THU-PM-274

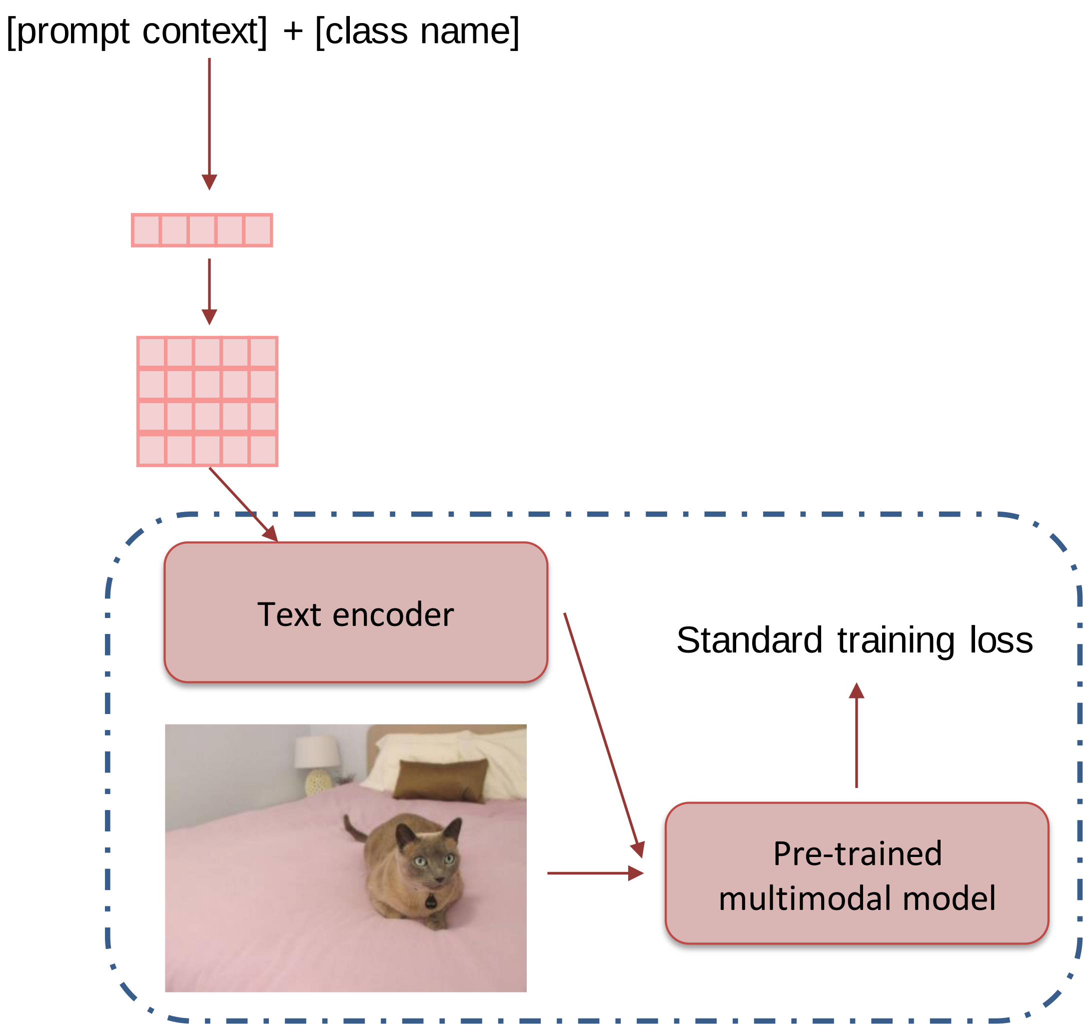
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Learning to Name Classes for Vision and Language Models

