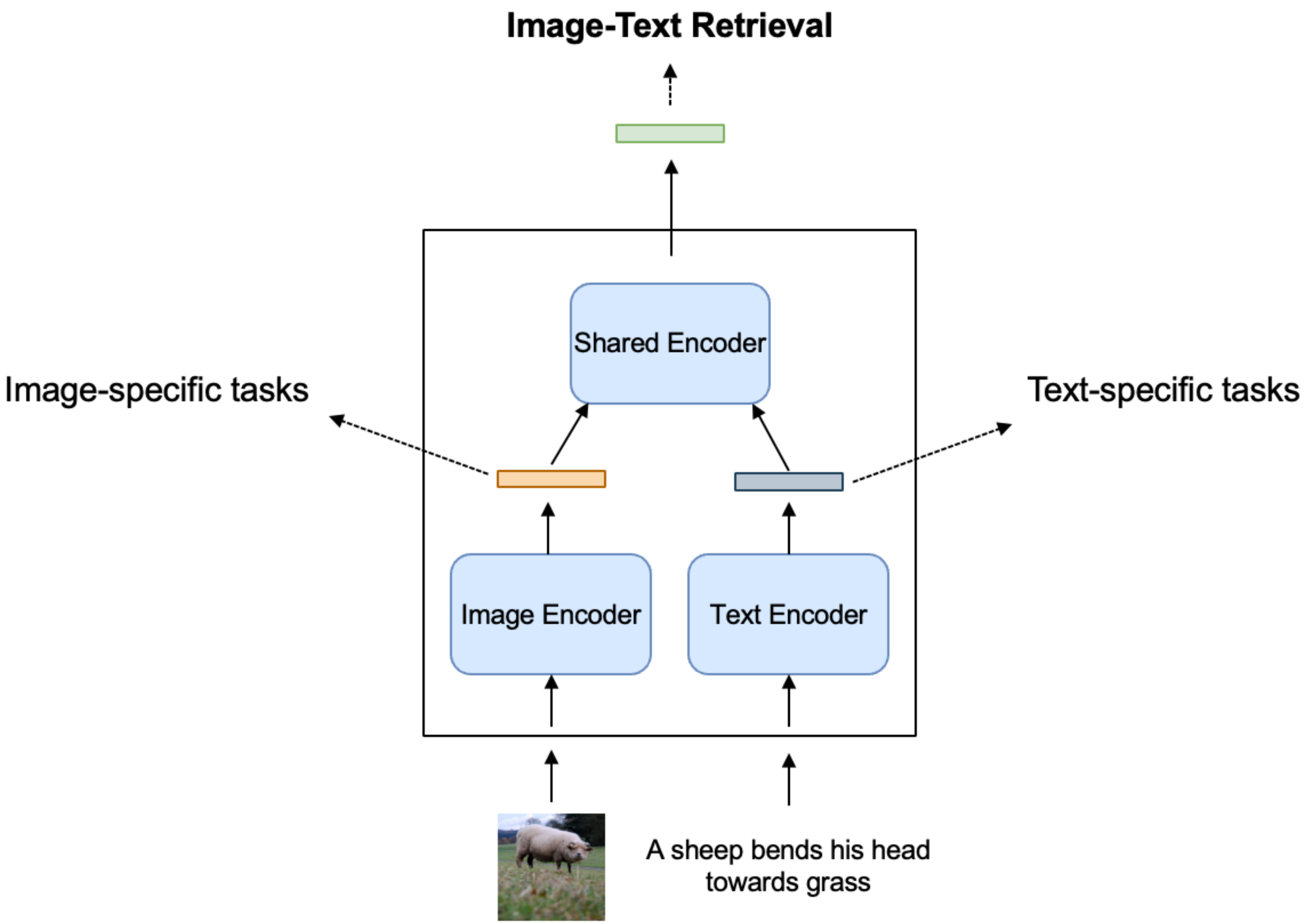
Image-Text Retrieval



Content

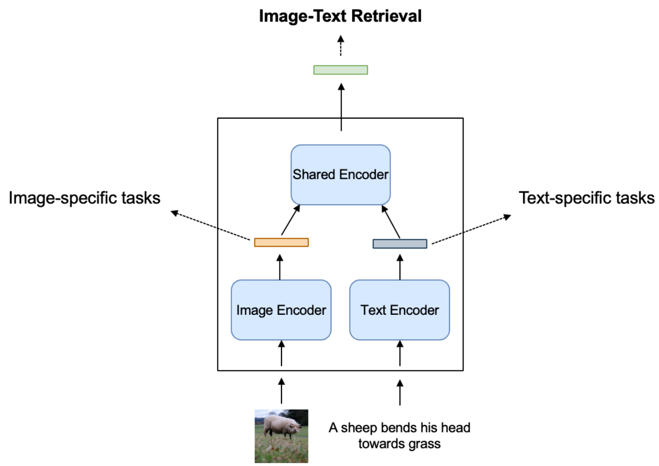
Limitations

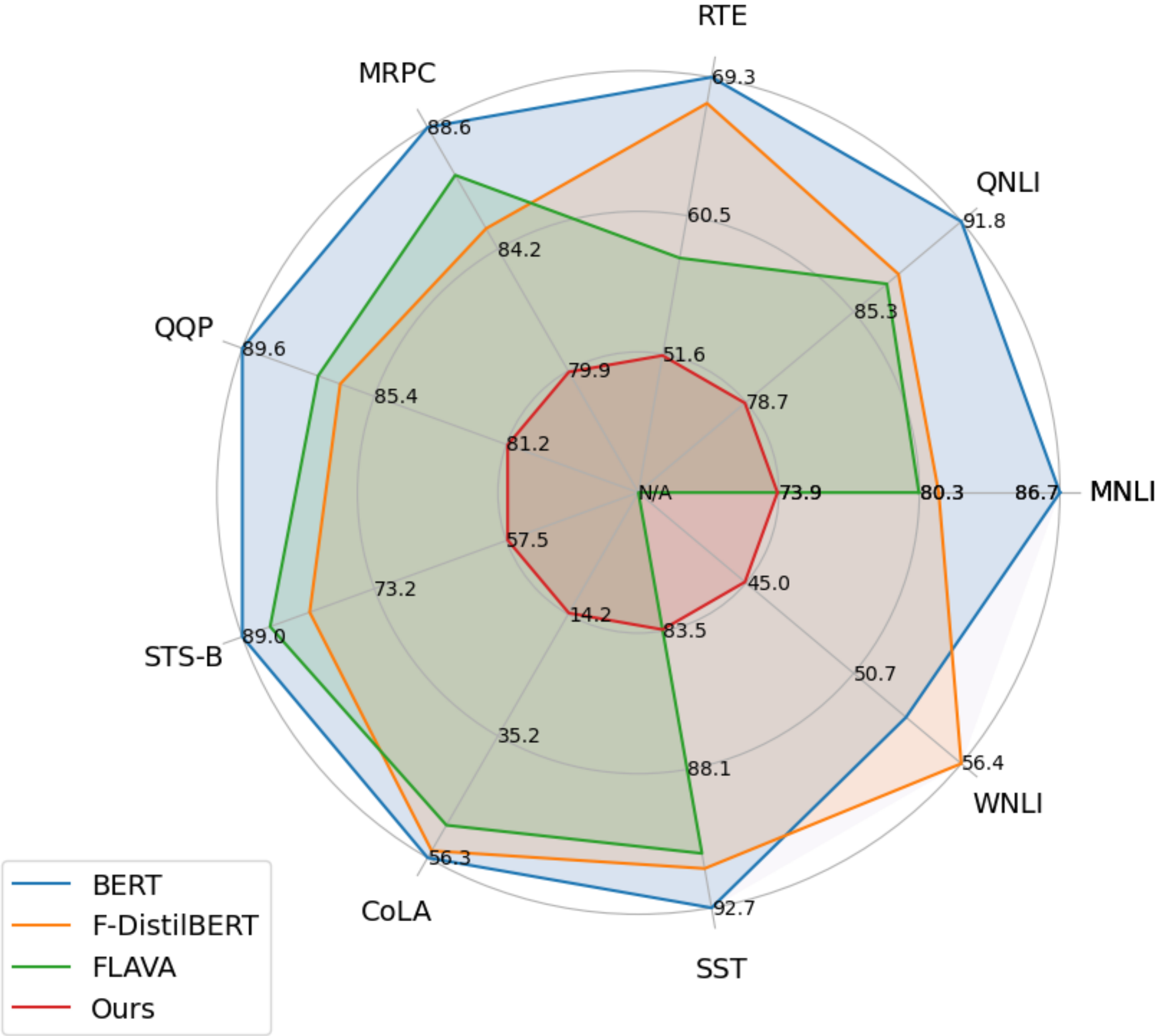
Unimodal Performance



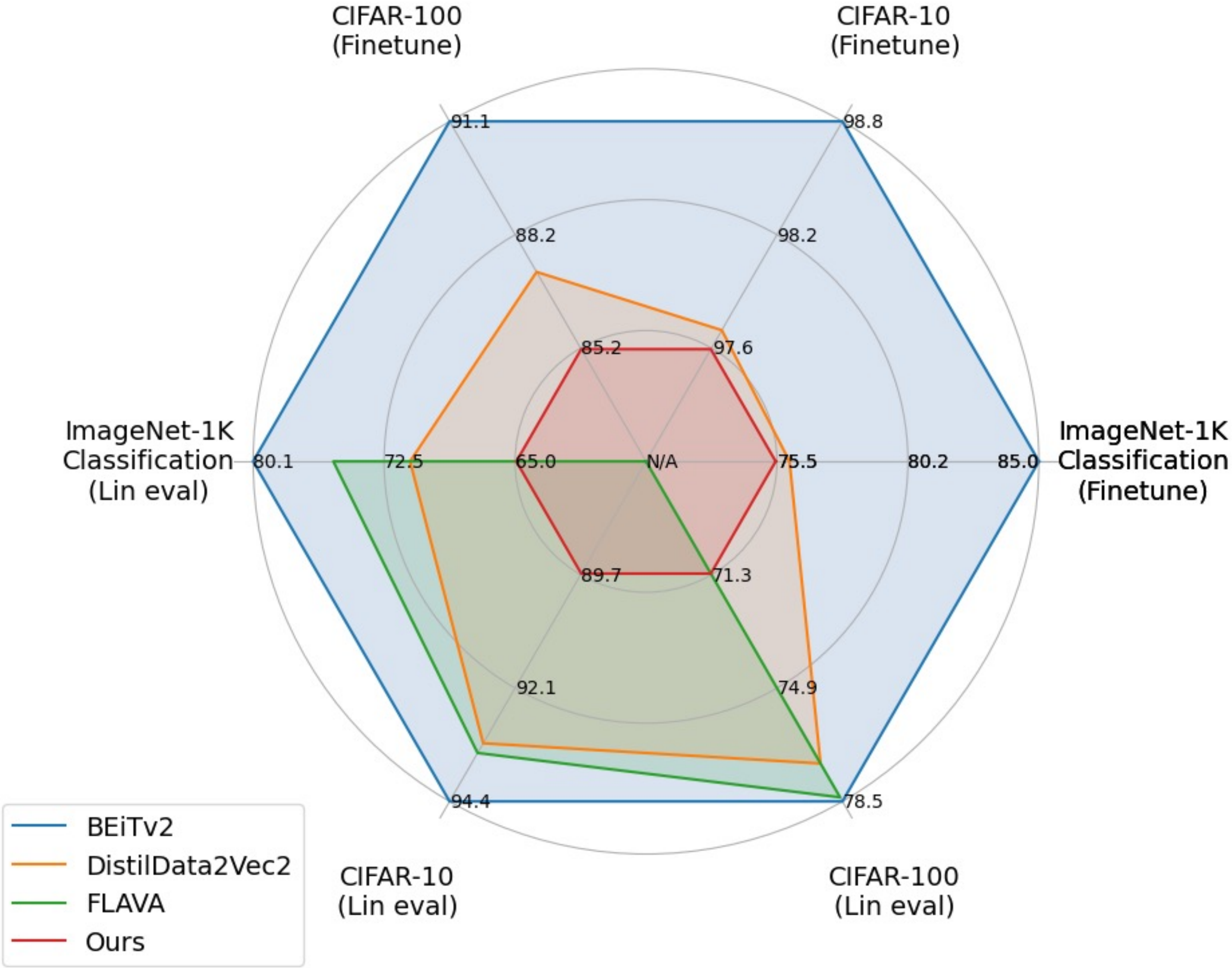
Limitations

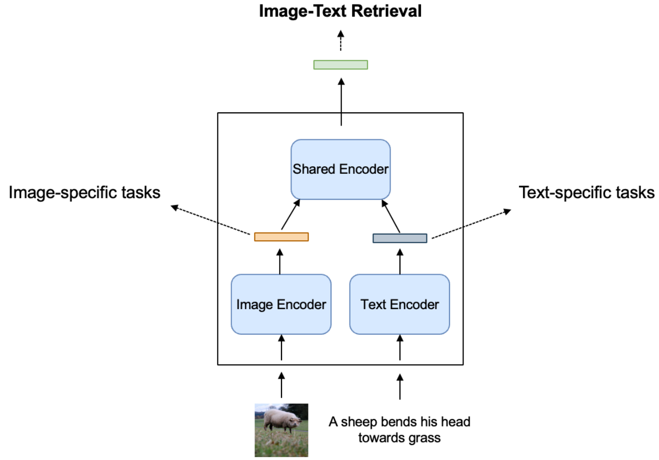
