How to AI (Almost) Anything

Lecture 6 – Crossmodal Learning

Paul Liang
Assistant Professor
MIT Media Lab & MIT EECS

https://pliang279.github.io
ppliang@mit.edu
@pliang279

Assignments for This Coming Week

For project:

- Instructions for midterm assignment posted on piazza
- Midterm report due April 1 (Tuesday), presentations April 3 (Thursday)
- For April 3 (Thursday), class from 1-3pm (we will be flexible when you attend and present)
- Finalized main ideas and experimental setup, have datasets and baseline models working, detailed error analysis, initial progress towards implementing new ideas.

Reading assignment due tomorrow Wednesday (3/19).

This Thursday (3/20): fourth reading discussion on **multimodal interactions**.

1. Ten myths of multimodal interaction
2. Mixture-of-experts fusion

Today's lecture

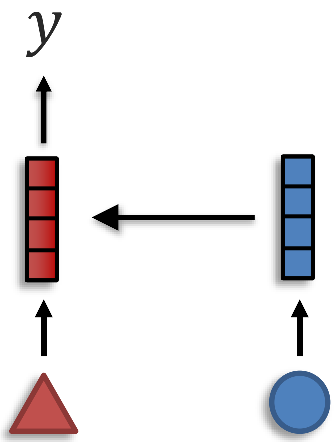
1. Basics of cross-modal transfer
2. Cross-modal transfer via fusion
3. Cross-modal transfer via alignment
4. Cross-modal transfer via translation

Transference

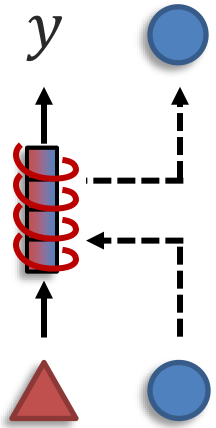
Definition: Transfer knowledge between modalities, usually to help the primary modality which may be noisy or with limited resources

Sub-challenges:

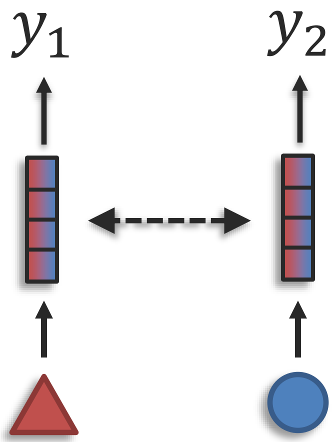
Transfer



Co-learning

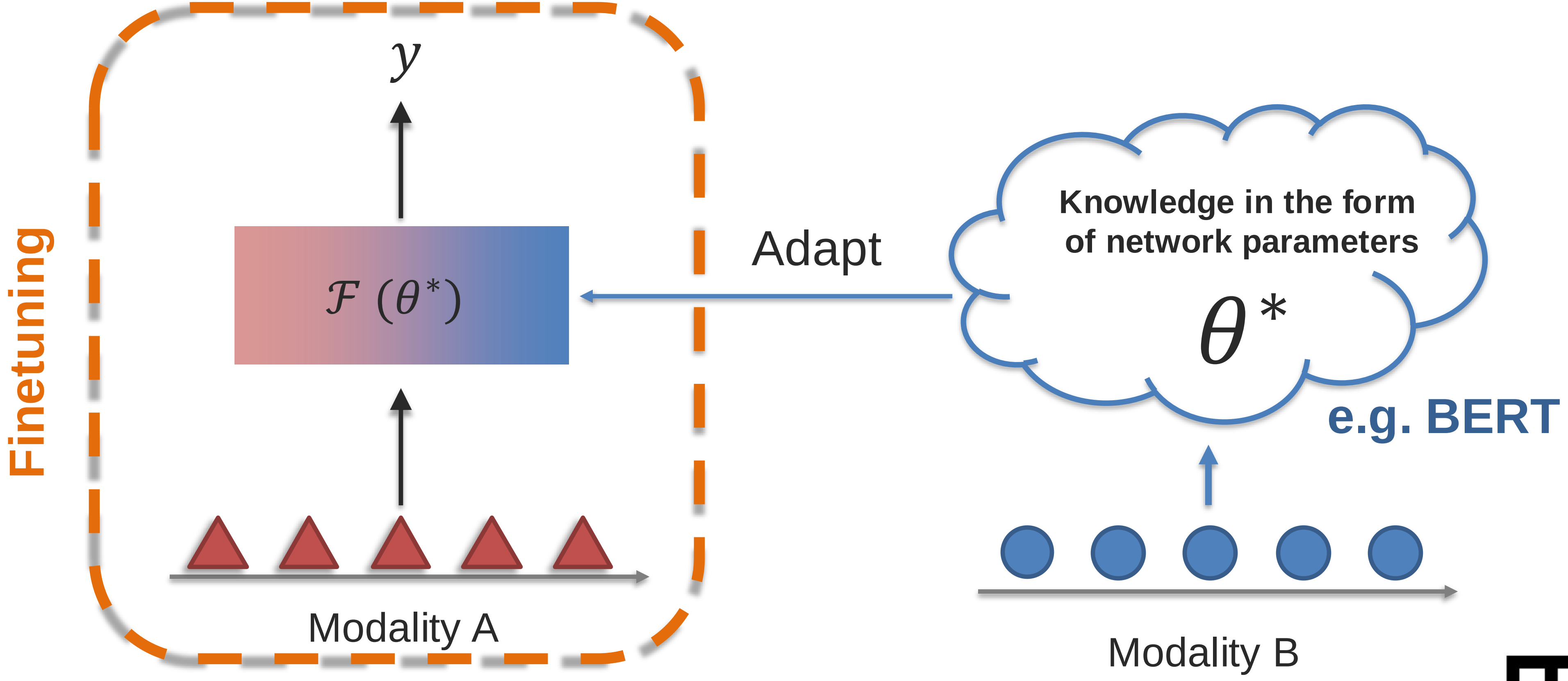


Model Induction



Part 1: Transfer via Pretrained Models

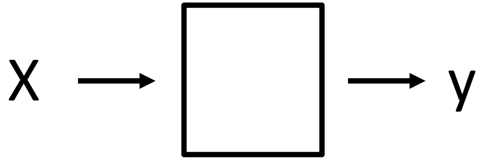
Definition: Transferring knowledge from large-scale pretrained models to downstream tasks involving the primary modality.



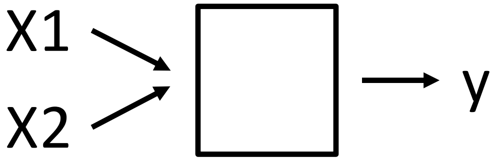
multisensory intelligence

Multitask and Transfer Learning

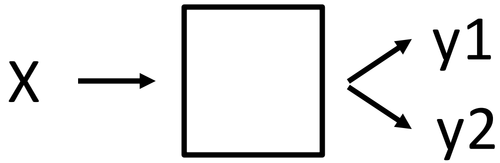
Supervised learning



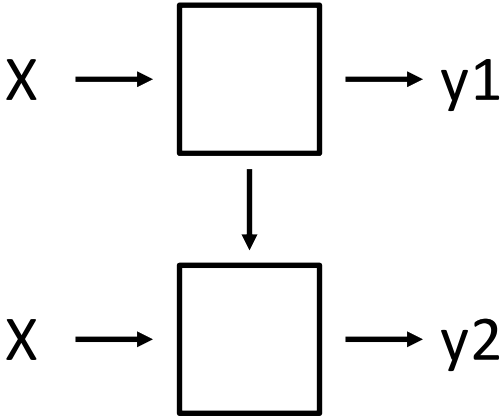
Multimodal (supervised) learning



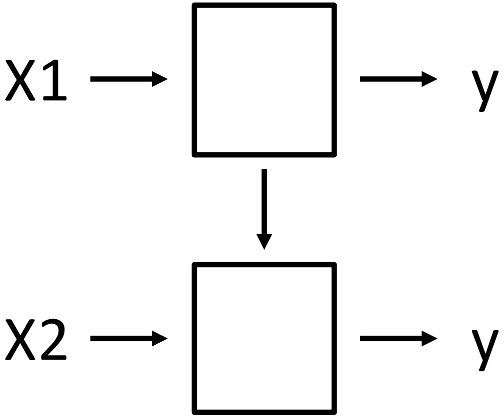
Multitask learning



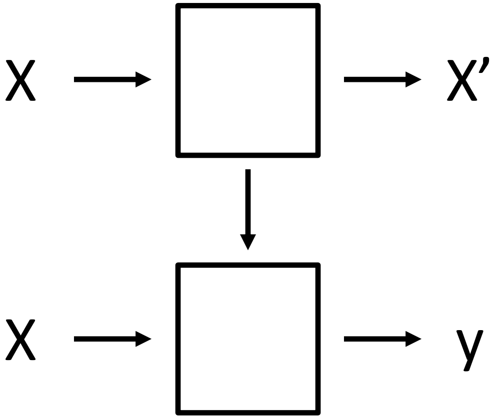
Transfer learning



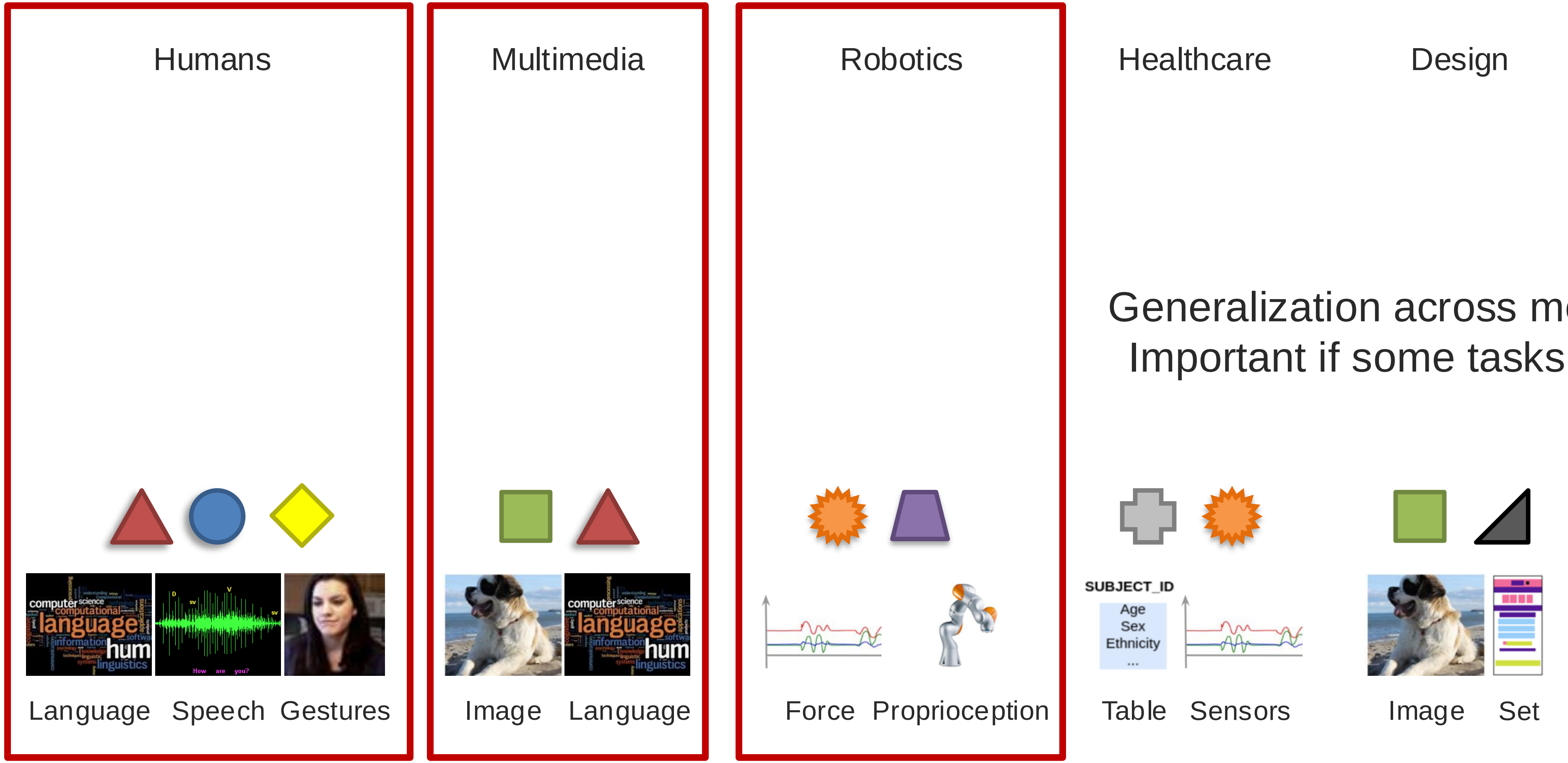
Cross-modal learning



Unsupervised/self-supervised pre-training



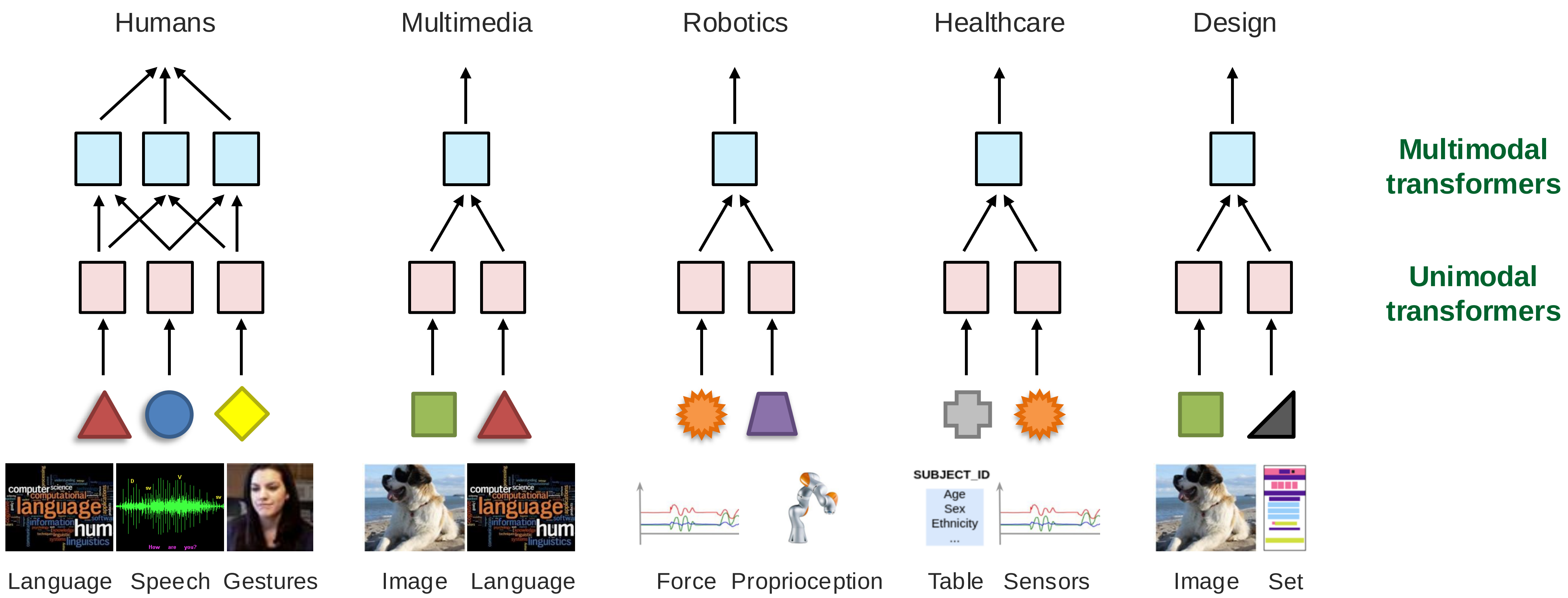
Multitask and Transfer Learning



Generalization across modalities and tasks
Important if some tasks are low-resource