Sarah Parisot, Yongxin Yang, Steven McDonagh

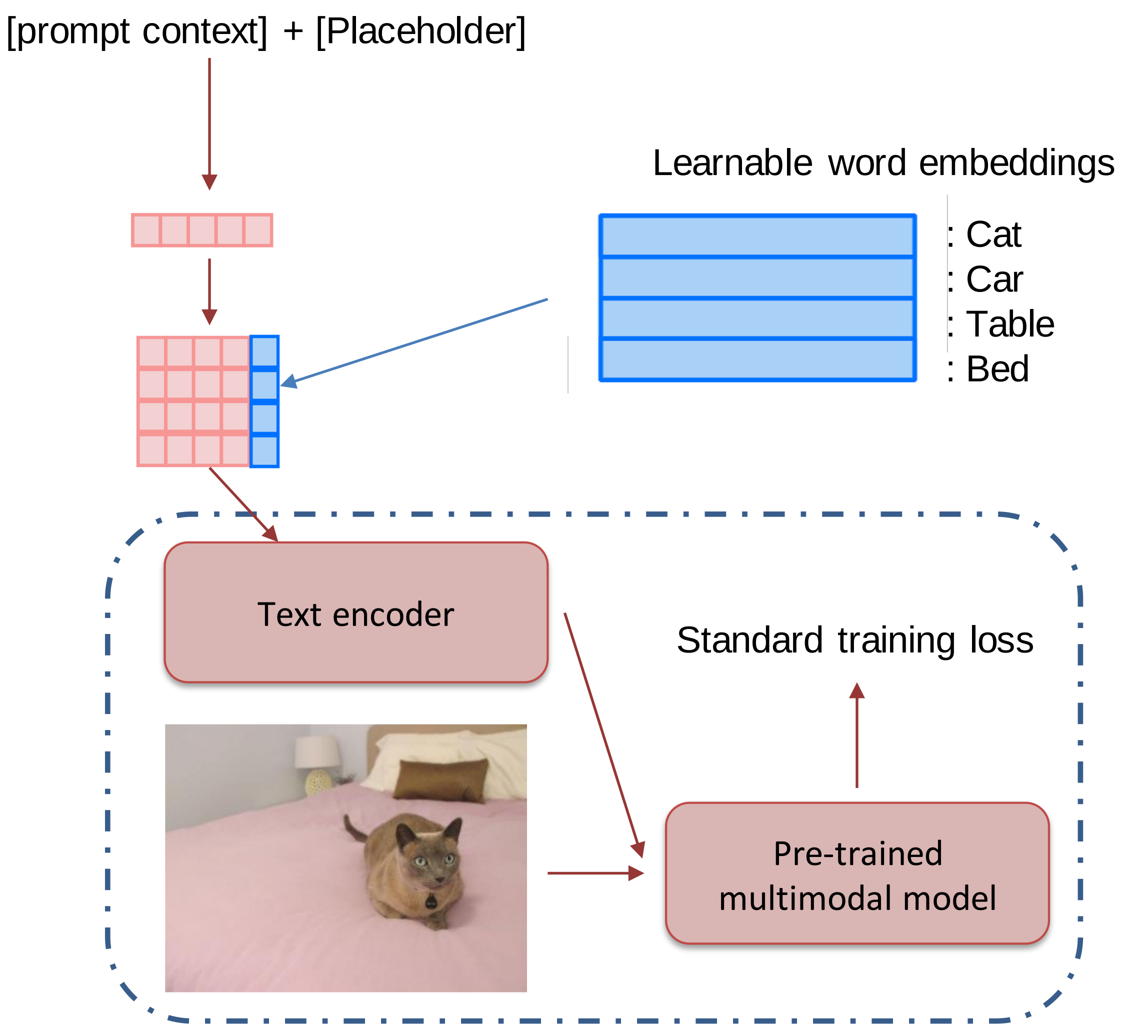
Huawei Noah's Ark Lab

Adapt vision-language models to new dataset by **learning class names**

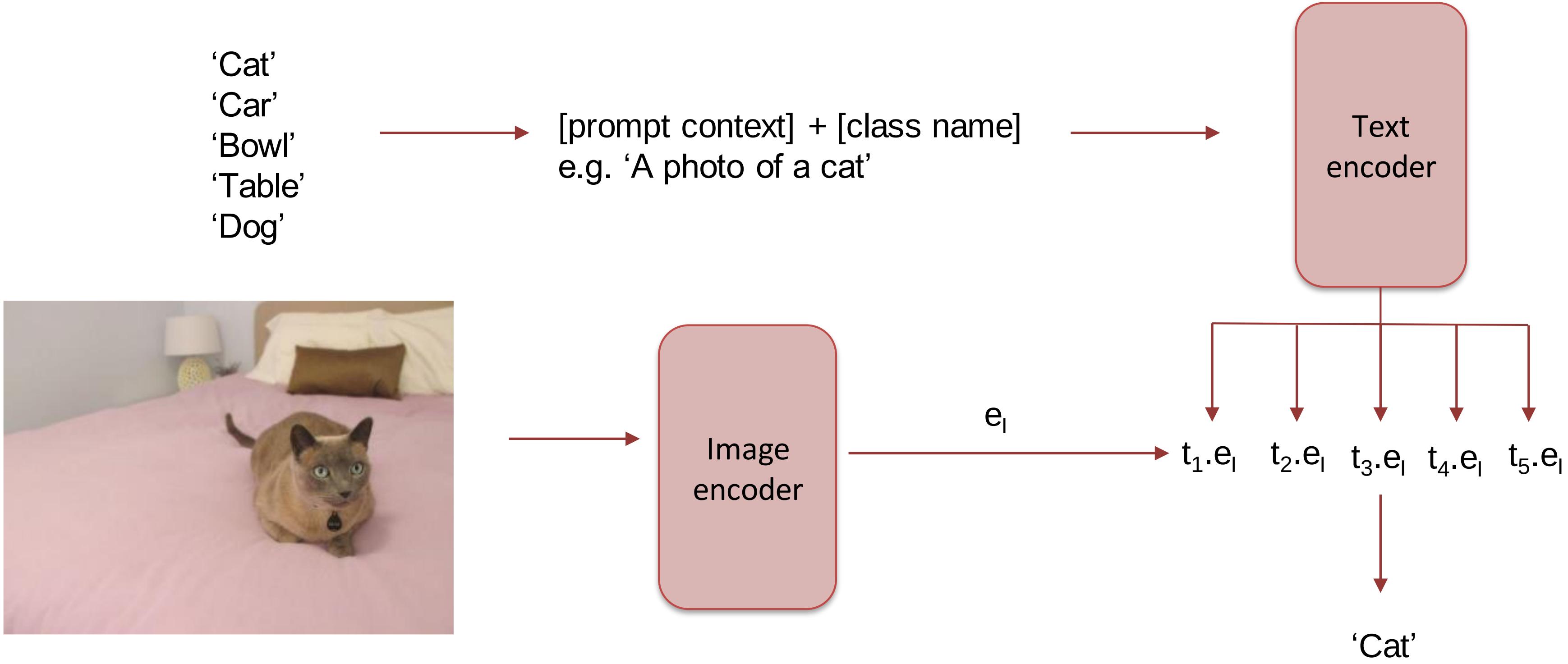


Adapt vision-language models to new dataset by **learning class names**

- Removes class name ambiguities
- Increases robustness to prompt context