Unimodal Performance





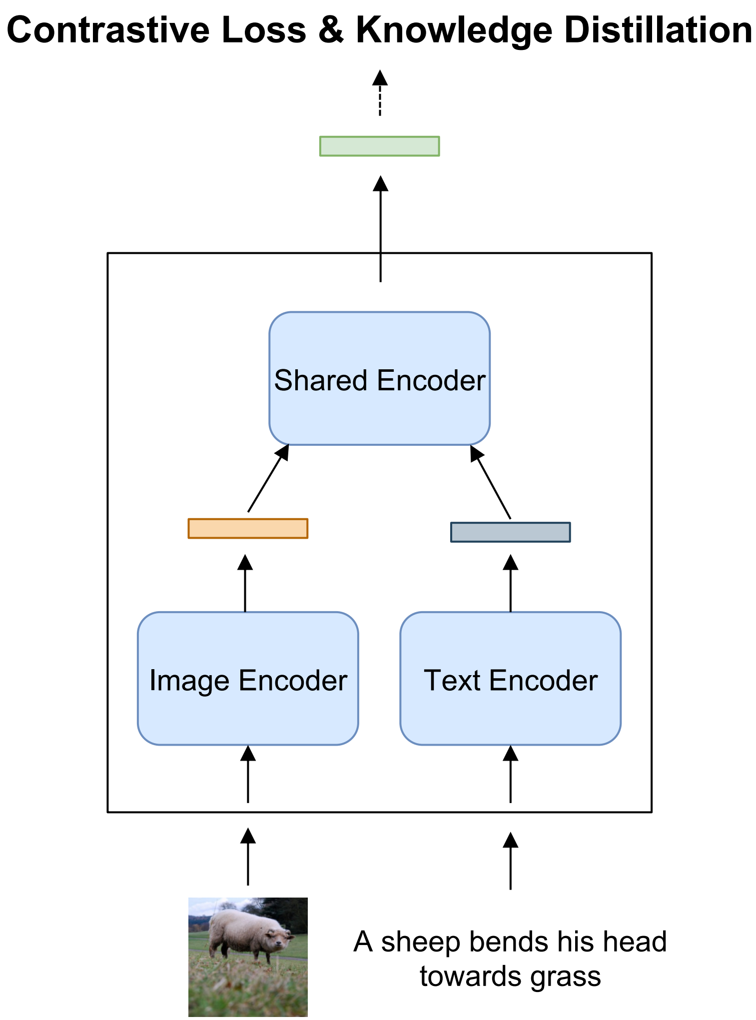
Limitations
Unimodal Performance





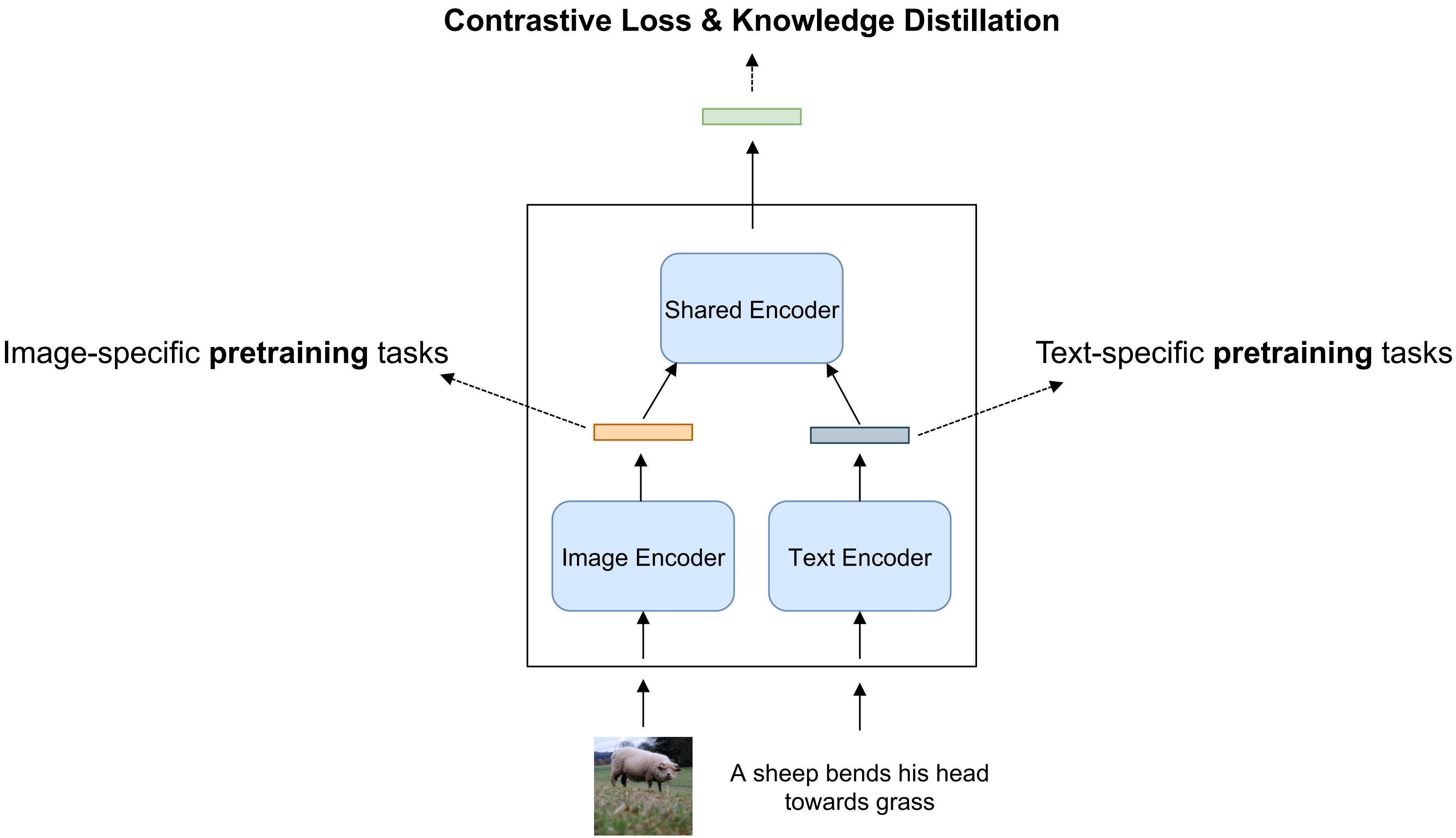
Limitations

Unimodal Performance – Future work



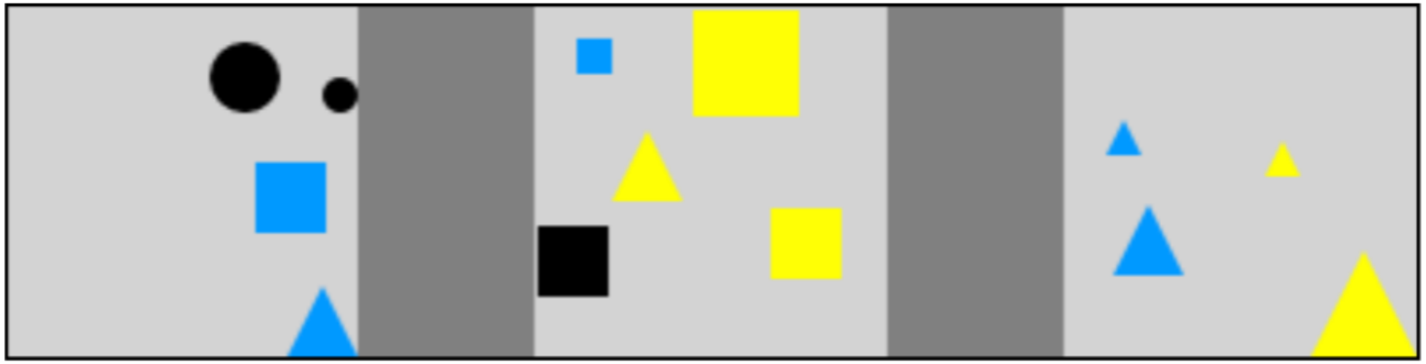
Limitations

Unimodal Performance – Future work



Limitations
Visual Reasoning – NLVR2

Image Captioning
Visual Question Answering
Text-to-Image
Visual Reasoning



One of the grey box has exactly six objects

Content

Conclusion

Research Objectives

Approach	# Params	Training data (Image-Text pairs)	Estim. Costs ($)
CLIP	428M	400M	>77k
VLMo	562M	1B	>>10k
CoCa	2.1B	>3B	>350k
Ours	117M	3.3M	15.5

Conclusion

Research Objectives

Criterion	Fulfilled	Note
End-to-end Self-supervised		
Smaller		
Cheaper		
Competitive in Performance		

99.84% cheaper than VLMo,
99.98% cheaper than CLIP

Main issues: Very restricted in multimodal tasks, poor performance on unimodal tasks.