[Reed et al., A Generalist Agent. TMLR 2022]
[Liang et al., HighMMT: Quantifying Modality and Task Heterogeneity for High-Modality Representation Learning. TMLR 2022]

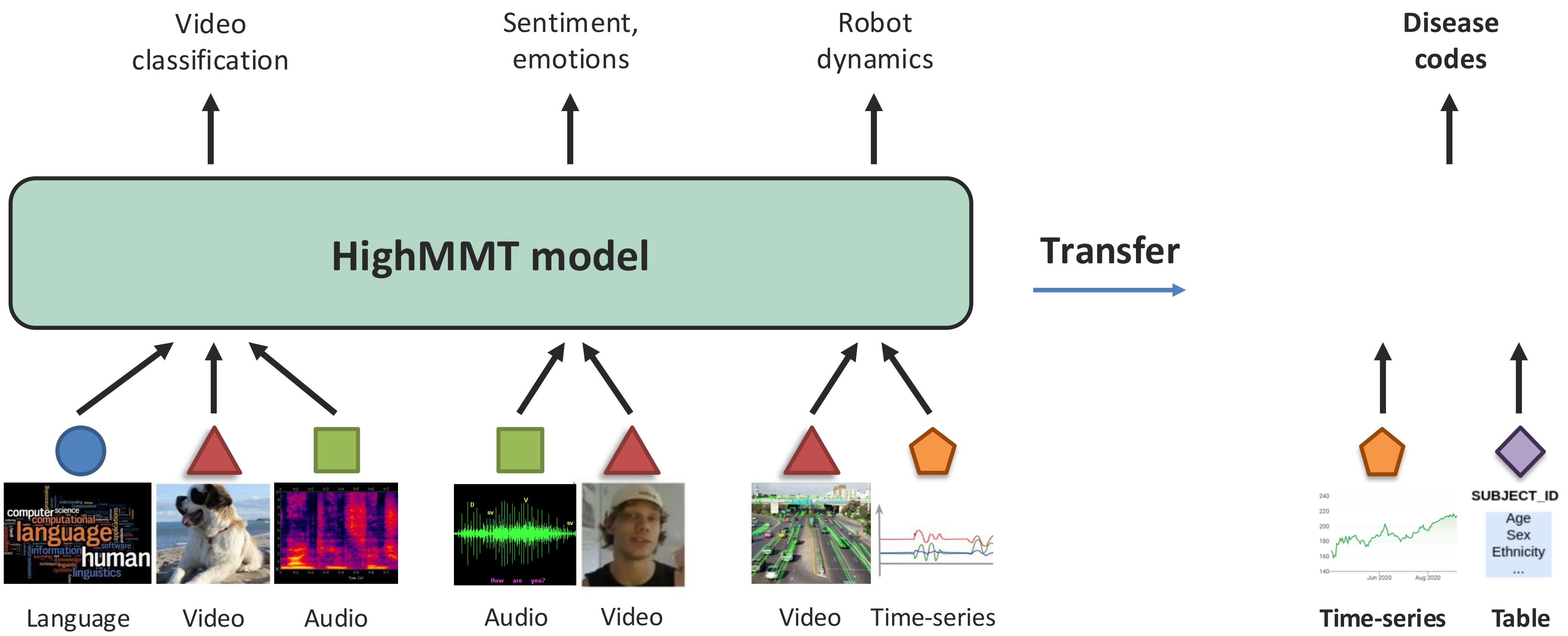
Multitask and Transfer Learning



[Reed et al., A Generalist Agent. TMLR 2022]
[Liang et al., HighMMT: Quantifying Modality and Task Heterogeneity for High-Modality Representation Learning. TMLR 2022]

High-Modality Multimodal Transformer

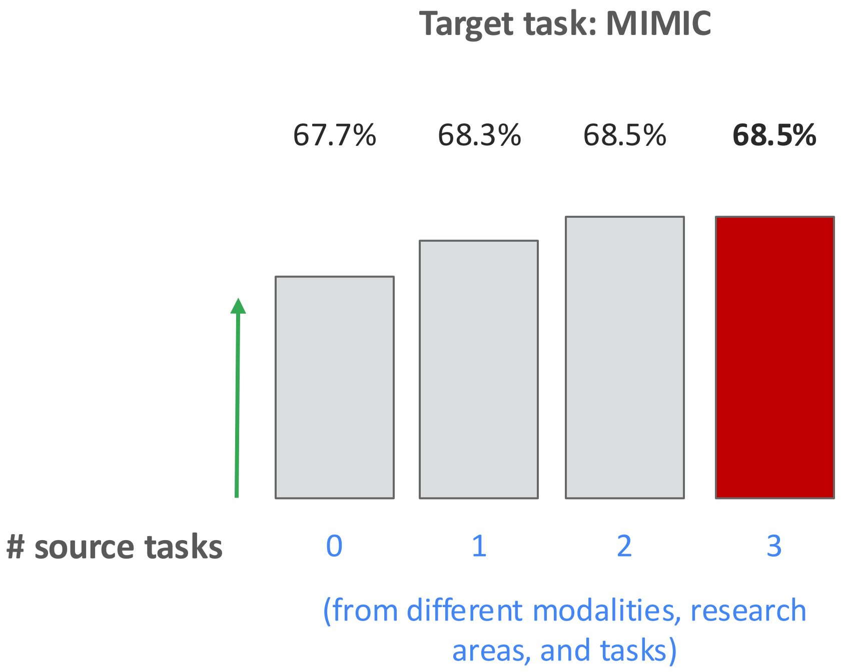
Transfer across partially observable modalities

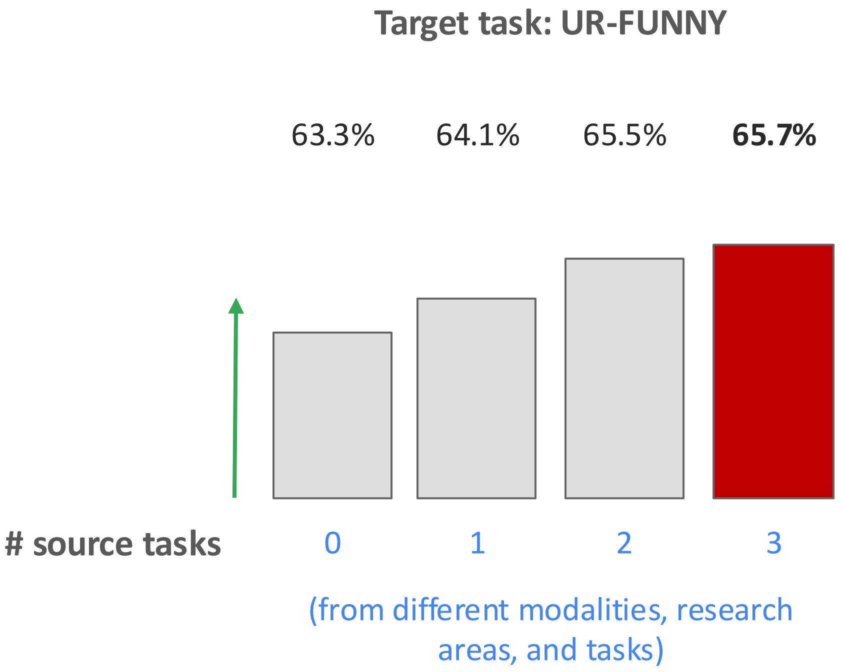


[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022]

Multitask and Transfer Learning

Transfer across partially observable modalities



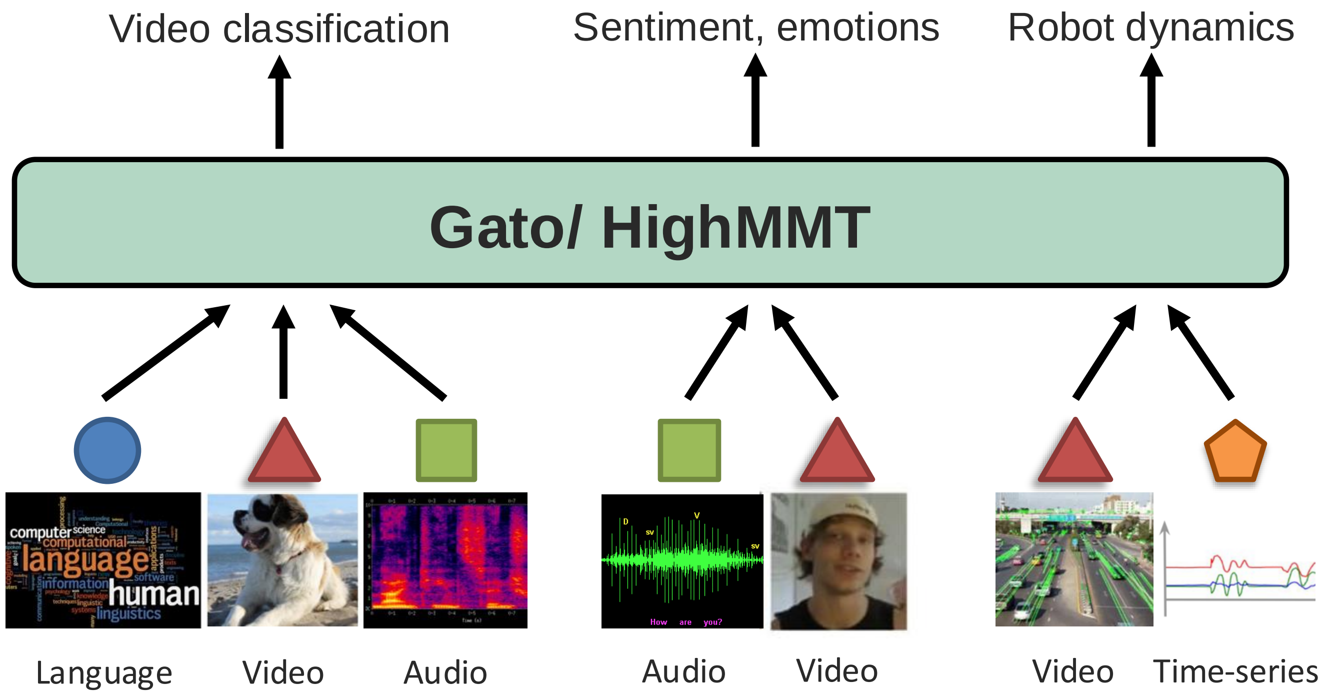


Achieves both multitask and transfer capabilities across modalities and tasks

[Liang et al., HighMMT: Towards Modality and Task Generalization for High-Modality Representation Learning. TMLR 2022]

High-Modality Models

Some implicit assumptions:
- All modalities can be represented as sequences without losing information.
- Dimensions of heterogeneity can be perfectly captured by modality-specific embeddings.
- Cross-modal connections & interactions are shared across modalities and tasks.



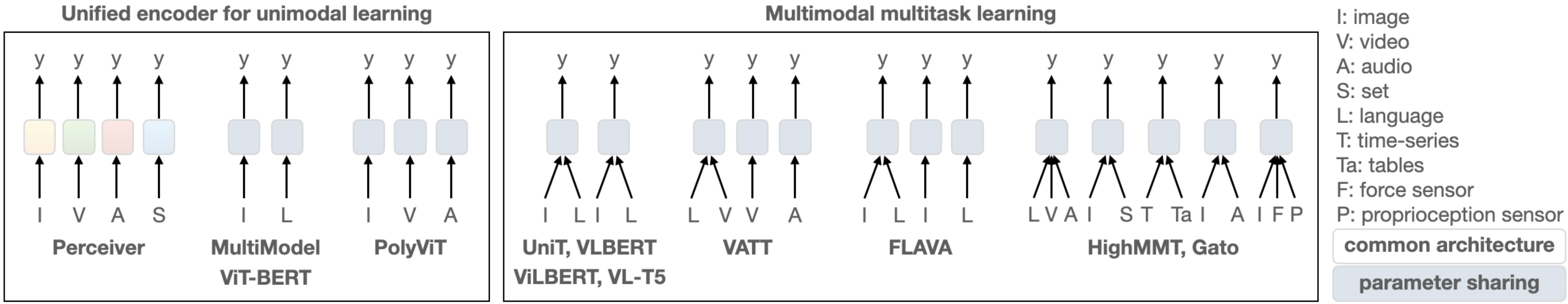
Shared multimodal model?

Modality-specific embeddings?

Standardized input sequence?