- Language agnostic: adapt to model's observed language
- Directly applicable to both classification and object detection tasks

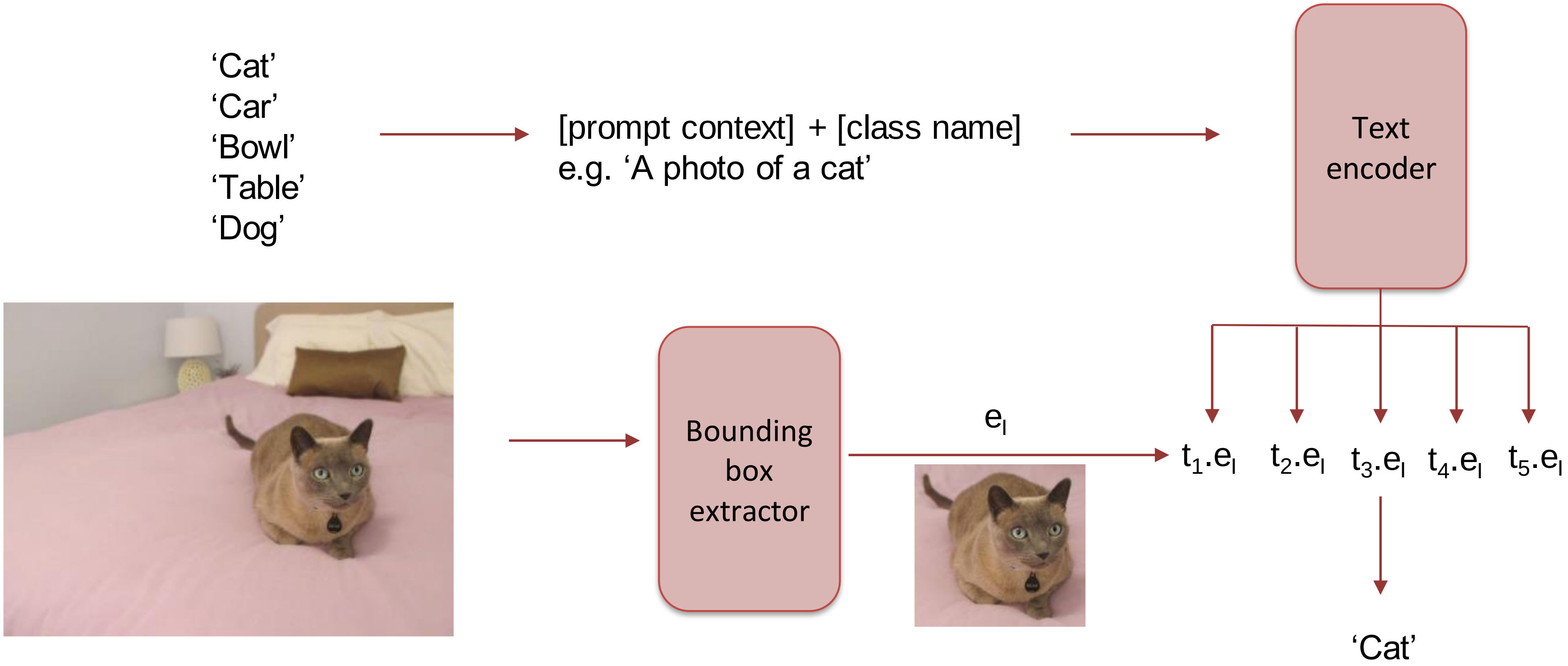


Vision-language classification models



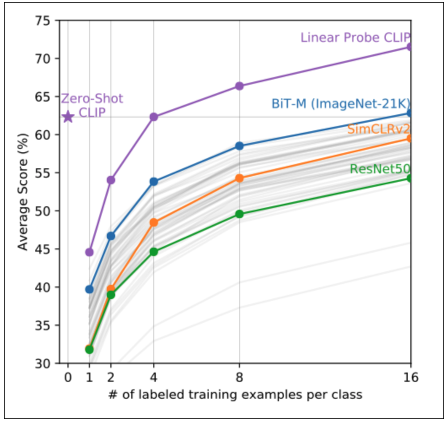
Radford et al. "Learning transferable visual models from natural language supervision." *International Conference on Machine Learning*. PMLR, 2021.