Thank you for your attention!

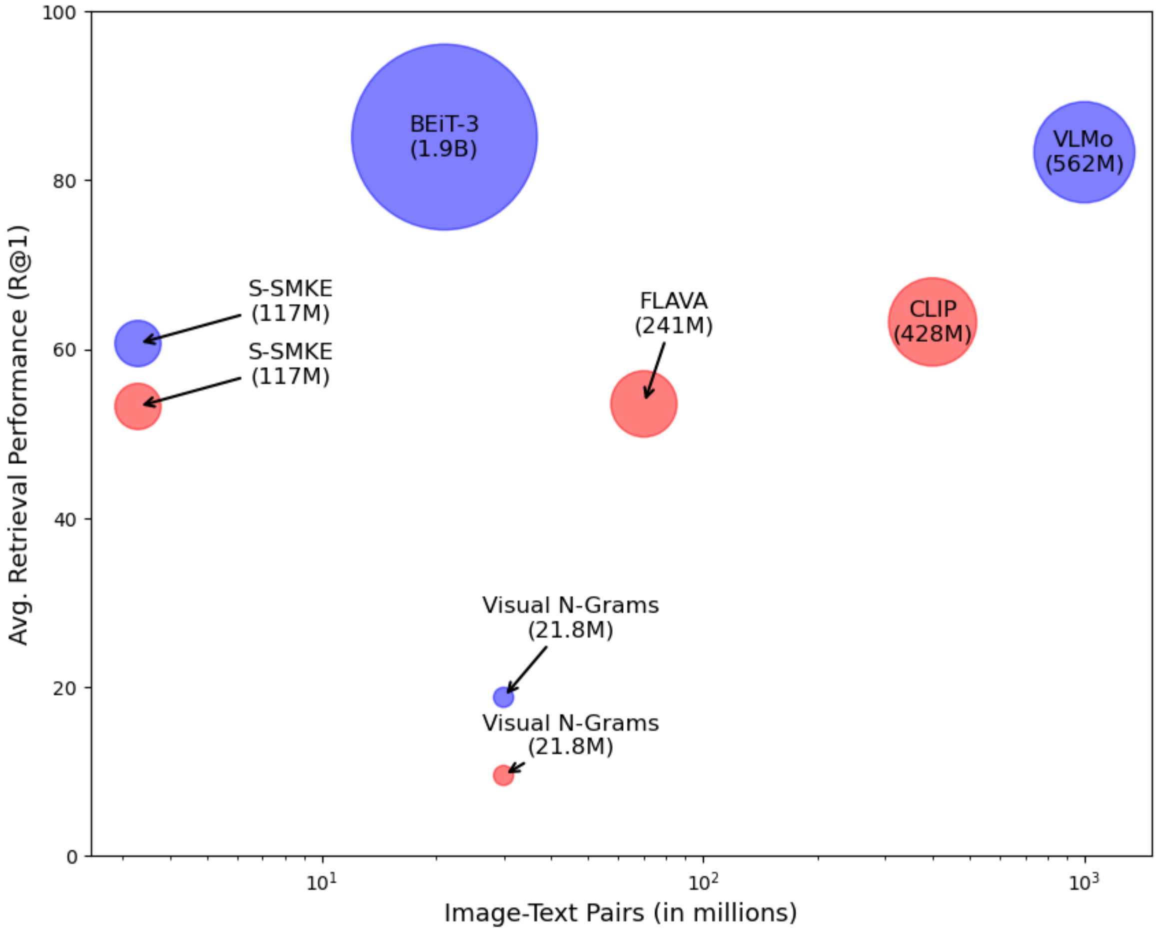
Literature

- T.-Y. Lin et al., “Microsoft COCO: Common Objects in Context,” in Computer Vision – ECCV 2014, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., 2014, pp. 740–755.
- J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, “CoCa: Contrastive Captioners are Image-Text Foundation Models,” Transactions on Machine Learning Research, 2022, [Online]. Available: https://openreview.net/forum?id=Ee277P3AYC
- A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” in Proceedings of the 38th International Conference on Machine Learning, M. Meila and T. Zhang, Eds., 2021, pp. 8748–8763.
- H. Bao et al., “VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts,” in Advances in Neural Information Processing Systems, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., Curran Associates, Inc., 2022, pp. 32897–32912.
- W. Wang et al., “Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks,” in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19175–19186.
- A. Singh et al., “FLAVA: A Foundational Language And Vision Alignment Model,” in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15617–15629.
- Y. Aytar, C. Vondrick, and A. Torralba, “See, Hear, and Read: Deep Aligned Representations,” CoRR, 2017, [Online]. Available: http://arxiv.org/abs/1706.00932
- A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi, “A Corpus for Reasoning about Natural Language Grounded in Photographs,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6418–6428.

