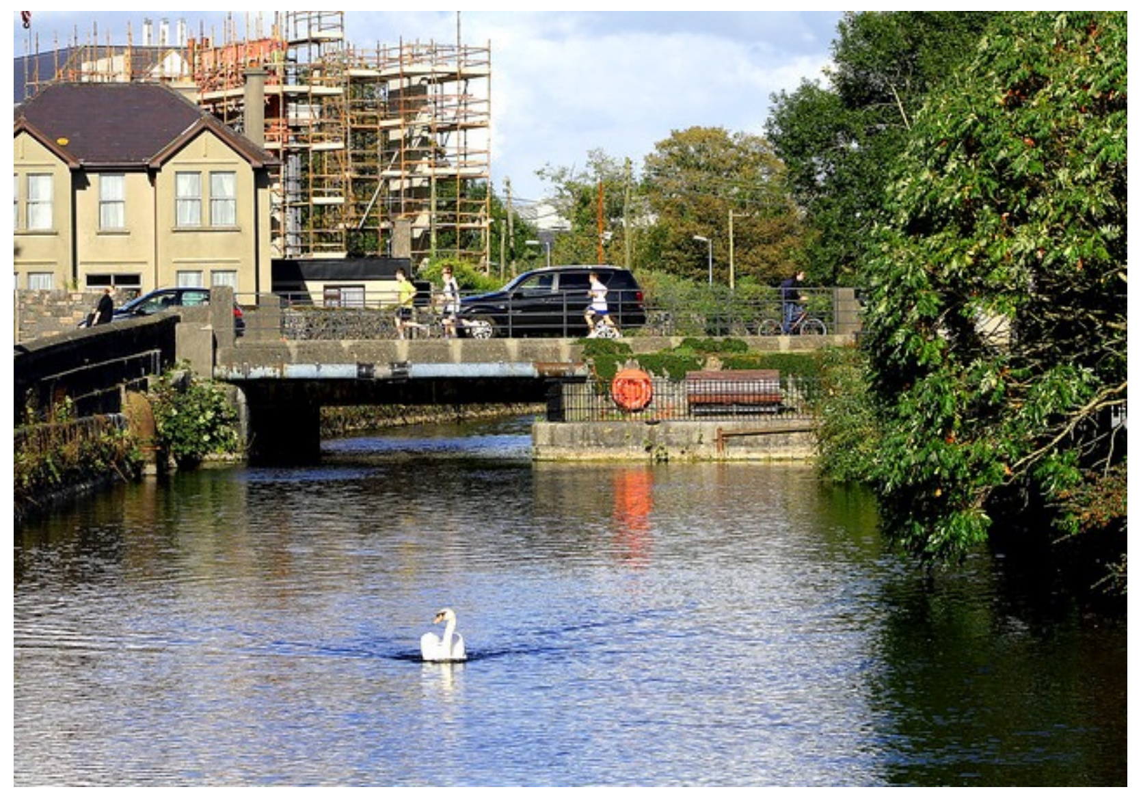
Self-Supervised Multimodal Representation Learning

Gabriel Rasskin Advisors: Paul Liang, Louis-Philippe Morency

Carnegie Mellon University

Problem: Data is Messy

Real-world data is multimodal and unlabeled. How can we learn useful representations?