Vision-language detection models



Minderer et al. "Simple open-vocabulary object detection with vision transformers." *ECCV* 2022

Fine-tuning



- Adapting vision-language models to new data: **challenging!**
  - Small dataset overfitting
  - Losing generalisation ability

- Linear probing
  - Train a standard linear classification layer using frozen image encoder

☑ Data efficient
☑ improves over zero-shot performance
☑ no hand-crafted text components

☒ Loses open-set and zero-shot properties

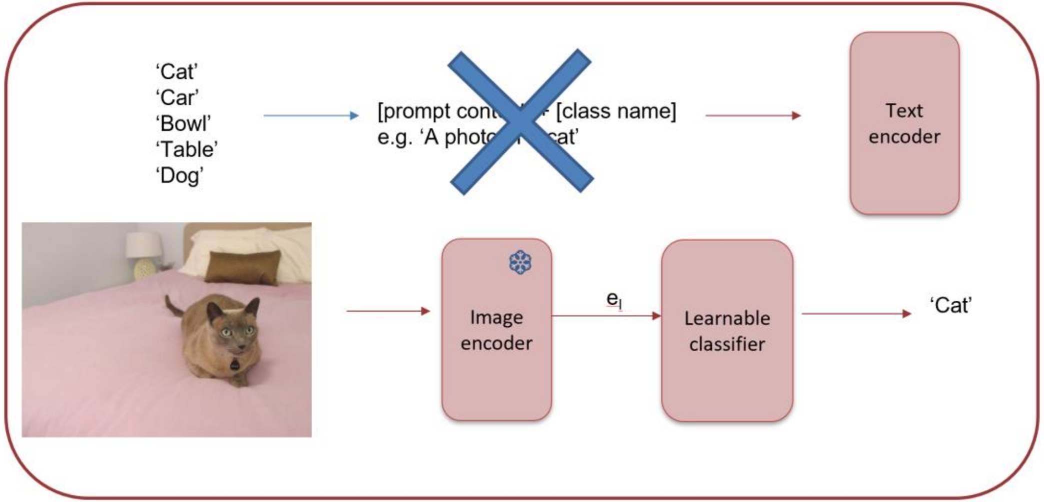
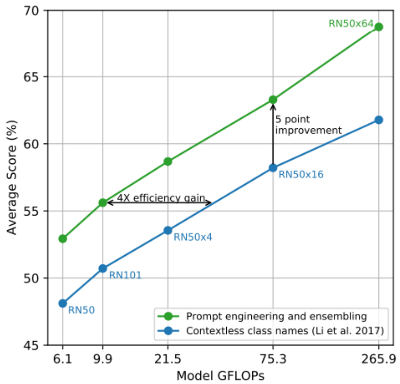


Image source: Radford et al. "Learning transferable visual models from natural language supervision." *International Conference on Machine Learning*. PMLR, 2021.

Sensitivity to prompt input

- Model performance is sensitive to text input



- Existing methods rely on *handcrafted* class names. Potentially:
  - Ambiguous
  - Too technical
  - Unrepresentative of image content

Ambiguous class names

Both named 'bow'

Both named 'bat'

Technical class names

Class name:
2007 Cadillac Escalade EXT Crew Cab

Class name:
A340-200

Image source: Radford et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

Prompt context learning

- Learn prompt context word embeddings (frozen vision-language)

Data efficient

☑ Improves over zero-shot performance
☑ Address prompt sensitivity limitations
☑ Maintain open-set properties

☒ Relies on handcrafted class names
☒ Difficult continual adaptation
☒ Weak object detection performance