Literature

- Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, “BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers,” CoRR, 2022, [Online]. Available: https://arxiv.org/abs/2208.06366
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, J. Burstein, C. Doran, and T. Solorio, Eds., Jun. 2019, pp. 4171–4186.
- J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge Distillation: A Survey,” International Journal of Computer Vision, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.

Conclusion
Vision-Language Landscape



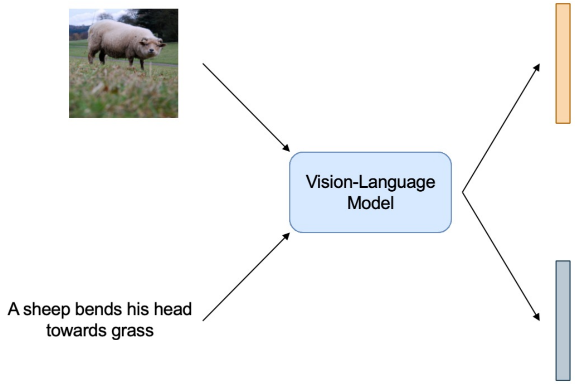
Conclusion

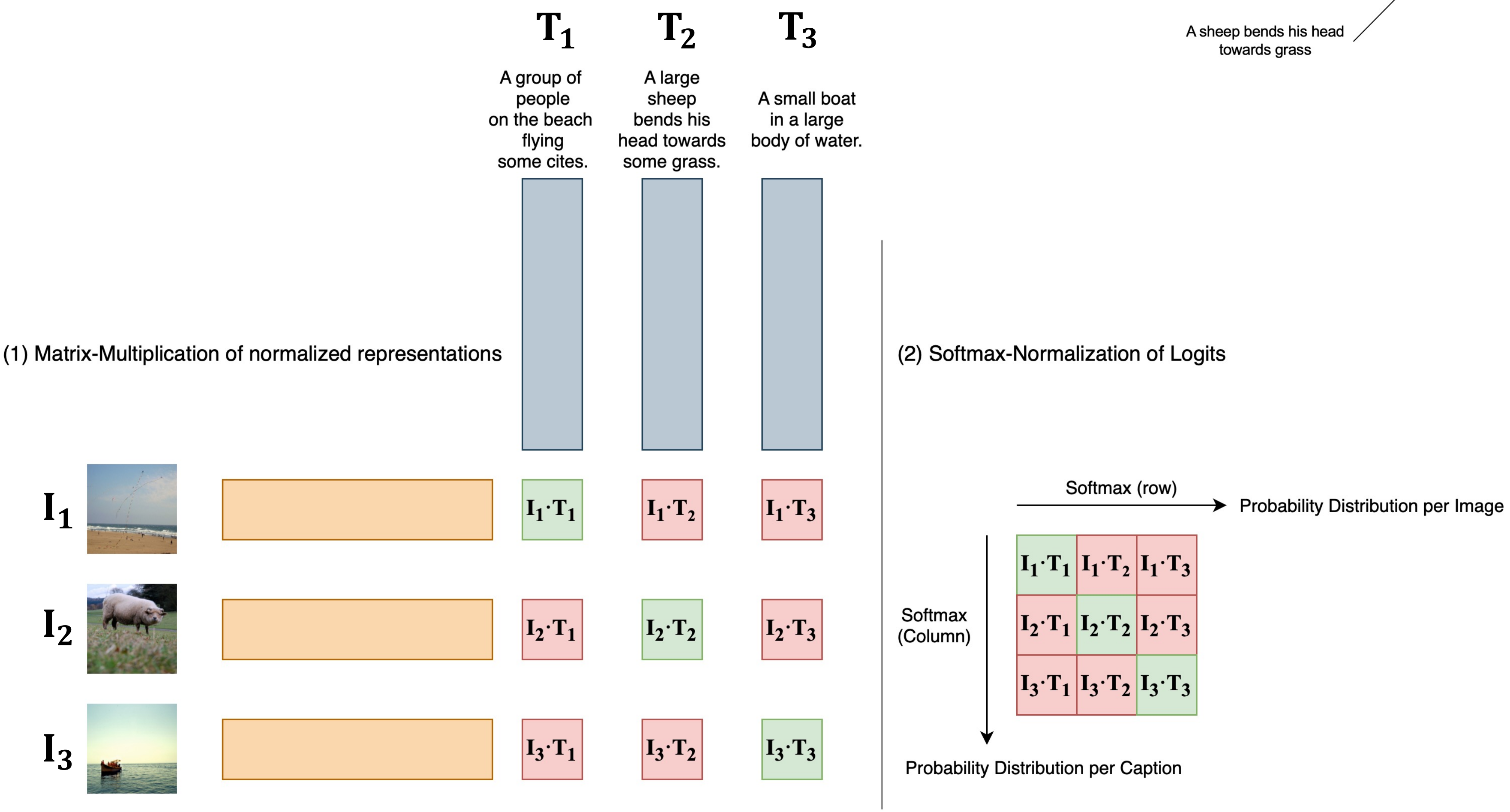
Bachelor vs. Master‘s thesis

Criterion	Bachelor thesis	Master’s thesis
Model Size	<11M	117M (202M)
Data	60k	>3.3M
Data Collection	Available ready to use	Complex collection/scraping and preprocessing from various sources
Performance	635th place	Beating papers from Meta and OpenAI in some benchmarks