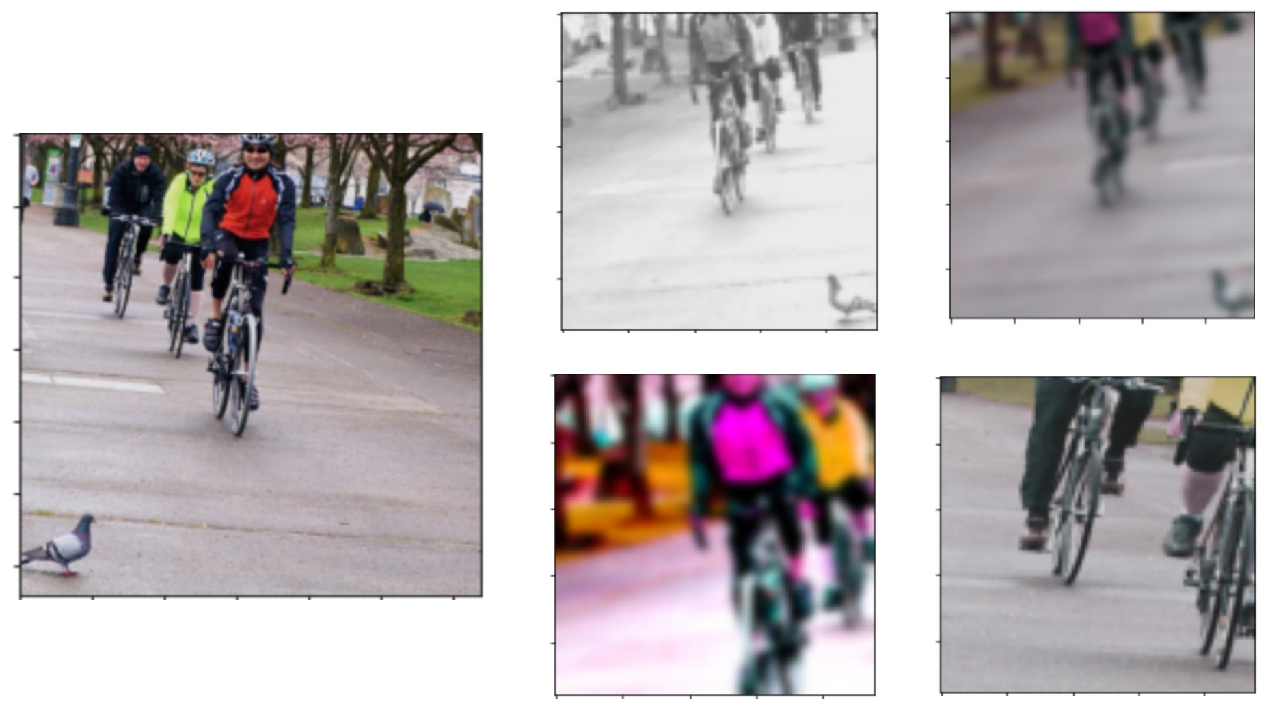


- A lone swan swims in a river near a bridge.
- One swam [sic] in the middle of the water, as car [sic] pass over the bridge.
- A ducks [sic] is swimming on the water while people walk in the background

Figure 1. Image and captions COCO #62276

Background Work: Multimodal + Contrastive Learning

Intuition: Maximize agreement between representations of the same objects, maximize disagreement with other objects.



Augmentations:

- **Resizing Crop**
- **Color transform**
- **Grayscaling**
- **Gaussian Blur**

Figure 2. Image augmentations

Caption Augmentation: Treat image captions as interchangeable. Current approaches work at the word level rather than the sentence level.

Dealing with memory and compute constraints made it necessary to precompute all sentence encodings, creating a “virtual” text head.