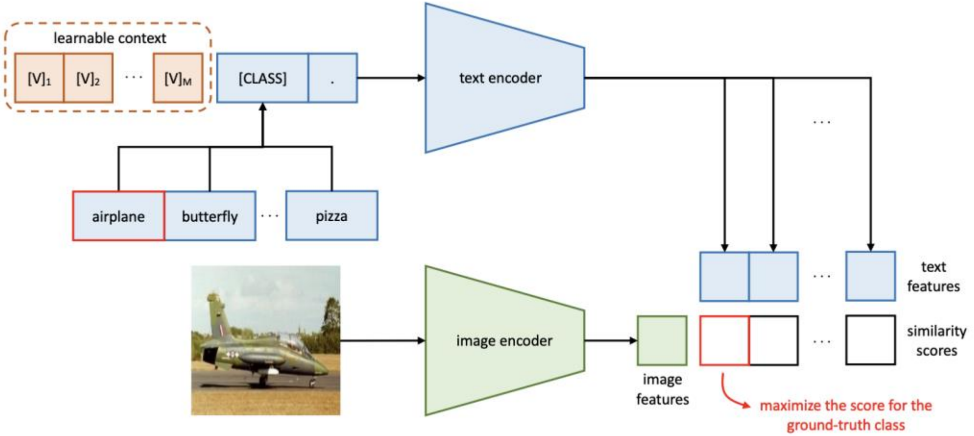
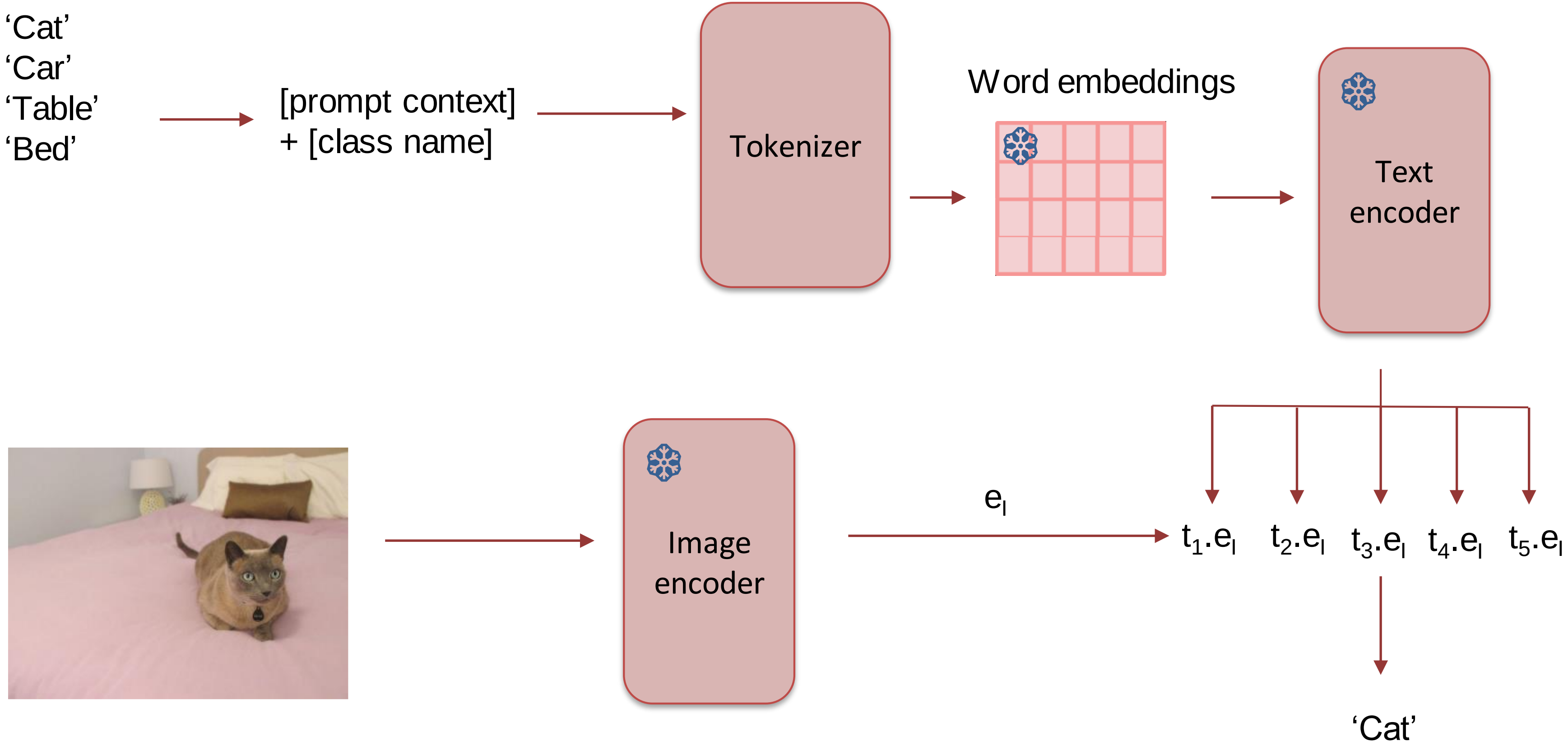