Multitask and Transfer Learning

Many more dimensions of transfer



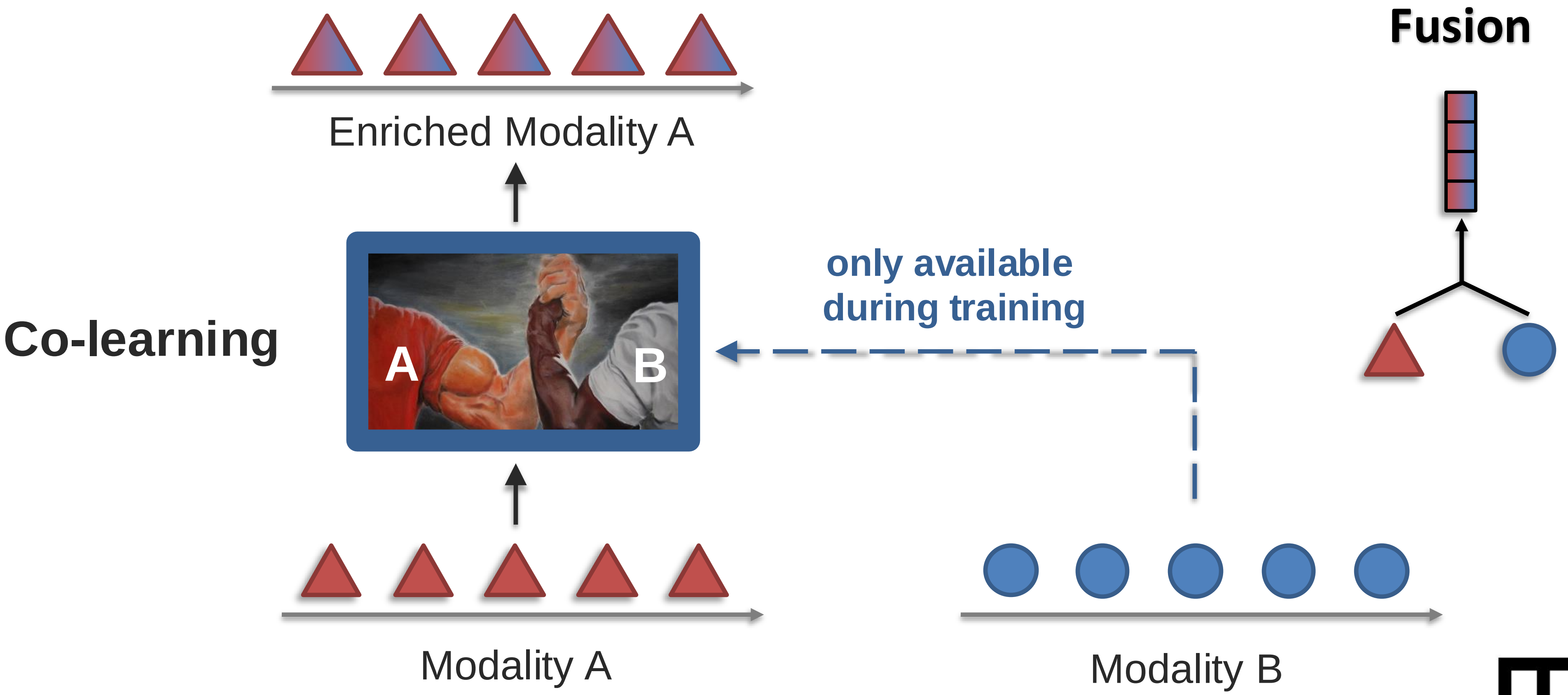
Open challenges:

- Low-resource: little downstream data, lack of paired data, robustness (next section)
- Beyond language and vision
- Settings where SOTA unimodal encoders are not deep learning e.g., tabular data
- Complexity in data, modeling, and training
- Interpretability (next section)

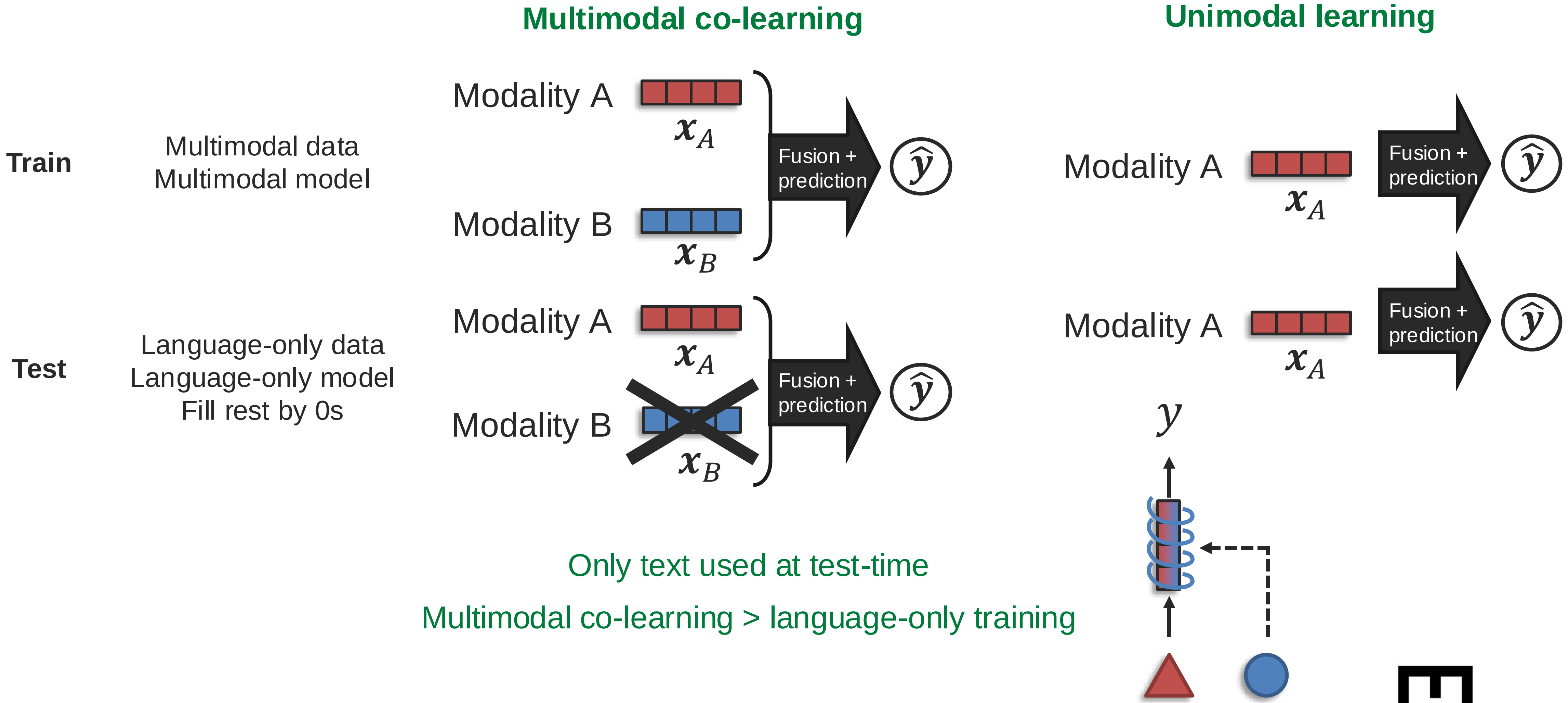
multisensory intelligence

Part 2: Co-learning

Definition: Transferring information from secondary to primary modality by sharing representation spaces between both modalities.



Co-learning via Fusion



Only text used at test-time

Multimodal co-learning > language-only training

[Zadeh et al., Foundations of Multimodal Co-learning. Information Fusion 2020]

Co-learning via Fusion

Generative model (Deep Boltzmann Machine) to learn joint representation and infer missing text.

Model	MAP	Prec@50
Image LDA (Huiskes et al., 2010)	0.315	-
Image SVM (Huiskes et al., 2010)	0.375	-
Image DBN	0.463 ± 0.004	0.801 ± 0.005
Image DBM	0.469 ± 0.005	0.803 ± 0.005
Multimodal DBM (generated text)	$\mathbf{0.531 \pm 0.005}$	$\mathbf{0.832 \pm 0.004}$

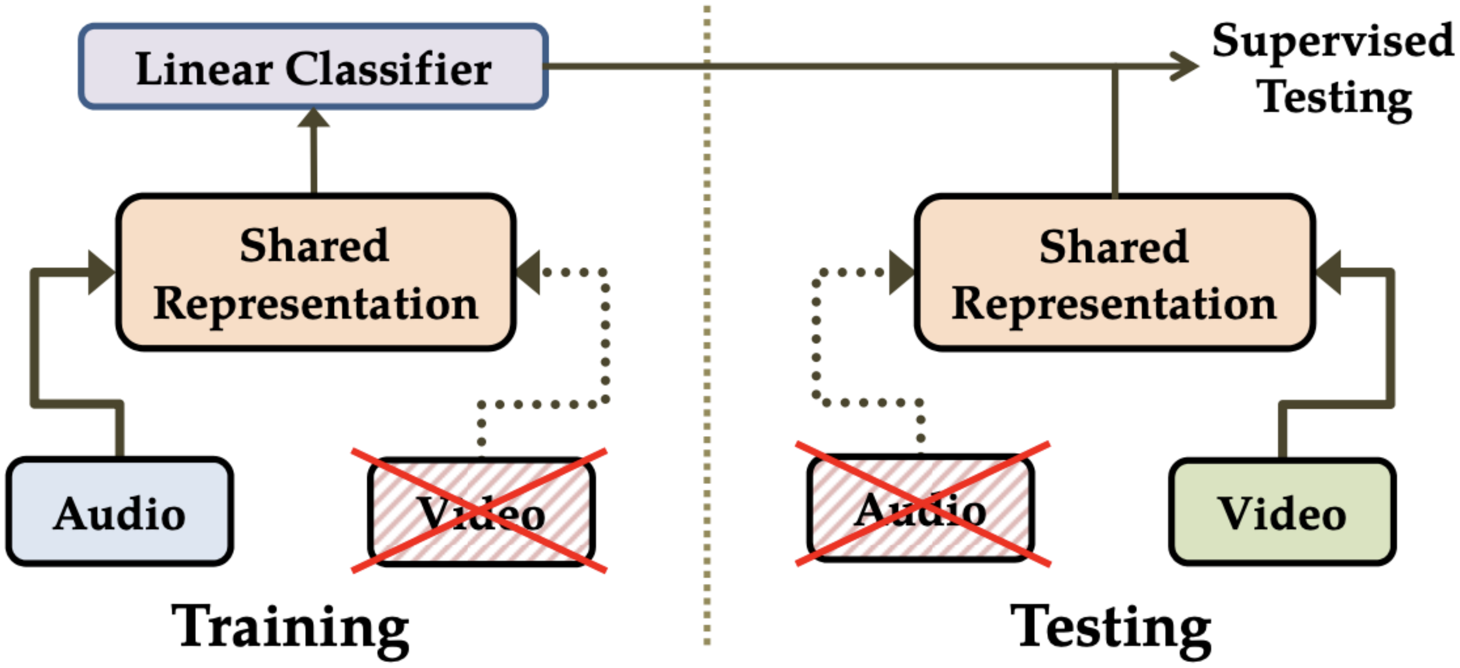
learning multimodal features helps even when some modalities are absent at test time.

[Srivastava and Salakhutdinov, Multimodal Learning with Deep Boltzmann Machines. JMLR 2014]

multisensory intelligence

Co-learning via Fusion

Train on some subset of modalities and test on another

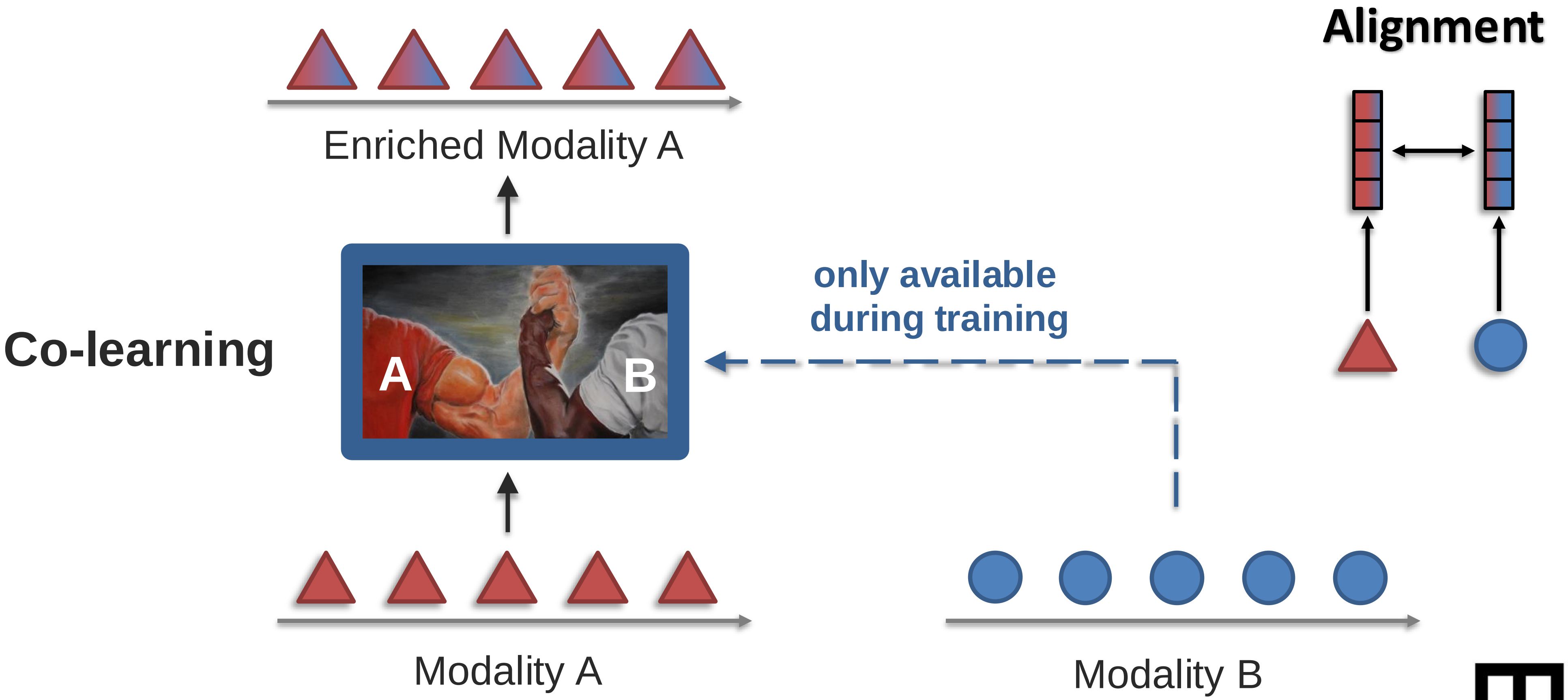


McGurk effect

[Ngiam et al., Multimodal Deep Learning. ICML 2011]

Co-learning via Alignment

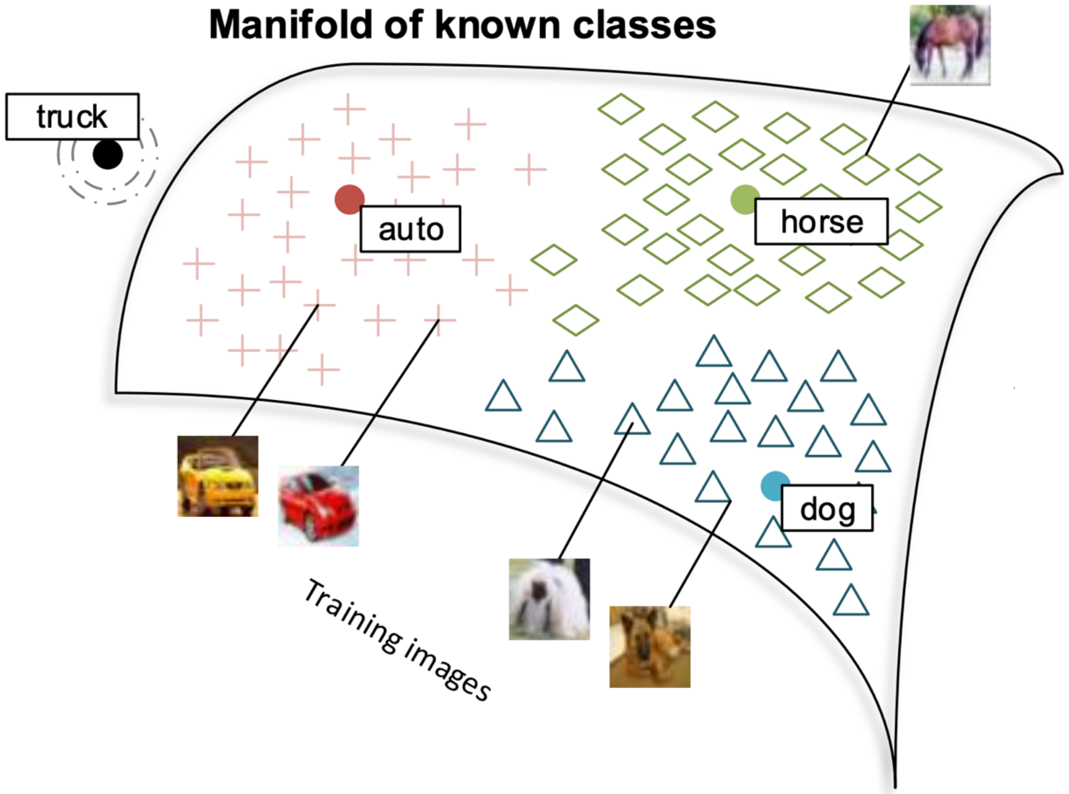
Definition: Transferring information from secondary to primary modality by sharing representation spaces between both modalities.