Novel Architecture: BERTSimCLR

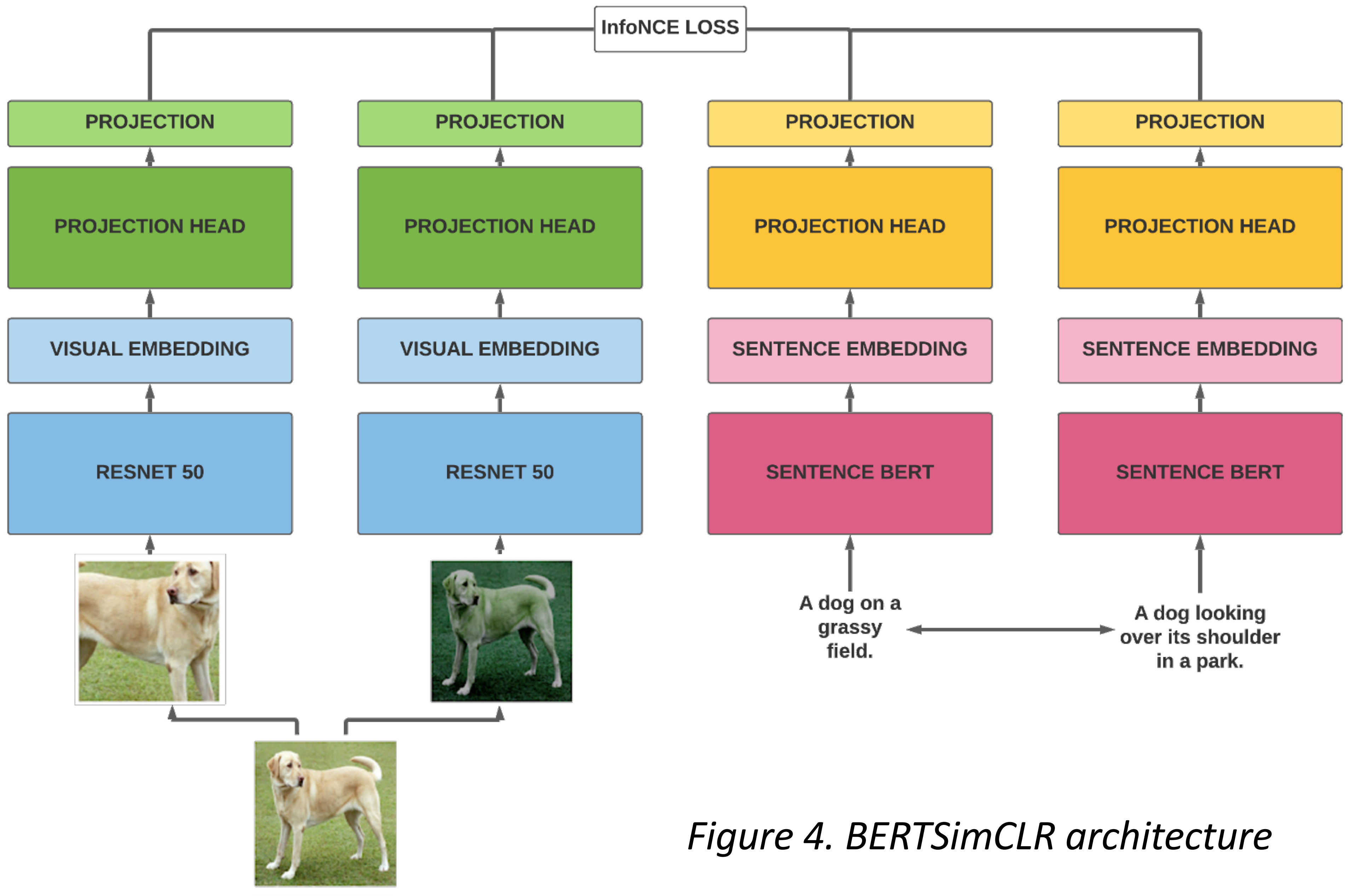


Figure 4. BERTSimCLR architecture

Ablation: Validation loss with and without text augmentation on 20% of MSCOCO

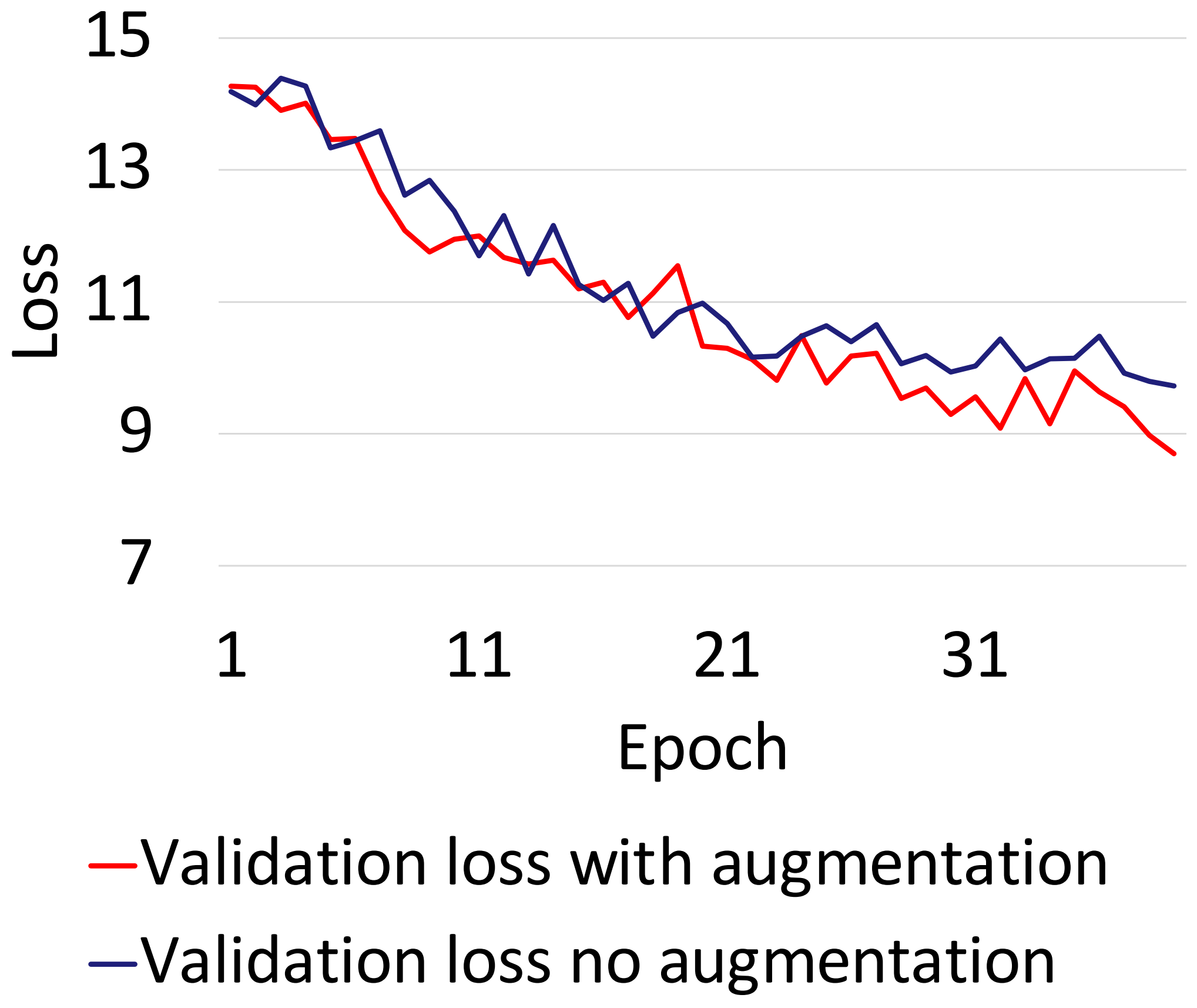


Figure 3. Text augmentation ablation

We’ve found that text augmentation accelerates training.

Linear Evaluation Ablation Testing on 20% of MSCOCO

Top 1 Accuracy. Best value bolded.

Method	With Text Augmentation	Without Text Augmentation	SimCLR Baseline
CIFAR10	**61.6%**	60.8%	48.8%
CIFAR100	**28.7%**	26.4%	17.4%

Top 5 Accuracy. Best value bolded.

Dataset	With Text Augmentation	Without Text Augmentation	SimCLR Baseline
CIFAR10	**97. 2%**	97.1%	93.1%
CIFAR100	**61.5%**	58.5%	43.2%

We perform linear evaluation on our ResNet50 after pretraining with BERTSimCLR. Our method shows improvements on unimodal baseline (SimCLR) and no text augmentation.

Downstream Linear Evaluation Results

Dataset	Top 1 Acc	Top 5 Acc
CIFAR10	36%	85%
CIFAR100	14%	37%

Other Results

COCO captions encoded by SentenceBERT. Reduced memory footprint and sorted multi-GPU compatibility issues.

Maximize GPU utilization: gradient accumulation, mixed precision, multi-GPU, multi-machine, Pytorch lightning.

Architecture exploration: Easier to incorporate contrastive loss for image than text.