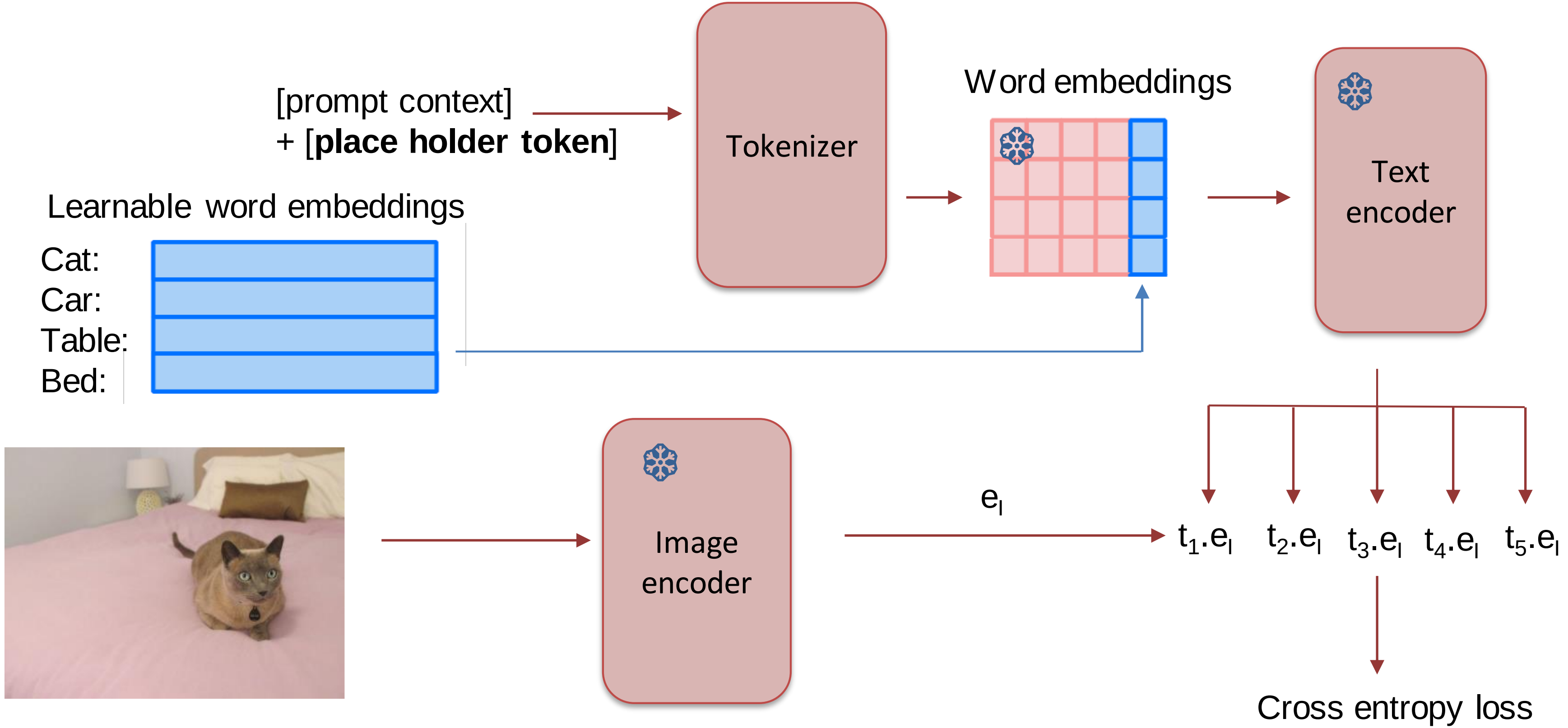
Figure 2: Overview of context optimization (CoOp).

Zhou et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision* (2022)