Method

Contrastive Loss

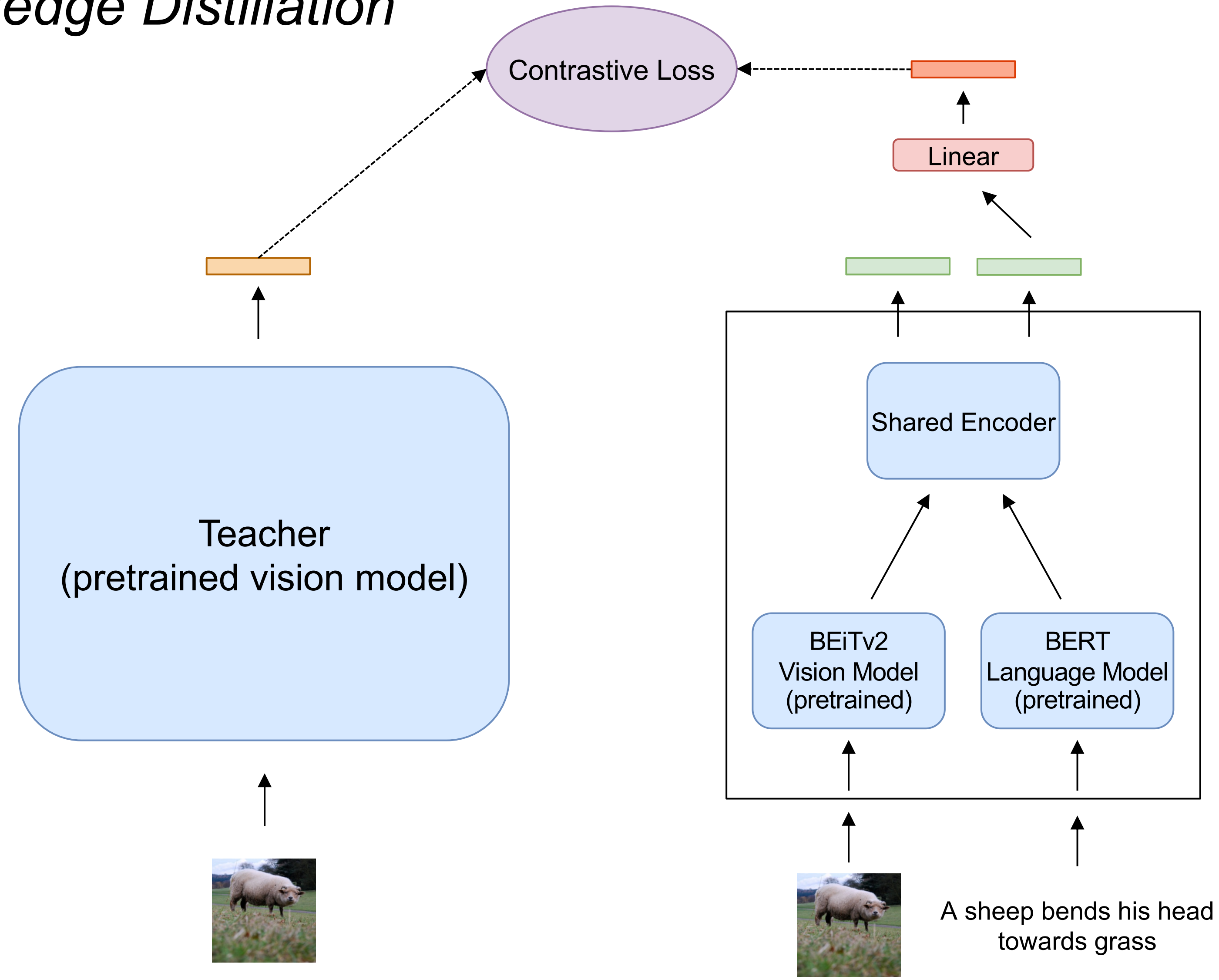




Goal: Find suitable image-text pair in a set of images and texts
=> Ensures *alignment*