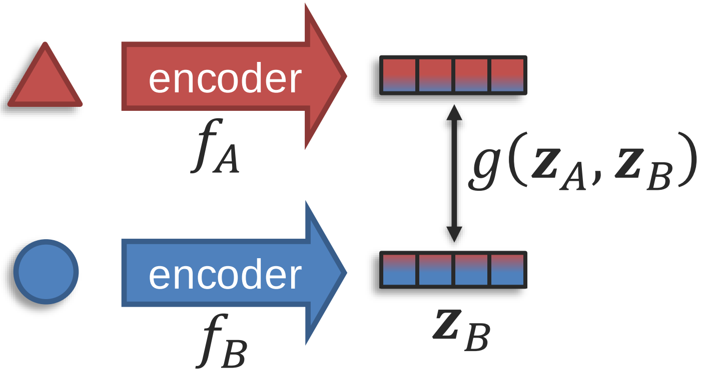
multisensory intelligence

Co-learning via Alignment

Representation alignment: word embedding space for zero-shot visual classification



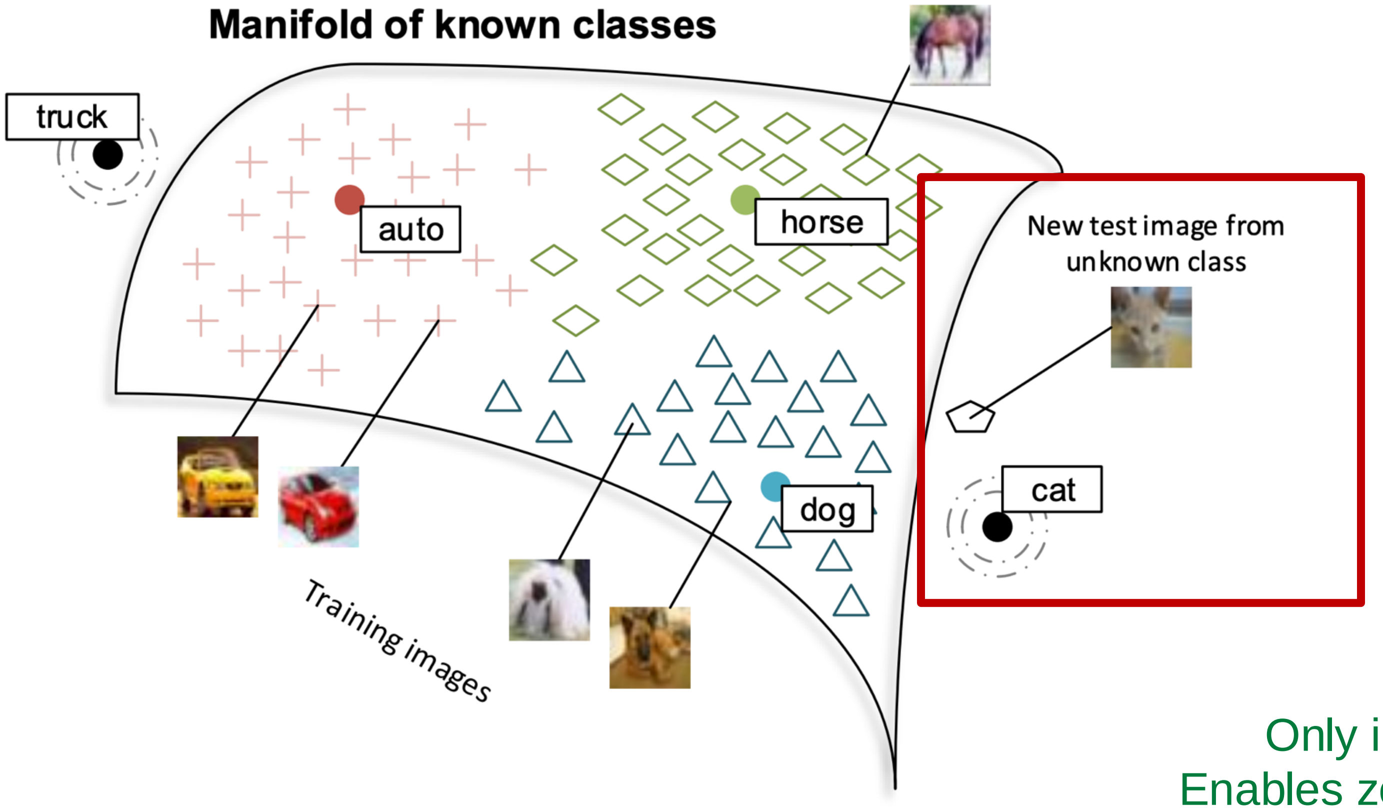
Recall representation alignment!



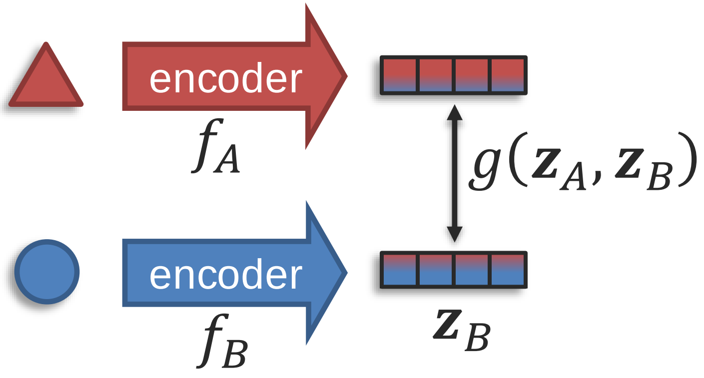
[Socher et al., Zero-Shot Learning Through Cross-Modal Transfer. NeurIPS 2013]

Co-learning via Alignment

Representation alignment: word embedding space for zero-shot visual classification



Recall representation alignment!

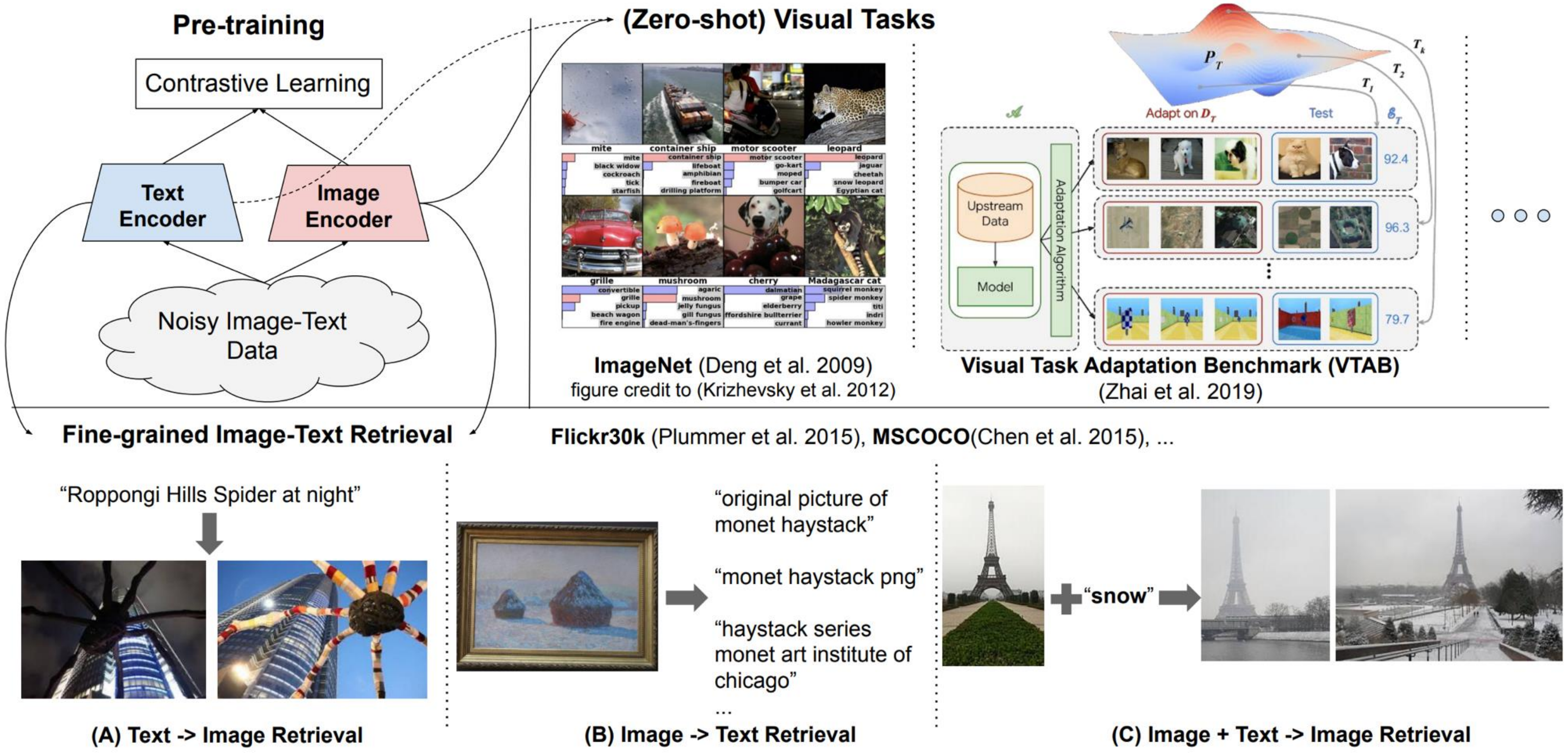


Only images used at test-time
Enables zero-shot image classification

[Socher et al., Zero-Shot Learning Through Cross-Modal Transfer. NeurIPS 2013]

Co-learning via Alignment

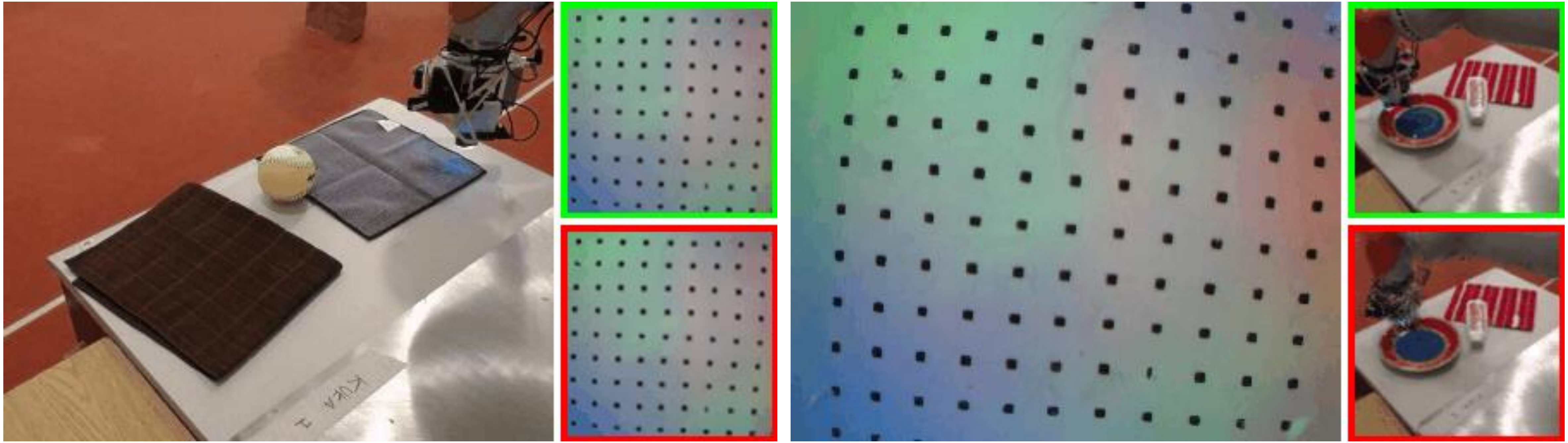
Representation alignment at scale



[Jia et al., Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. ICML 2021]

Vision-Touch Alignment

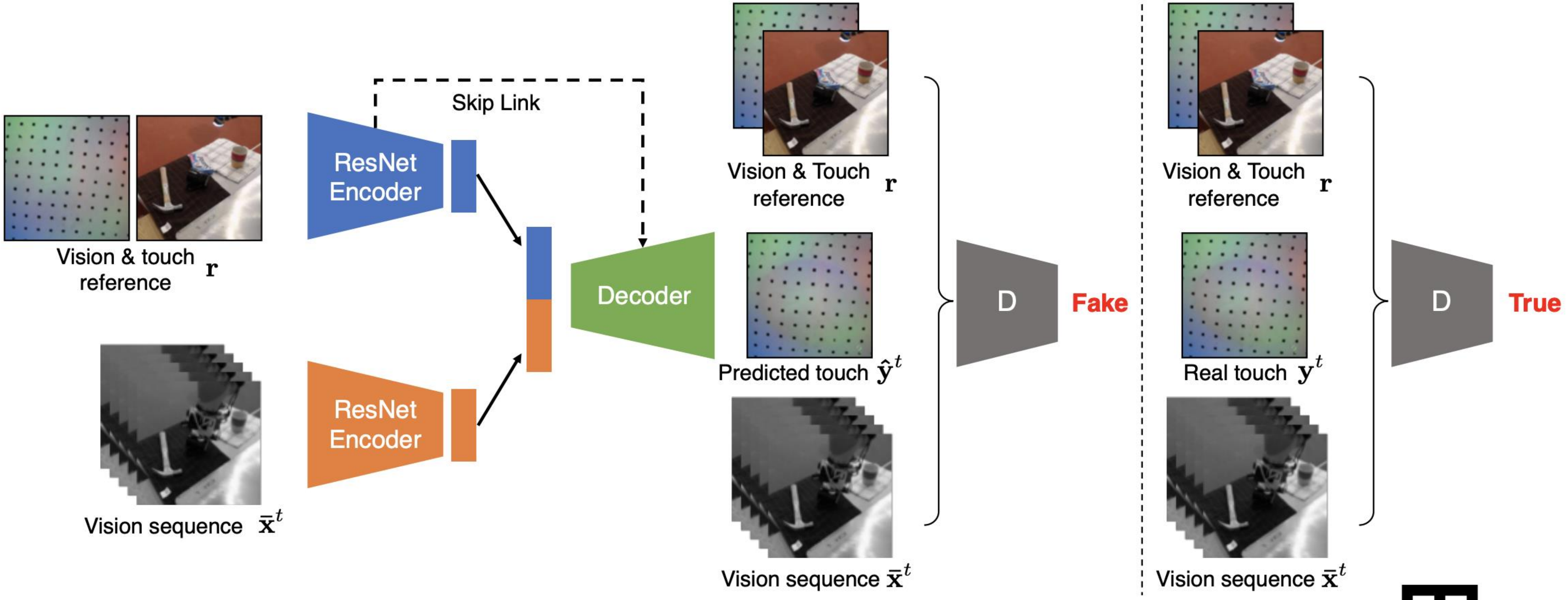
Aligning vision and touch in robotics



[Li et al., Connecting Vision and Touch via Cross-modal Prediction. CVPR 2019]