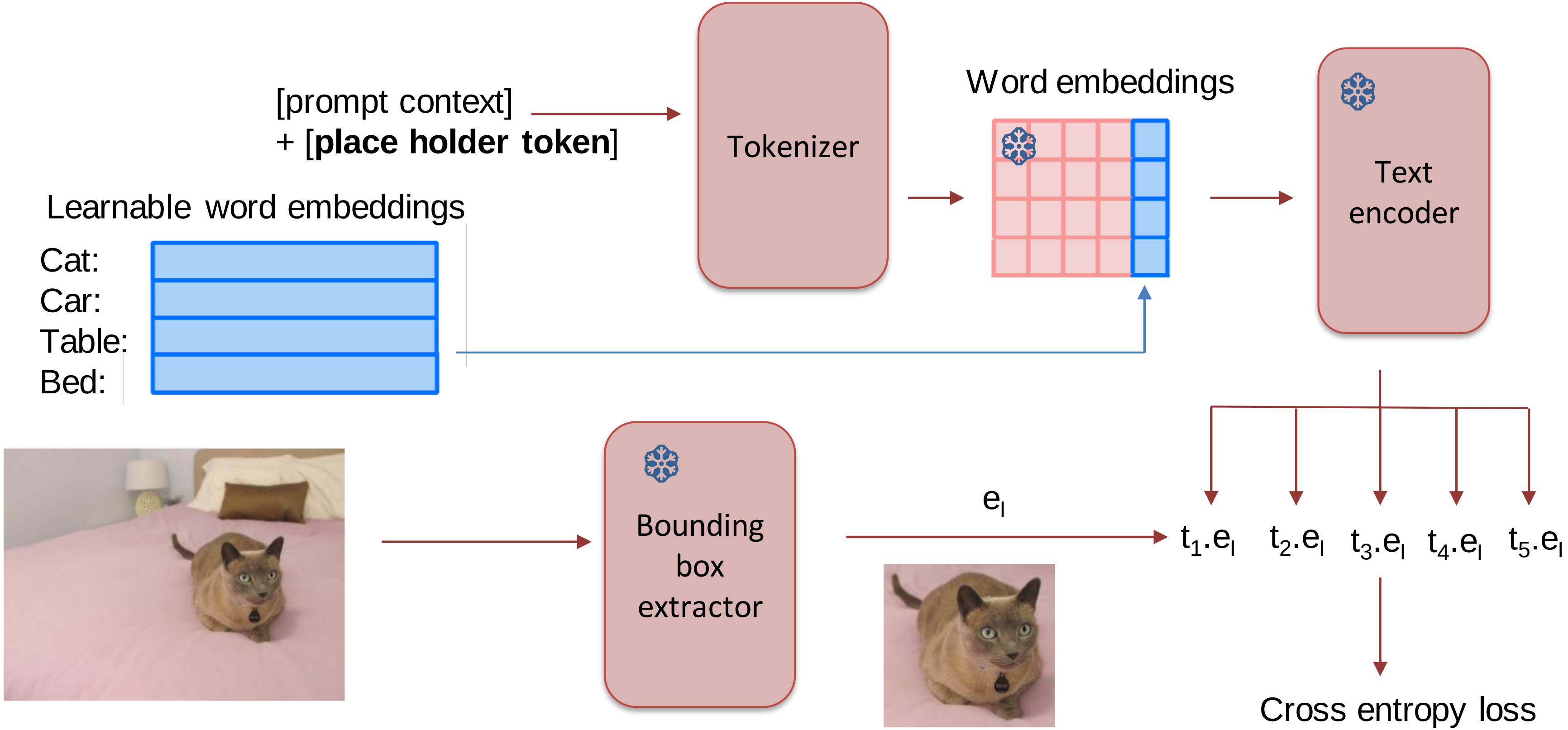
Proposed solution



Proposed solution



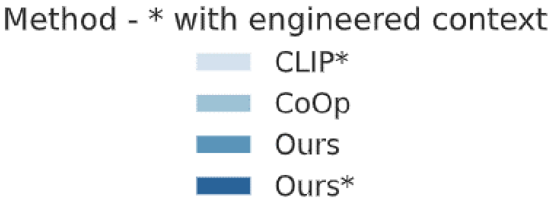
Proposed solution



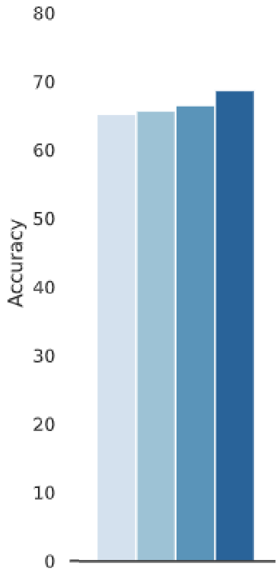
Experiments: classification with CLIP

- Outperforms SOTA in open-vocabulary and sequential training settings

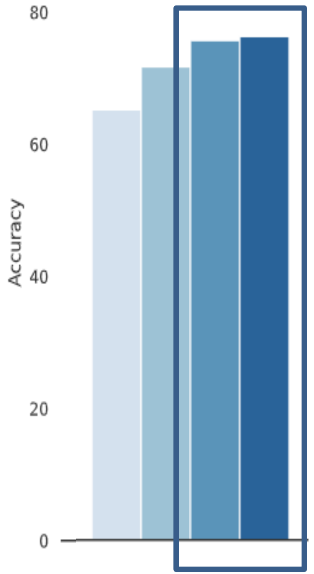
- Learning all class names strongly reduces dependency on prompt context