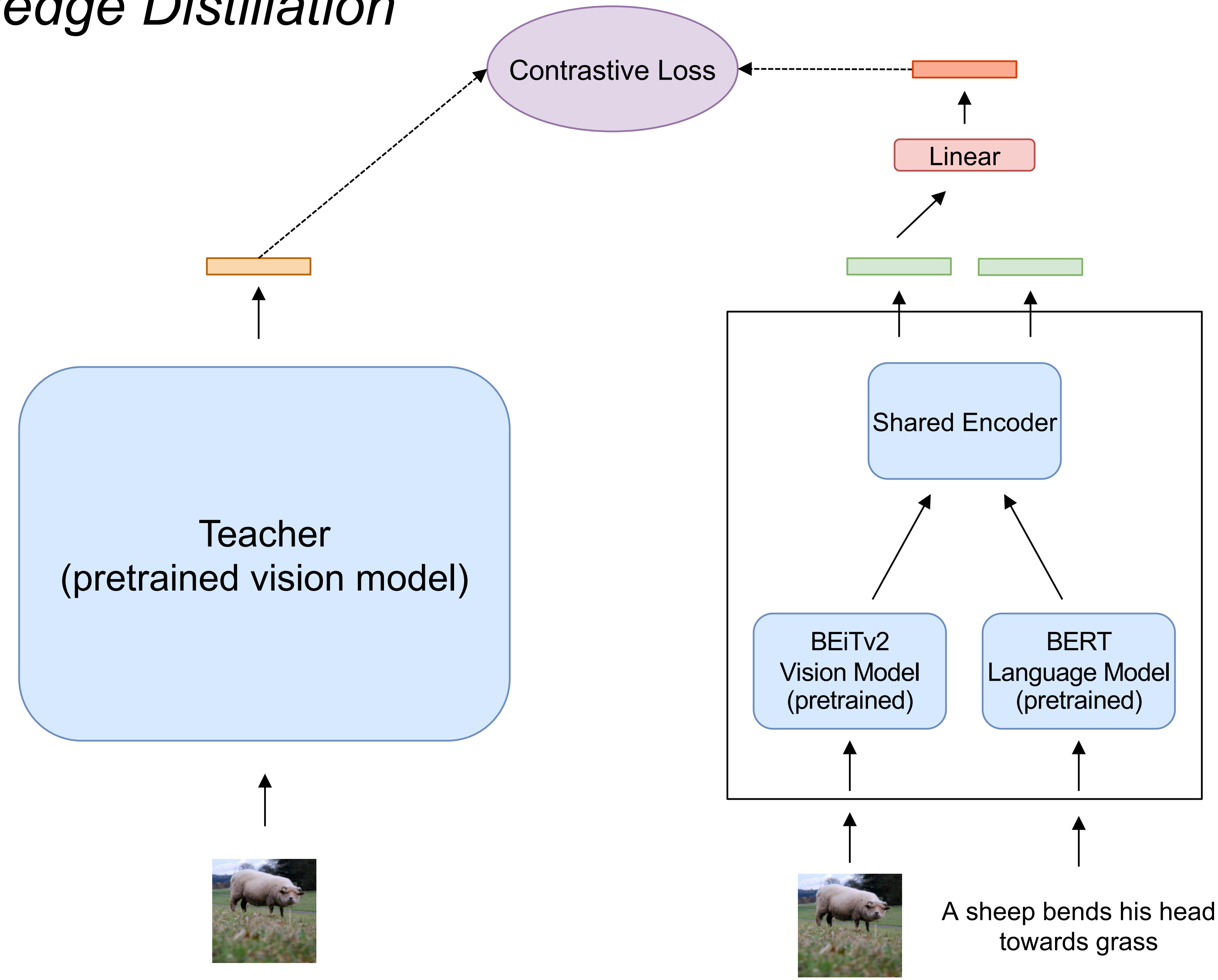
Method

Knowledge Distillation



Method

Knowledge Distillation



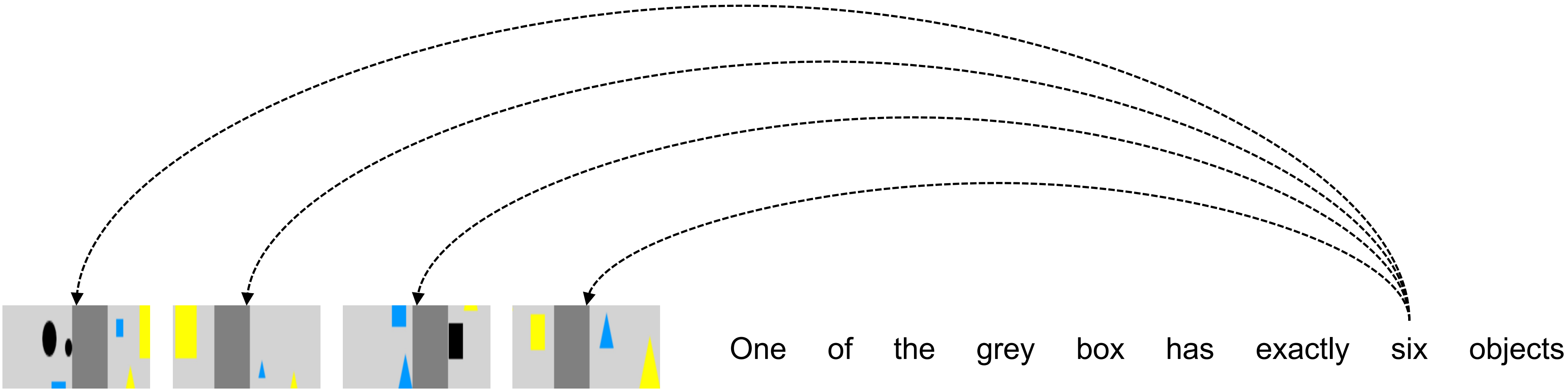
Results
Image-Text Retrieval

Model	MSCOCO (5K test set)						Flickr30K (1K test set)					
	Image → Text			Text → Image			Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
FLAVA	42.74	76.76	-	38.38	67.47	-	67.7	94.0	-	65.22	89.38	-
CLIP	58.4	81.5	88.1	37.8	62.4	72.2	88.0	98.7	99.4	68.7	90.6	95.2
BEiT-3	**84.8**	**96.5**	**98.3**	**67.2**	**87.7**	**92.8**	**98.0**	**100.0**	**100.0**	**90.3**	**98.7**	**99.5**
S-SMKE	53.54	81.1	89.52	35.65	66.0	77.77	70.9	92.1	96.0	52.72	80.2	87.46
S-SMKE finetuned	56.2	83.3	91.1	39.8	69.2	79.8	82.0	95.4	98.0	64.6	87.5	93.1

Metric	Meaning
R@1	Percentage of images where the correct text is the top-ranked result, or vice versa.
R@5	Percentage of images where the correct text is found within the top-5 results, or vice versa.
R@10	Percentage of images where the correct text is found within the top-10 results, or vice versa.

Limitations

Visual Reasoning – NLVR2



Limitations

Visual Reasoning – NLVR2

Image Captioning

Visual Question Answering

Text-to-Image

Visual Reasoning