Vision-Touch Alignment

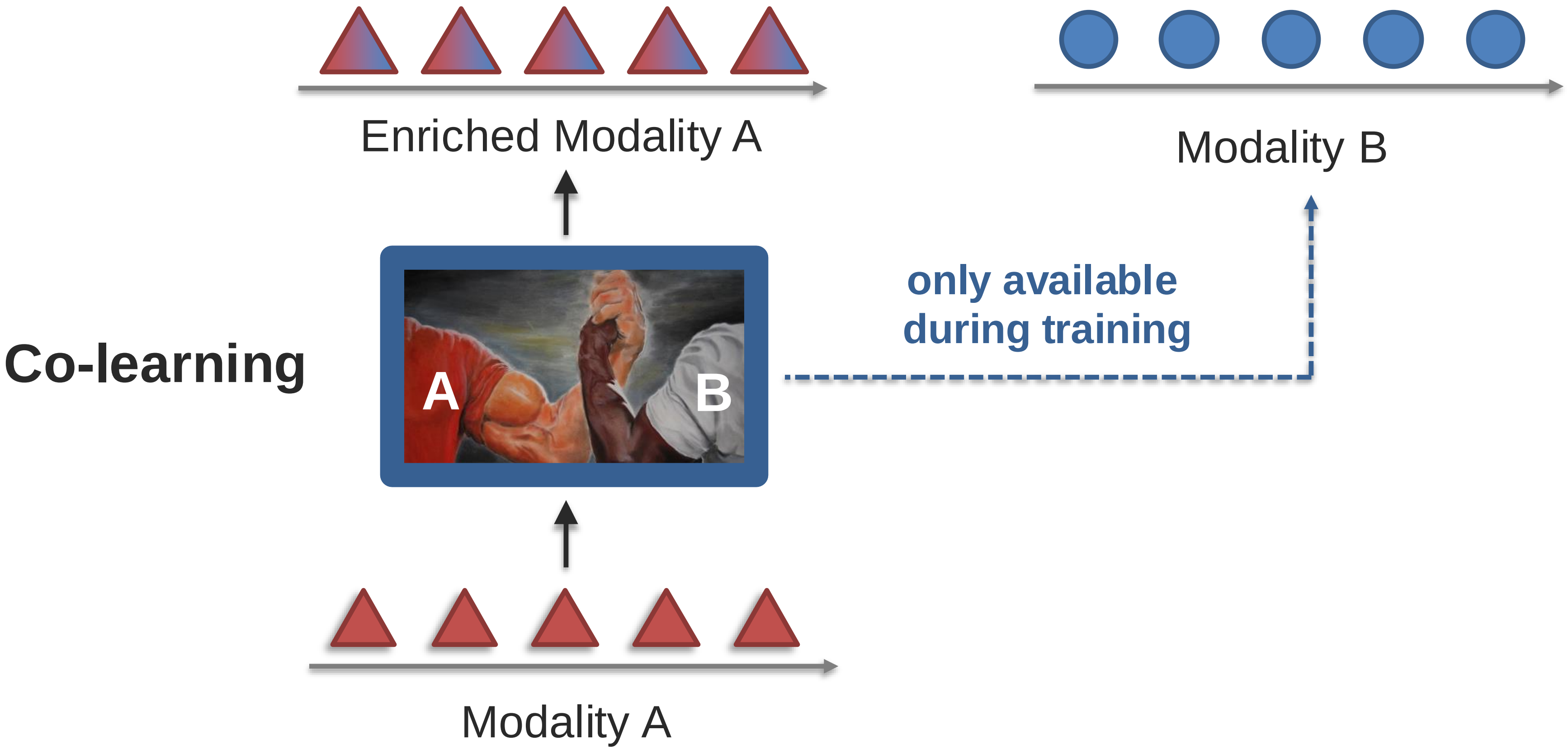
Aligning vision and touch in robotics



[Li et al., Connecting Vision and Touch via Cross-modal Prediction. CVPR 2019]

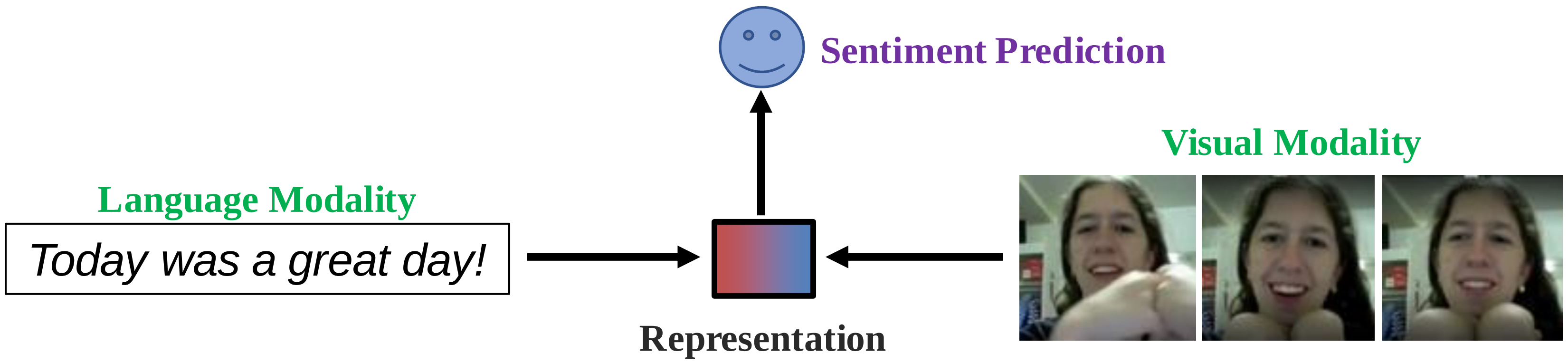
Co-learning via Translation

Definition: Transferring information from secondary to primary modality by using the secondary modality as a generation target.



Co-learning via Translation

Bimodal translations



Both modalities required at test time!
Sensitive to noisy/missing visual modality.

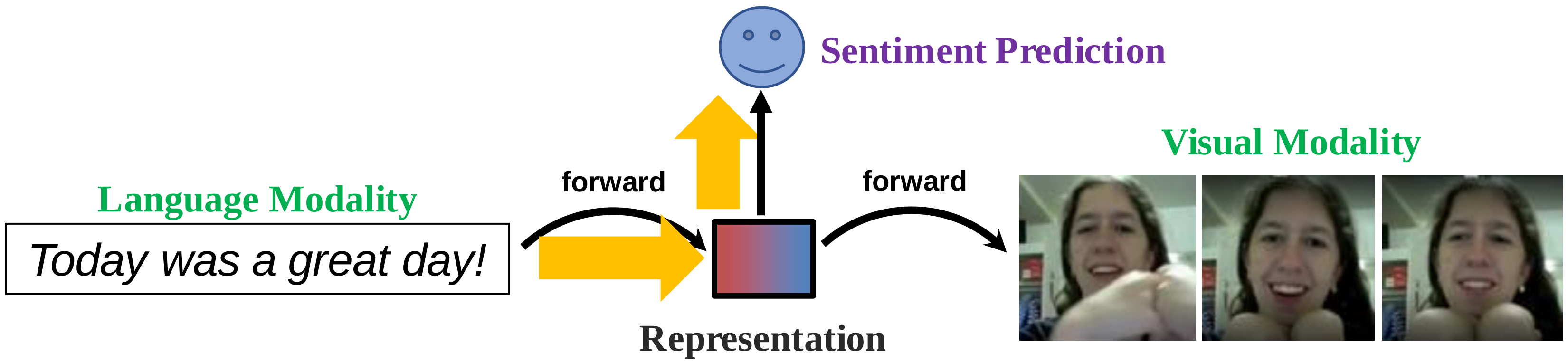
We want to leverage information from visual modality
while being robust to it during test-time.

[Pham et al., Found in Translation: Learning Robust Joint Representations via Cyclic Translations Between Modalities. AAAI 2019]

multisensory intelligence

Co-learning via Translation

Bimodal translations

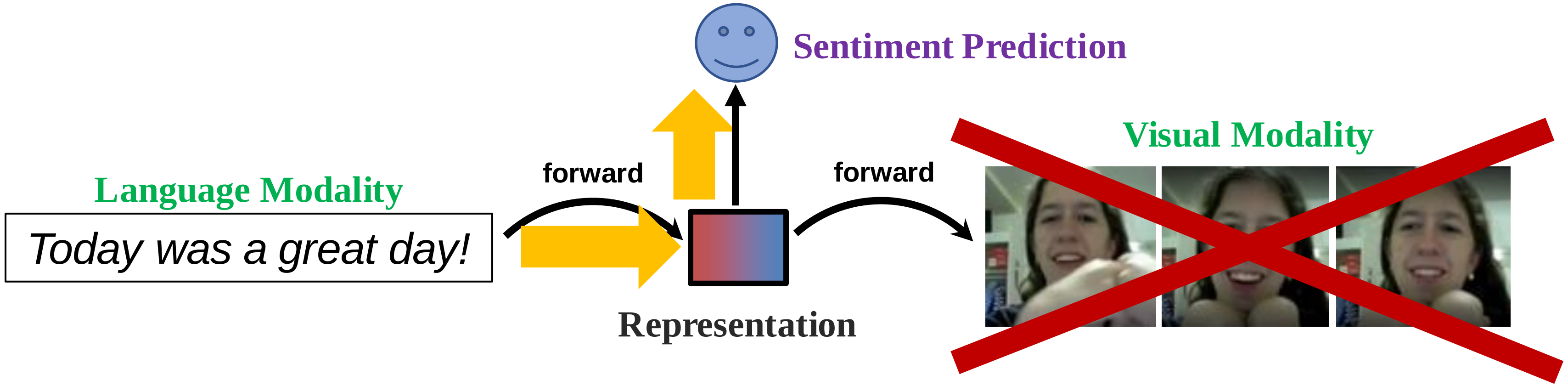


Cross-modal translation during training
Only language modality required at test time!

[Pham et al., Found in Translation: Learning Robust Joint Representations via Cyclic Translations Between Modalities. AAAI 2019]

Co-learning via Translation

Bimodal translations

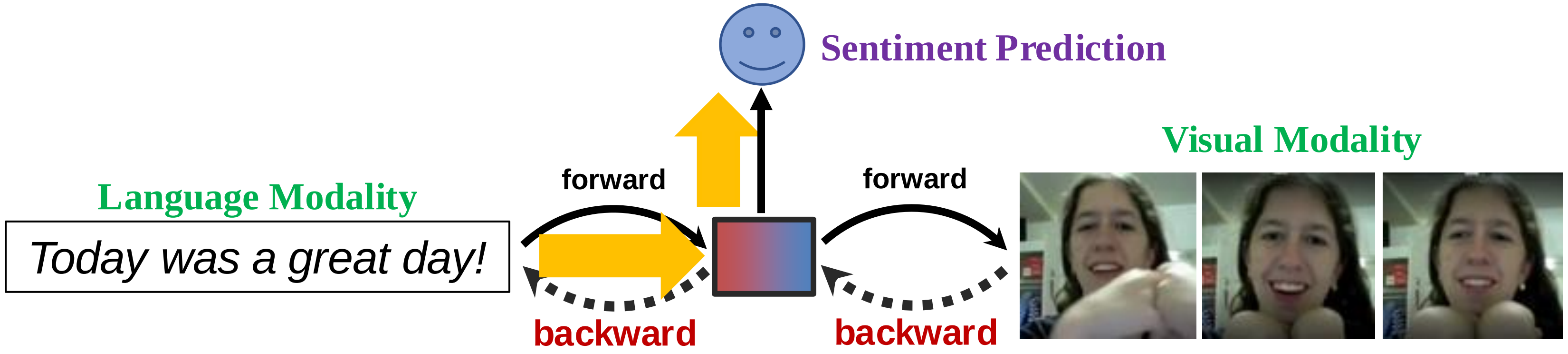


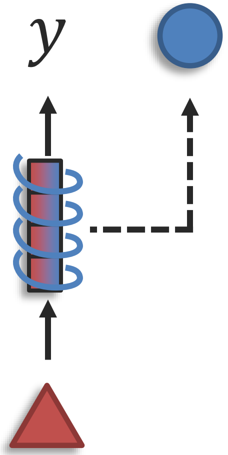
Problem: how do you ensure that both modalities are being used?

[Pham et al., Found in Translation: Learning Robust Joint Representations via Cyclic Translations Between Modalities. AAAI 2019]

Co-learning via Translation

Bimodal cyclic translations





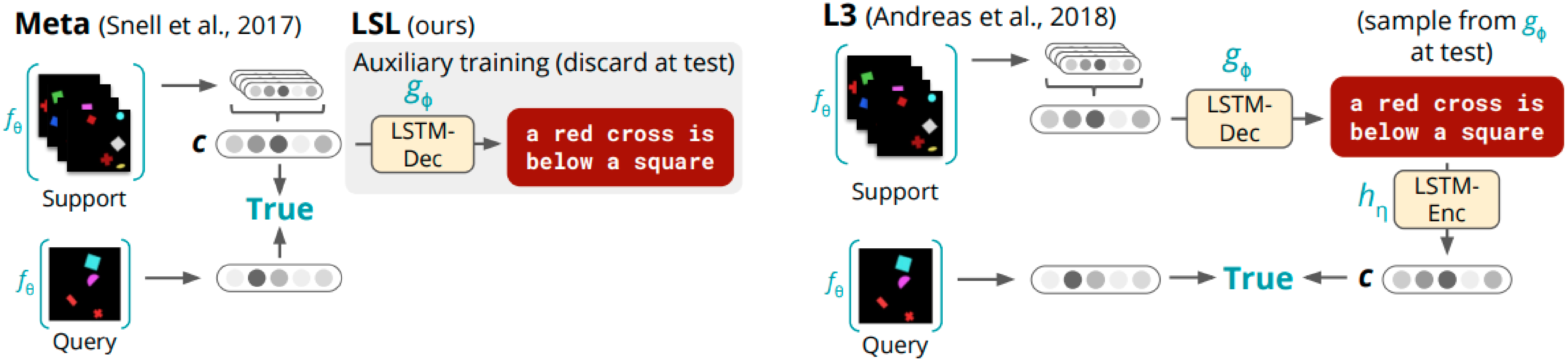
Solution: cyclic translations from visual back to language

Cross-modal translation during training
Only language modality required at test time!

[Pham et al., Found in Translation: Learning Robust Joint Representations via Cyclic Translations Between Modalities. AAAI 2019]

Co-learning for Compositionality

Image to text translation



[Mu et al., 2019. Shaping Visual Representations with Language for Few-Shot Classification]
[Andreas et al. 2017, Learning with Latent Language]
[Sharma et al. 2021. Skill Induction and Planning with Latent. Language]

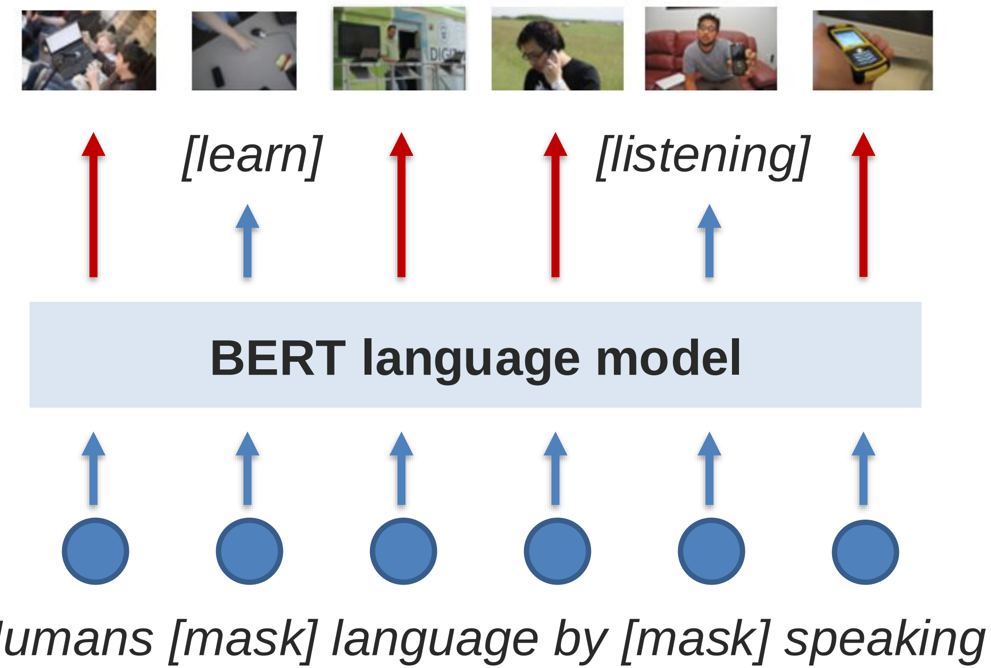
multisensory intelligence

Co-learning for Pre-training

Predicting images from corresponding language

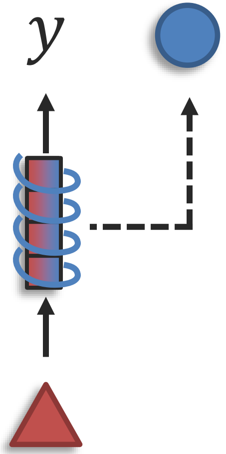
Voken (visual token) classification

Masked language modeling



Only text used at test-time

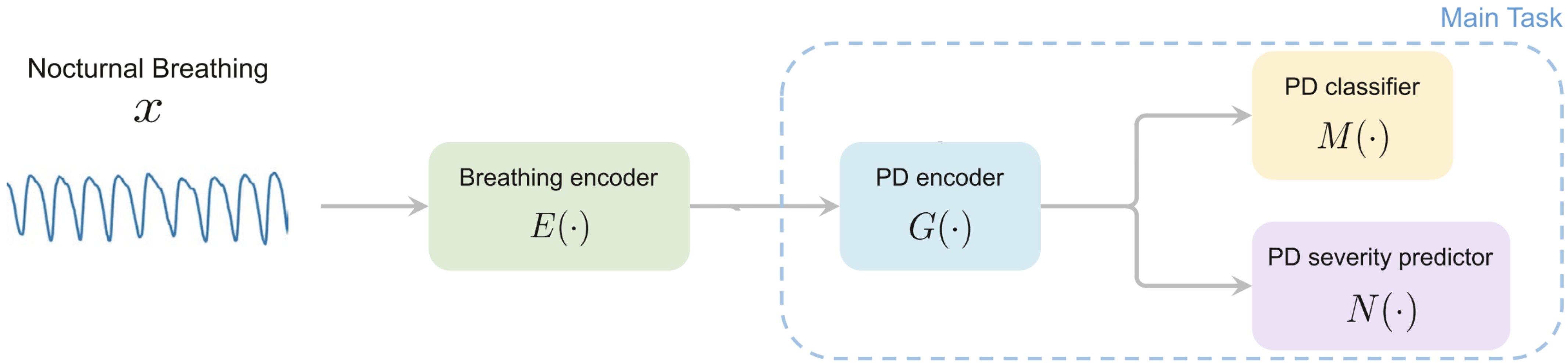
Multimodal co-learning > language-only training



[Tan and Bansal, Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. EMNLP 2020]

Co-learning for Dense Supervision

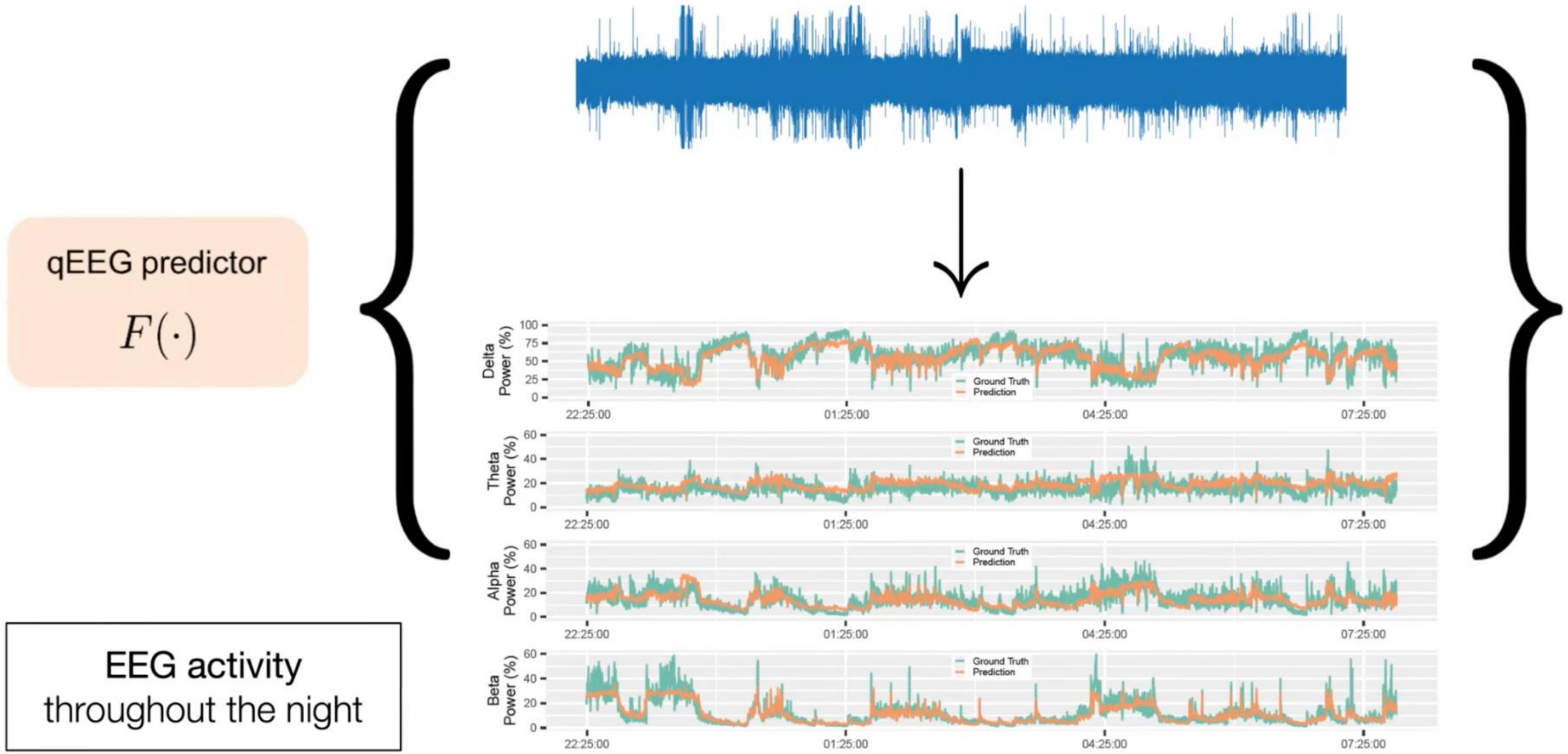
10hours of breathing data to detect Parkinson's Disease (PD) – sparse 1 bit signal



[Yang et al., Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. Nature Medicine 2022]

Co-learning for Dense Supervision

Predicting paired EEG data as auxiliary task provides dense supervision



[Yang et al., Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. Nature Medicine 2022]

Limits of Co-learning

Vision-language pretrained models on lexical grounding

Sentence-level semantic tasks

Encoder	SRL	Coref.	SPR	Rel.
$\text{BERT}_{\text{base}}$	90.10 ± 0.20	95.90 ± 0.00	83.70 ± 0.00	76.25 ± 0.05
$\text{VideoBERT}_{\text{text}}$	84.33 ± 0.05	92.47 ± 0.05	78.23 ± 0.05	65.83 ± 0.21
$\text{VideoBERT}_{\text{VL}}$	84.73 ± 0.05	92.82 ± 0.05	78.80 ± 0.00	66.37 ± 0.80
$\text{VisualBERT}_{\text{text}}$	89.00 ± 0.00	94.87 ± 0.05	82.27 ± 0.05	74.37 ± 0.19
$\text{VisualBERT}_{\text{VL}}$	89.57 ± 0.21	95.13 ± 0.05	82.17 ± 0.09	74.83 ± 0.05

Not much improvements with visual co-learning

Semantic Role Labeling *“The **carrots** are then pureed in the food processor”*

Entity Coreference *“After the **apples** are chopped, put **them** in the bowl”*

[Yun et al., Does Vision-and-Language Pretraining Improve Lexical Grounding? EMNLP 2021]

Limits of Co-learning

Vision-language pretrained models on seemingly multimodal tasks

Physical commonsense QA

Encoder	Linear	MLP	Trans.
$\text{BERT}_{\text{base}}$	55.43 ± 0.31	57.98 ± 0.16	60.12 ± 1.43
$\text{VideoBERT}_{\text{text}}$	57.87 ± 0.64	58.97 ± 0.44	62.35 ± 1.23
$\text{VideoBERT}_{\text{VL}}$	58.51 ± 0.20	58.56 ± 0.27	63.66 ± 1.31
$\text{VisualBERT}_{\text{text}}$	54.81 ± 0.19	56.81 ± 0.24	58.63 ± 0.79
$\text{VisualBERT}_{\text{VL}}$	55.83 ± 0.27	59.10 ± 0.11	61.66 ± 1.08

Marginal improvements with visual co-learning

“How to remove gloss from furniture?”

↓

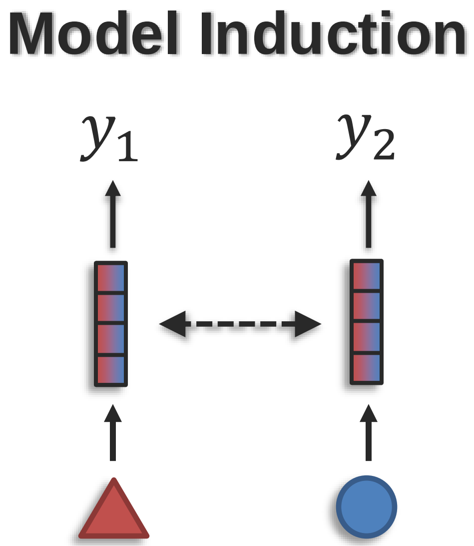
“Rub furniture with steel wool/cotton ball”

[Yun et al., Does Vision-and-Language Pretraining Improve Lexical Grounding? EMNLP 2021]

multisensory intelligence

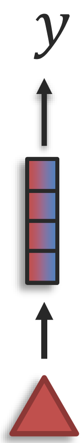
Part 3: Model Induction

Definition: Keeping individual unimodal models separate but inducing common behavior across separate models.



Self-training

Warmup: a single view – Self-training



Assume:
1. Labeled data $\{X_1^L, Y\}$.
2. Unlabeled data $\{X_1^U\}$.

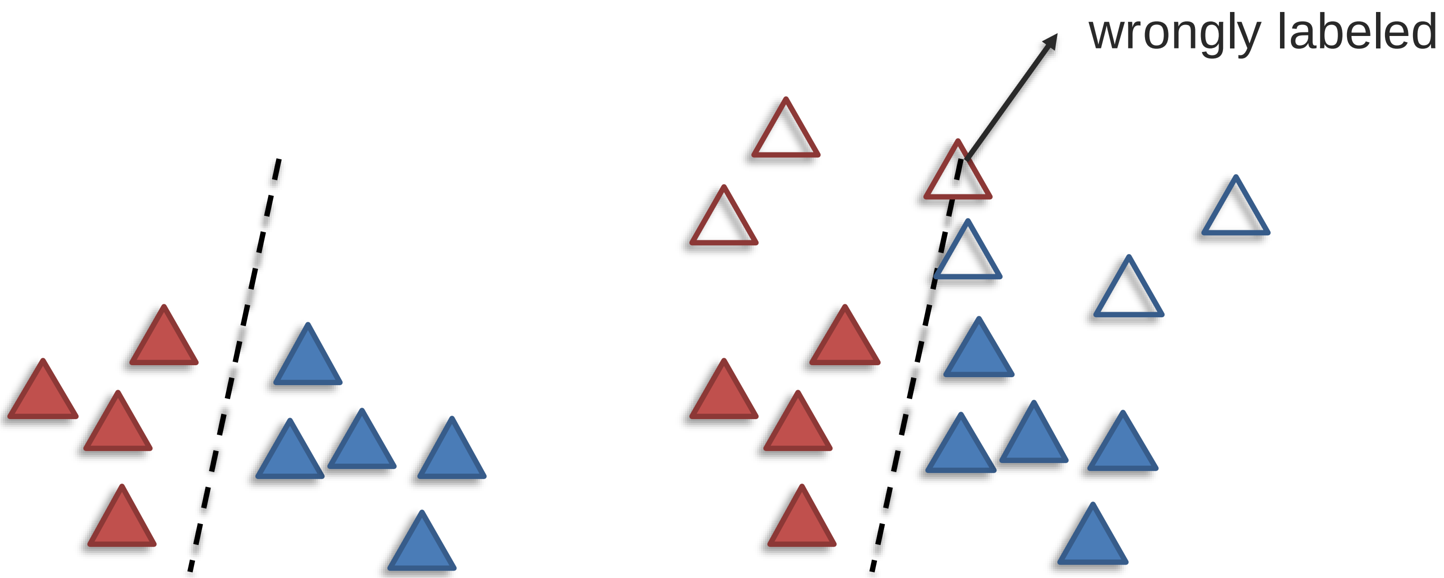
Train:
1. Train classifier f_1 on $\{X_1^L, Y\}$.
2. Use classifier f_1 to label the most confident examples in $\{X_1^U\}$ and add it to the labeled set $\{X_1^U, Y = f_1(X_1^U)\}$.
3. Go to 1, and repeat until there are no more unlabeled samples.

Test:
1. For a new unlabeled sample $\{X_1\}$, output $f_1(X_1)$.

[Wei et al., Theoretical Analysis of Self-Training with Deep Networks on Unlabeled Data. ICLR 2021]

Self-training

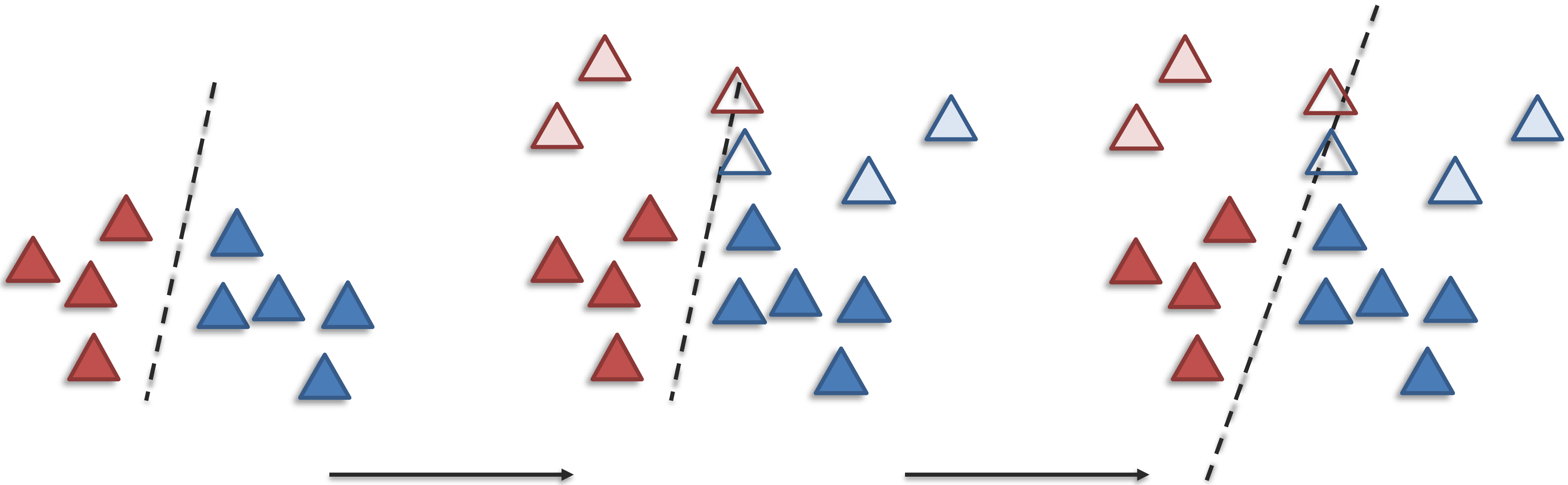
Warmup: a single view – Self-training



[Wei et al., Theoretical Analysis of Self-Training with Deep Networks on Unlabeled Data. ICLR 2021]

Self-training

Warmup: a single view – Self-training

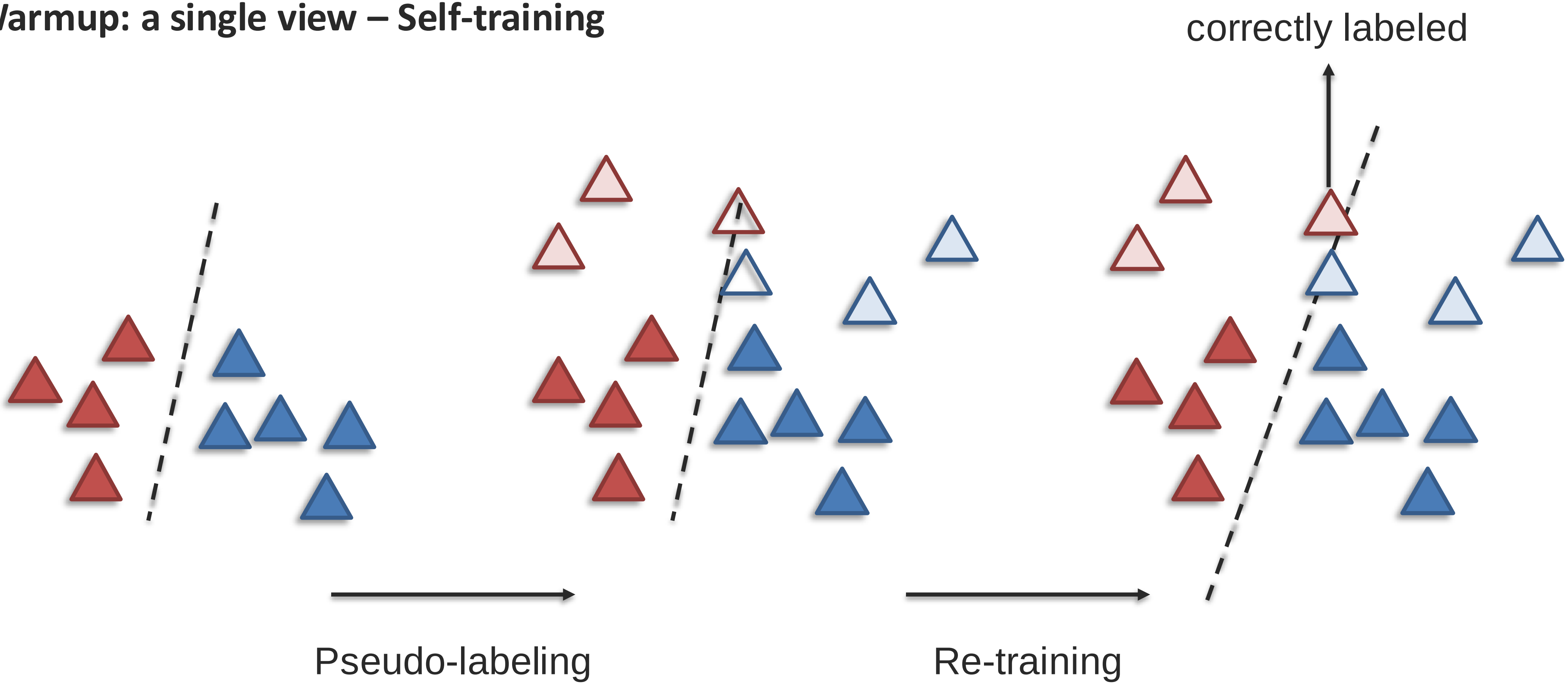


Pseudo-labeling Re-training

[Wei et al., Theoretical Analysis of Self-Training with Deep Networks on Unlabeled Data. ICLR 2021]

Self-training

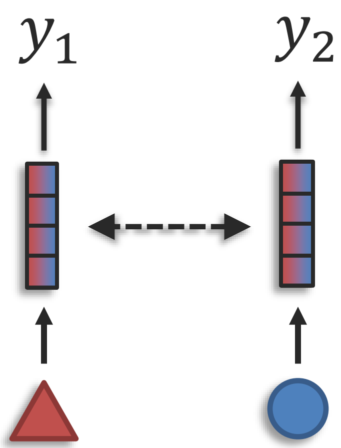
Warmup: a single view – Self-training



[Wei et al., Theoretical Analysis of Self-Training with Deep Networks on Unlabeled Data. ICLR 2021]

Self-training

From self-training to co-training



Ingredients:

- Two views on the data: x_1 and x_2
- Two classifiers: $x_1 \rightarrow y$ and $x_2 \rightarrow y$
- A bit of labeled data (x_1, x_2, y); lots of unlabeled data (x_1, x_2)

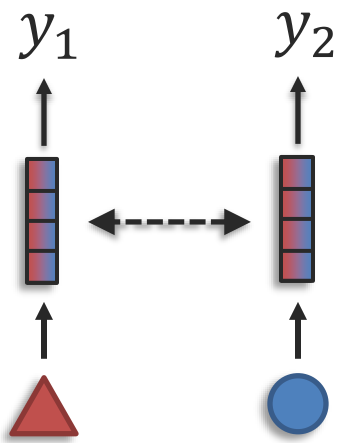
Assumptions:

1. Multi-view redundancy: either view is sufficient to predict the label alone, with enough data.

[Blum and Mitchell, Combining Labeled and Unlabeled Data with Co-Training. COLT 1998]

Co-training

Algorithm



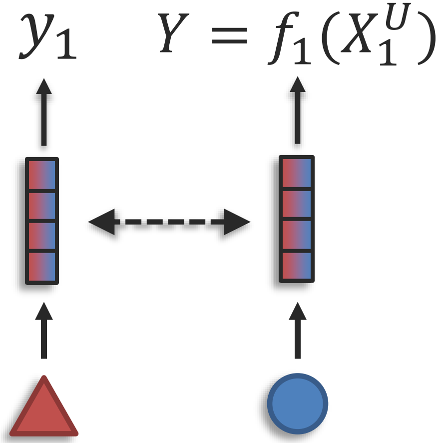
Assume:
1. **Small** amount of labeled data $\{X_1^L, X_2^L, Y\}$.
2. **Lots** of unlabeled data $\{X_1^U, X_2^U\}$.

Train:
1. Train classifier f_1 on $\{X_1^L, Y\}$ and f_2 on $\{X_2^L, Y\}$.

[Blum and Mitchell, Combining Labeled and Unlabeled Data with Co-Training. COLT 1998]

Co-training

Algorithm

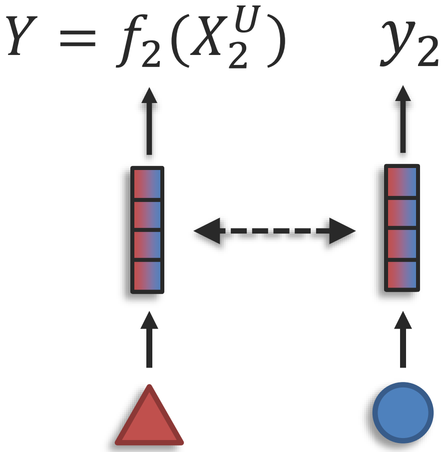


Assume:
1. **Small** amount of labeled data $\{X_1^L, X_2^L, Y\}$.
2. **Lots** of unlabeled data $\{X_1^U, X_2^U\}$.

Train:
1. Train classifier f_1 on $\{X_1^L, Y\}$ and f_2 on $\{X_2^L, Y\}$.
2. Use classifier f_1 to label the most confident examples in $\{X_1^U\}$ and add it to the labeled set to train f_2 $\{X_2^U, Y = f_1(X_1^U)\}$.

Co-training

Algorithm



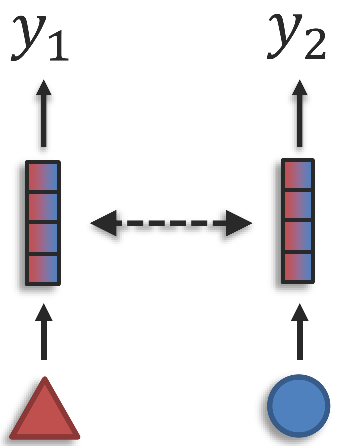
Assume:
1. **Small** amount of labeled data $\{X_1^L, X_2^L, Y\}$.
2. **Lots** of unlabeled data $\{X_1^U, X_2^U\}$.

Train:
1. Train classifier f_1 on $\{X_1^L, Y\}$ and f_2 on $\{X_2^L, Y\}$.
2. Use classifier f_1 to label the most confident examples in $\{X_1^U\}$ and add it to the labeled set to train f_2 $\{X_2^U, Y = f_1(X_1^U)\}$.
3. Use classifier f_2 to label the most confident examples in $\{X_2^U\}$ and add it to the labeled set to train f_1 $\{X_1^U, Y = f_2(X_2^U)\}$.
4. Go to 1, and repeat until there are no more unlabeled samples.

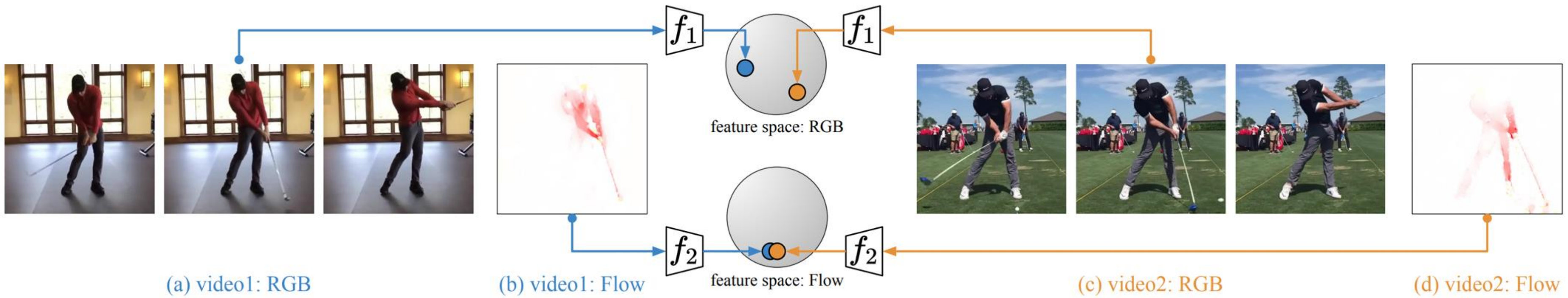
Test:
1. For a new unlabeled sample $\{X_1, X_2\}$, ensemble $f_1(X_1)$ and $f_2(X_2)$.

[Blum and Mitchell, Combining Labeled and Unlabeled Data with Co-Training. COLT 1998]

Modern Co-training

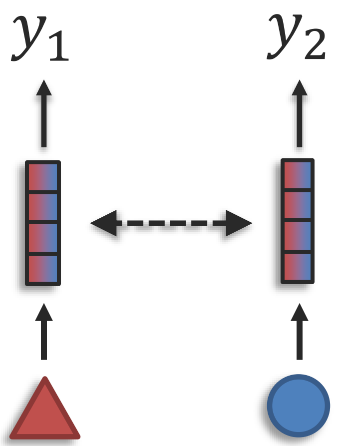


Co-training between RGB and optical flow for activity recognition
→ Positive samples hard to discover in RGB space can be easily found in flow space, and vice-versa (e.g., RGB sensitive to background differences but not flow).
→ Can use co-training between 2 RGB and flow contrastive learning modules.

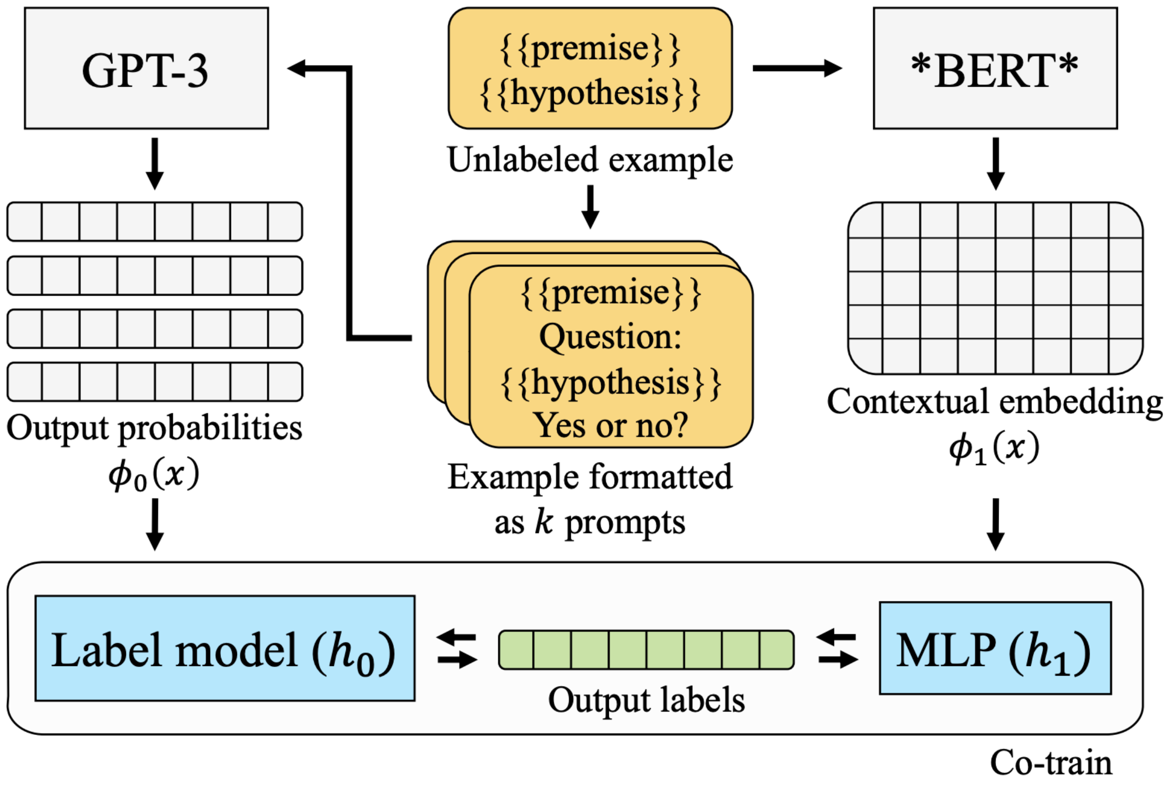


[Han et al., Self-supervised Co-training for Video Representation Learning. NeurIPS 2020]

Modern Co-training



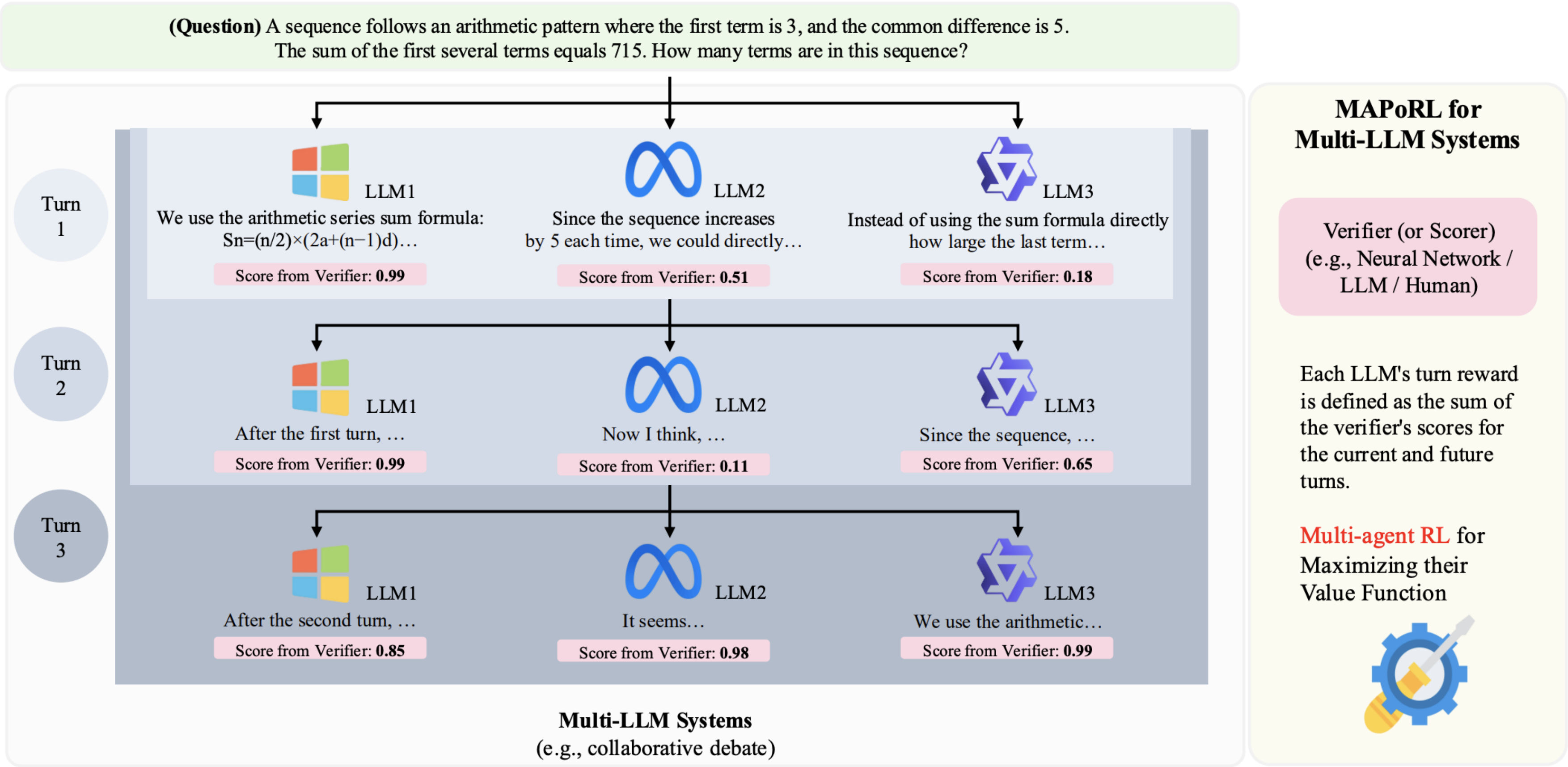
Language-model prompting



[Lang et al., Co-training Improves Prompt-based Learning for Large Language Models. ICML 2022]

Modern Co-training

Multi-agent LLMs, debate, co-training

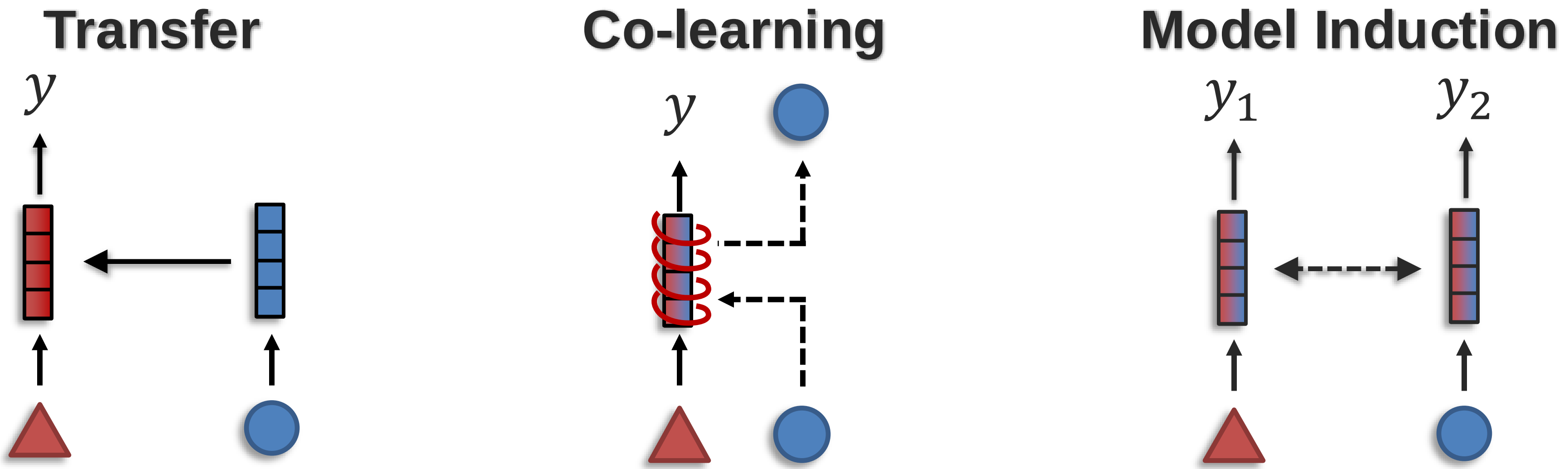


[Park et al., MAPoRL2: Multi-Agent Post-Co-Training for Collaborative Large Language Models with Reinforcement Learning. arXiv 2025]
[Du et al., Improving Factuality and Reasoning in Language Models through Multiagent Debate. ICML 2024]

multisensory intelligence

Summary: How to Cross-modal Learning

Definition: Transfer knowledge between modalities, usually to help the primary modality which may be noisy or with limited resources



1. Decide on secondary modalities
2. Decide on auxiliary input or auxiliary output
3. Decide on modifying model or using APIs only

multisensory intelligence

Lecture Topics

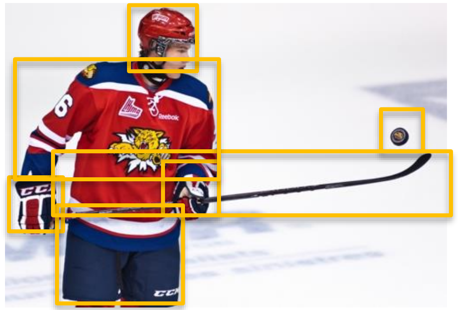
(subject to change, based on student interests and course discussions)

Module 1: Foundations of AI

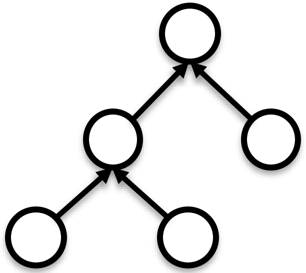
Week 1 (2/4): Introduction to AI and AI research

Week 2 (2/11): Data, structure, and information

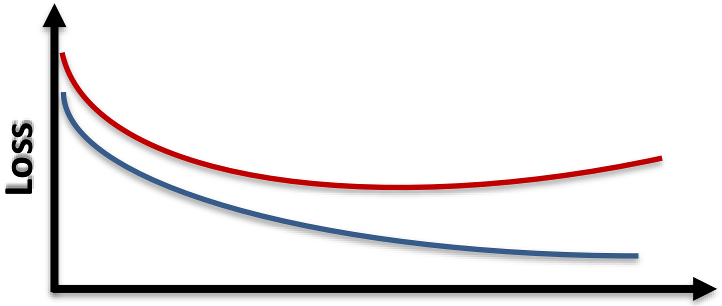
Week 4 (2/25): Common model architectures



Spatial



Hierarchical



Lecture Topics

(subject to change, based on student interests and course discussions)

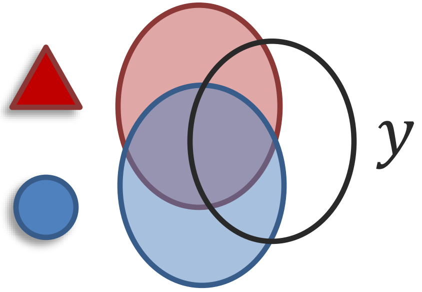
Module 2: Foundations of multimodal AI

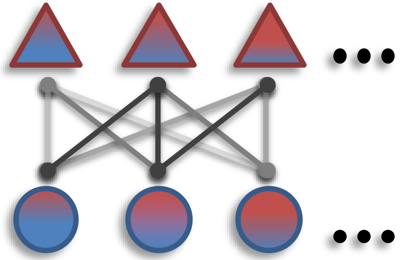
Week 5 (3/4): Multimodal connections and alignment

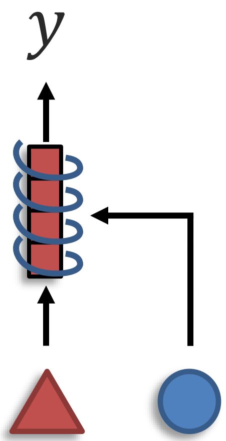
Week 6 (3/11): Multimodal interactions and fusion

Week 7 (3/18): Cross-modal transfer

Week 8 – No class, spring break







Lecture Topics

(subject to change, based on student interests and course discussions)

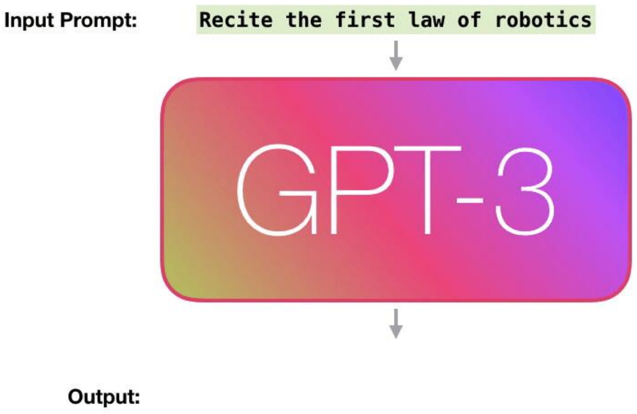
Module 3: Large models and modern AI

Week 9 (4/1): Pre-training, scaling, fine-tuning LLMs

Week 10 – No class, member's week

Week 11 (4/15): Large multimodal models

Week 12 (4/22): Modern generative AI



An armchair in the shape of an avocado

multisensory intelligence

Lecture Topics

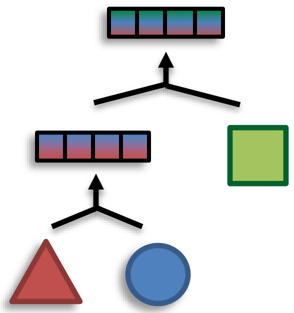
(subject to change, based on student interests and course discussions)

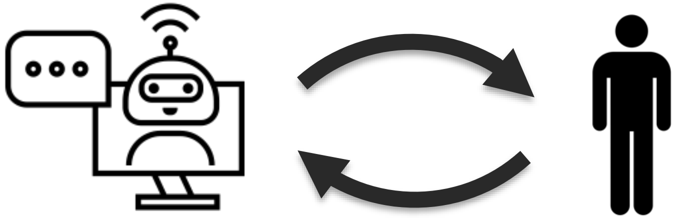
Module 4: Interactive AI

Week 13 (4/29): Interactive AI agents

Week 14 (5/6): Multi-step reasoning

Week 15 (5/13): Human-AI interaction and safety





multisensory intelligence

Assignments for This Coming Week

For project:

- Instructions for midterm assignment posted on piazza
- Midterm report due April 1 (Tuesday), presentations April 3 (Thursday)
- For April 3 (Thursday), class from 1-3pm (we will be flexible when you attend and present)
- Finalized main ideas and experimental setup, have datasets and baseline models working, detailed error analysis, initial progress towards implementing new ideas.

Reading assignment due tomorrow Wednesday (3/19).

This Thursday (3/20): fourth reading discussion on **multimodal interactions**.

1. Ten myths of multimodal interaction
2. Mixture-of-experts fusion

multisensory intelligence