Open-vocabulary setting: learning half of the dataset class names



Sequential training setting: learning two sets of class names sequentially

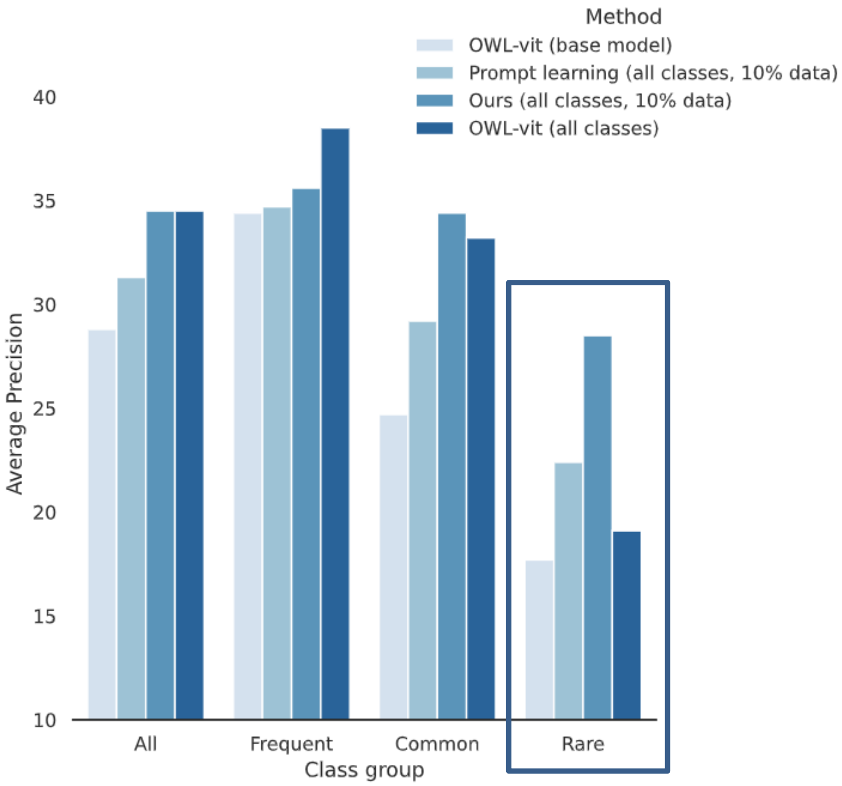


Experiments: Object detection with OWL-vit

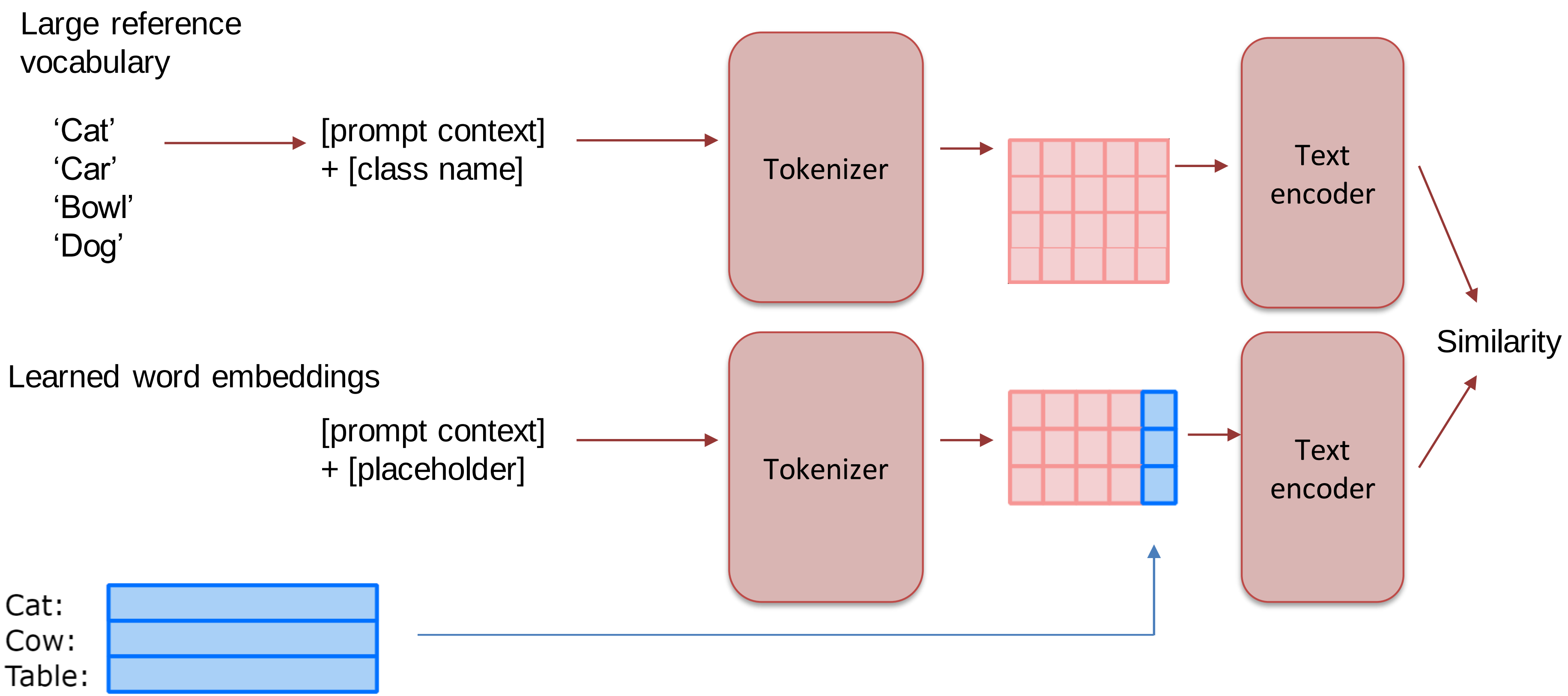
- Learning class names (10% of data) – match performance of fully fine-tuned model

- Significant performance improvement for rare classes

- Significant gains compared to prompt context learning



Interpretability



Interpretability

Original name:
Arctic

Boot, ski boot

Original name:
Tricycle

Cart, rickshaw

Original name:
Miscellaneous

wheel, waterwheel

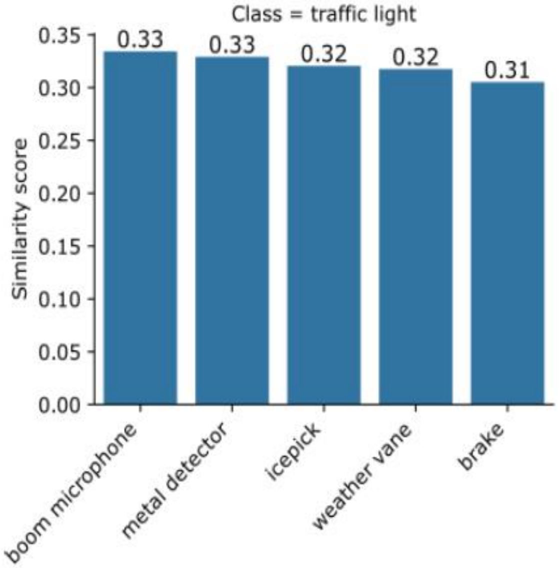
Interpretability

Identifying model biases: American English over British English

Original name:	*Original name*:	*Original name*:	*Original name*:
Clothes hamper	Wall socket	Postbox	Trousers
↓	↓	↓	↓
Laundry basket	Power outlet	Mailbox	Clothes, Pants

Interpretability

Potential to identify mislabelled data and failures modes of our method



class examples

Conclusion

- Novel data efficient adaptation for vision-language models
  - Removes dependency on hand-crafted class names
  - Learn optimal class word embeddings from visual content

- Out of the box usage on classification, detection models

- Complementary to prompt context learning methods

- High interpretability

HUAWEI

THU-PM-274

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Learning to Name classes for Vision and Language Models

Sarah Parisot, Yongxin Yang, Steven McDonagh

Huawei Noah’s Ark Lab

Paper: