Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 11.2: Transference

Paul Liang

** Co-lecturer: Louis-Philippe Morency. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk. Spring 2023 edition taught by Yonatan and Daniel Fried*

Administrative Stuff

Final Project Report (Due Sunday 12/10 at 8pm)

Main goals:

1. Produce a research paper which will motivate your research problem, describe the prior work, present your research contributions, explain the details of your experiments, and discuss your results.
2. Novel research ideas (N-1 new ideas for N students)
   - Novel algorithm
   - Novel application
   - Can you explain your idea in a few sentences, without reference to baselines?
3. Incorporate feedback from previous milestones
4. Compare to multimodal baselines from midterm report
   1. Did the proposed ideas solve the errors highlighted in error analysis?
   2. Broader implications of proposed ideas.

Final Project Report (Due Sunday 12/10 at 8pm)

Some suggestions:

- Proposed ideas
  - Explain how it tackles the challenges identified through error analysis
  - Formally explain the method and novelty
- Experimental setup
  - Datasets, metrics, baselines, methodology
  - Ablation studies
- Results
  - One subsection for each research question
  - The most important part is the discussion: what do the results mean, what implications they have, how should they be interpreted in the broader context?

Final Project Report (Due Sunday 12/10 at 8pm)

Some suggestions:

- Clear motivated research questions
- Clear ablation studies, revisit error analysis, add visualizations
- Not about results, but discussion
    - If it works, why does it work
    - If it doesn't idea, why did it not work and how can we fix it
- If your dataset is too large:
    - You can use a subset of your data or train for fewer epochs
    - But be consistent between experiments
- 3 students: 8 pages, 4 students: 9 pages, 5 students: 10 pages

Final Project Presentations (Tuesday 12/5 and Thursday 12/7)

Main objective:

- Present your research ideas and get feedback from classmates
- Focus on **only one** of your new research ideas
- All students should present and answer questions
- Be sure to be on time! We have many presentations each day ☺
- All presentations are in person (no remote presentations)

Presentation length:

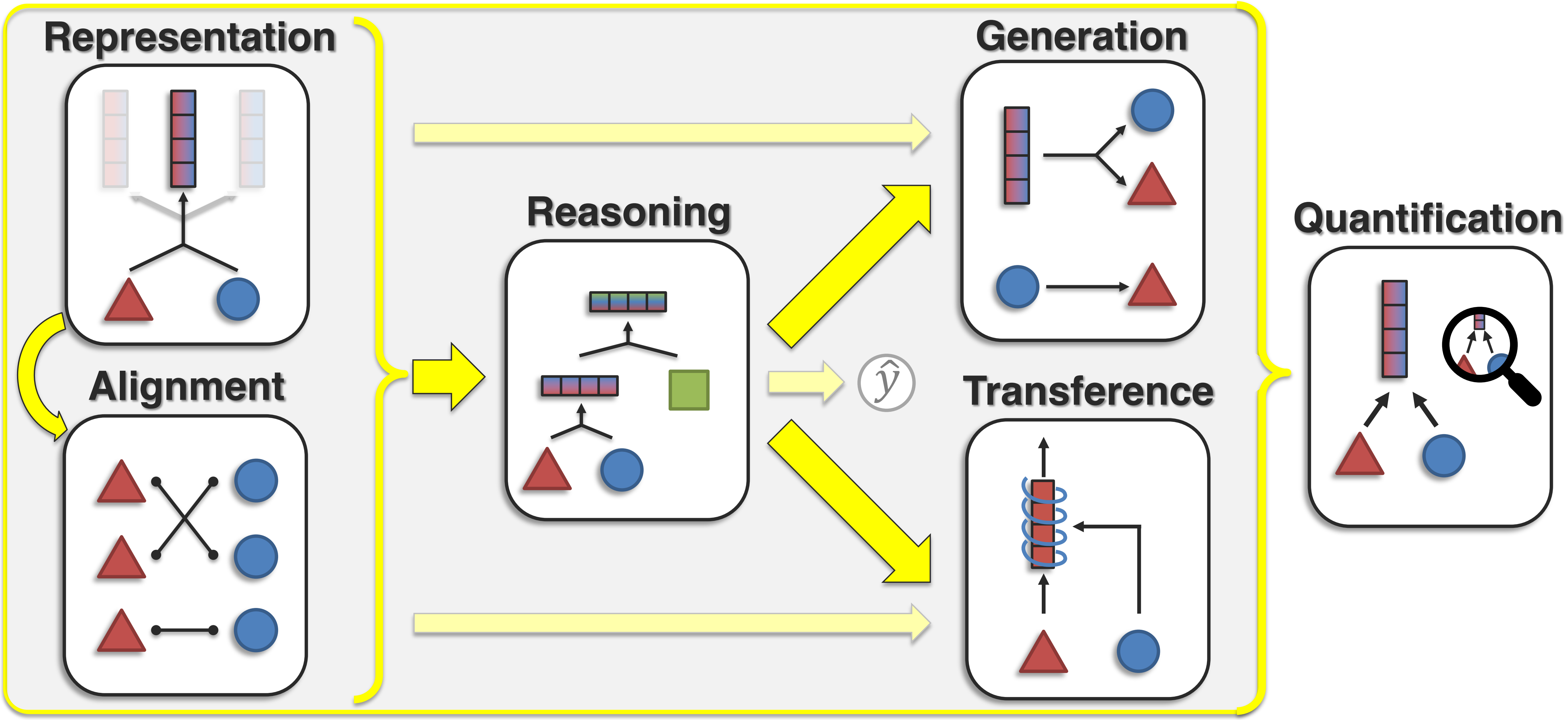
- 30-seconds elevator pitch
- 4-minute full presentation – all students should present

- Following each presentation, audience will be asked to share feedback

Final Project Presentations (Tuesday 12/5 and Thursday 12/7)

We will give more details about grading, presentation order, etc.

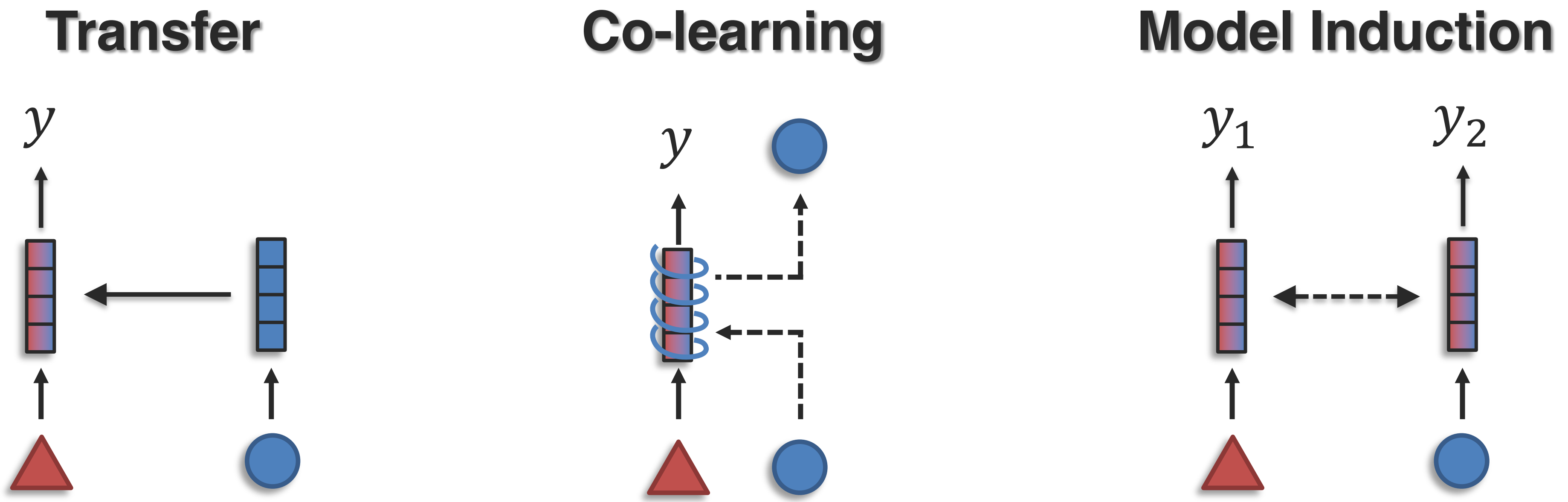
Core Multimodal Challenges



Transference

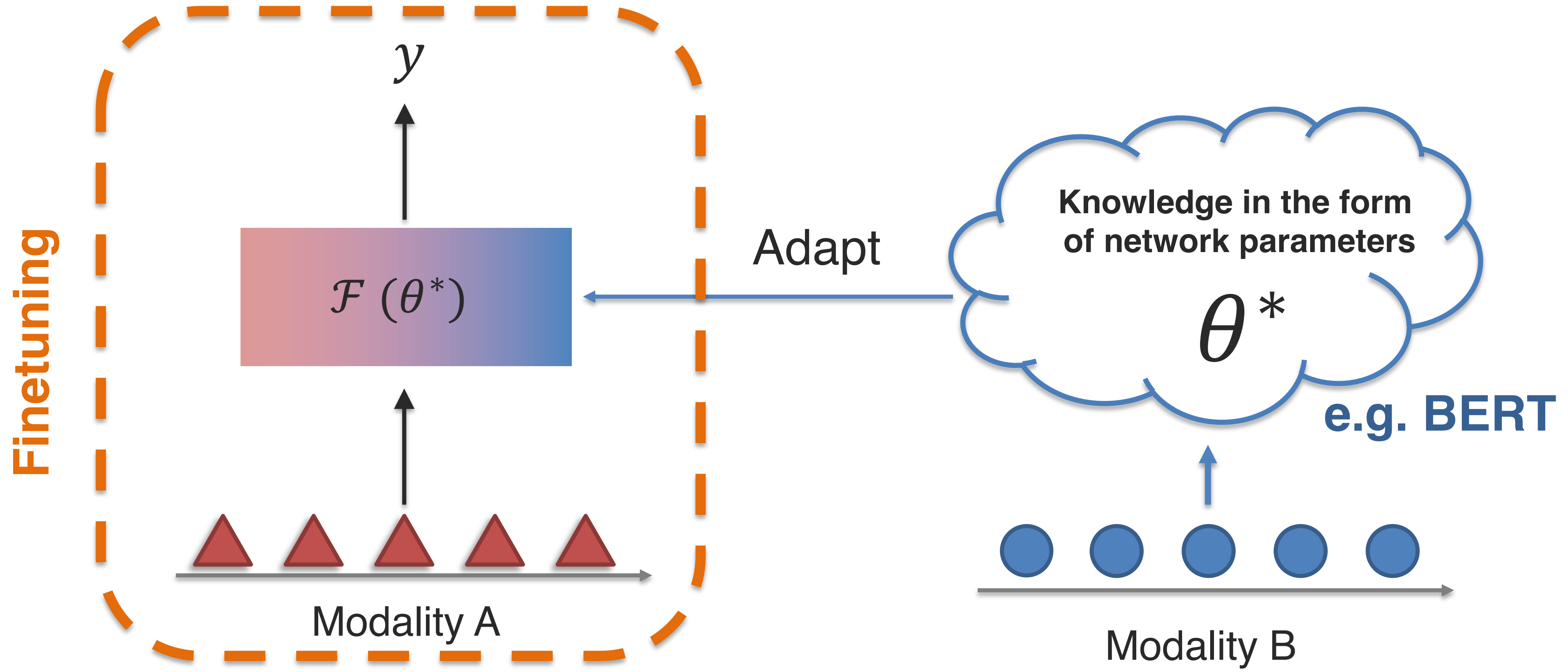
Definition: Transfer knowledge between modalities, usually to help the primary modality which may be noisy or with limited resources

Sub-challenges:



Sub-Challenge 5a: Transfer via Pretrained Models

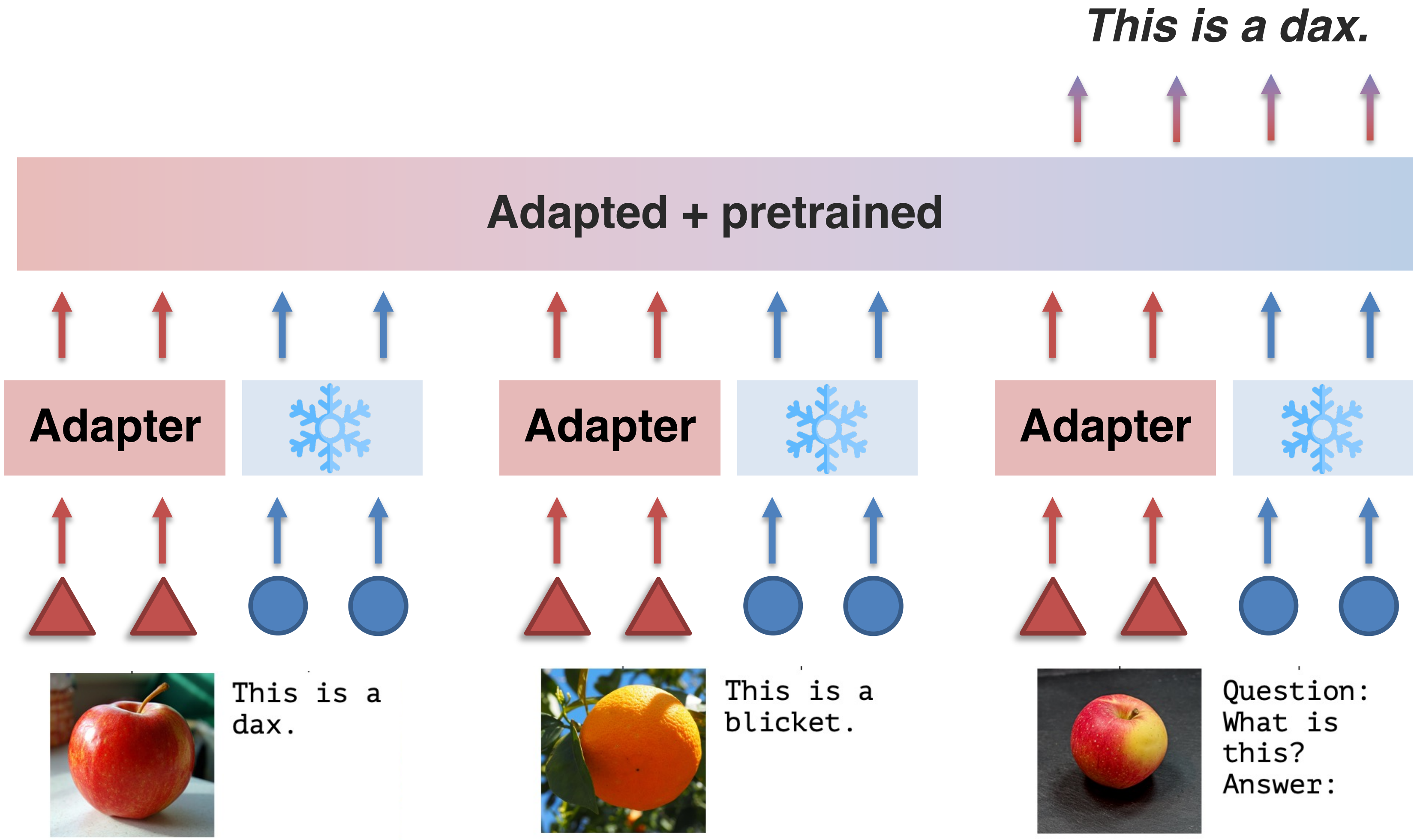
Definition: Transferring knowledge from large-scale pretrained models to downstream tasks involving the primary modality.



Sub-Challenge 5a: Transfer via Pretrained Models

Transfer via prefix tuning

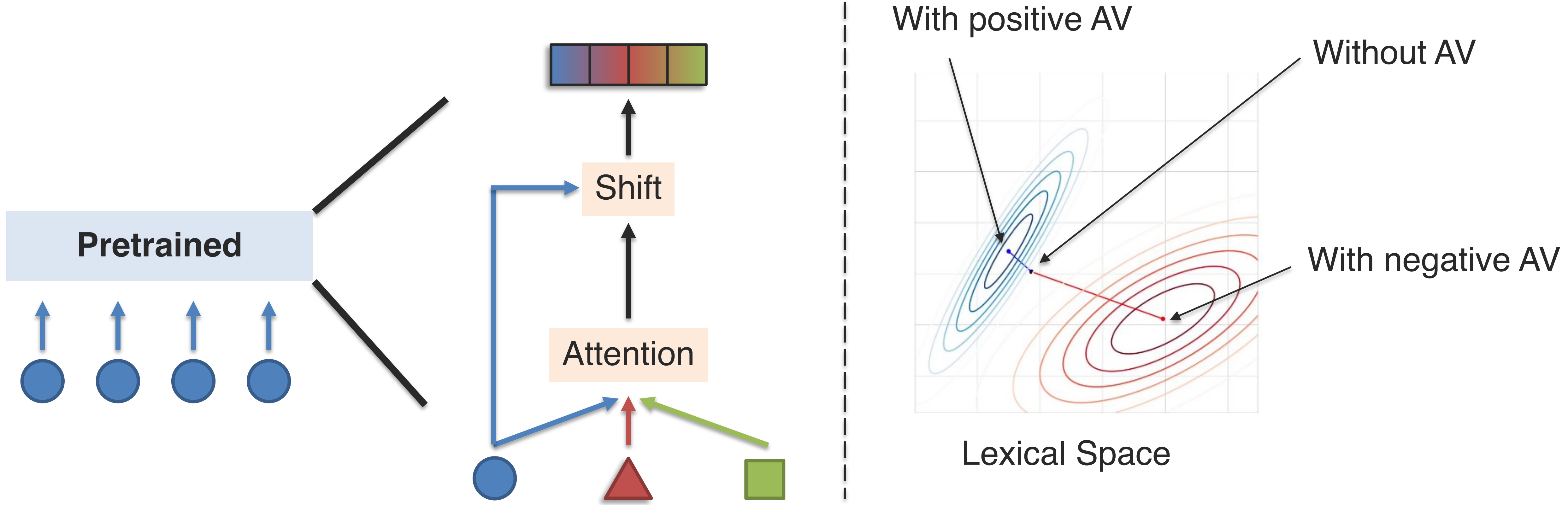
Few-shot image classification:



[Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021]

Sub-Challenge 5a: Transfer via Pretrained Models

Transfer via representation tuning



[Ziegler et al., Encoder-Agnostic Adaptation for Conditional Language Generation. arXiv 2019]
[Rahman et al., Integrating Multimodal Information in Large Pretrained Transformers. ACL 2020]

Sub-Challenge 5a: Transfer via Pretrained Models

1. Disentanglement

$$\mathcal{L}_\beta(x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta \cdot \mathrm{KL}(q_\phi(z|x)||p(z))$$

2. Conditioning

$$p(\boldsymbol{x}_{0:T} \mid y) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, y)$$

3. Prompt tuning

4. Representation tuning

5. Classifier gradient tuning

$$\nabla \log p(\boldsymbol{x}_t \mid y) = \underbrace{\nabla \log p(\boldsymbol{x}_t)}_{\text{unconditional score}} + \gamma \underbrace{\nabla \log p(y \mid \boldsymbol{x}_t)}_{\text{classifier gradient}}$$

6. Classifier-free tuning

$$\nabla \log p(\boldsymbol{x}_t \mid y) = \underbrace{\gamma \nabla \log p(\boldsymbol{x}_t \mid y)}_{\text{conditional score}} + \underbrace{(1-\gamma) \nabla \log p(\boldsymbol{x}_t)}_{\text{unconditional score}}$$

Multitask and Transfer Learning

How can we transfer knowledge across multiple tasks, each over a different subset of modalities?

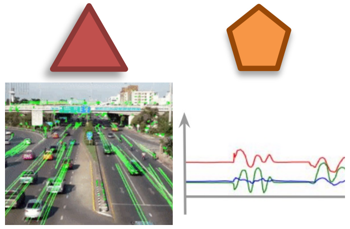
Video classification

Language Video Audio

Sentiment, emotions

Audio Video

Robot dynamics



Video Time-series

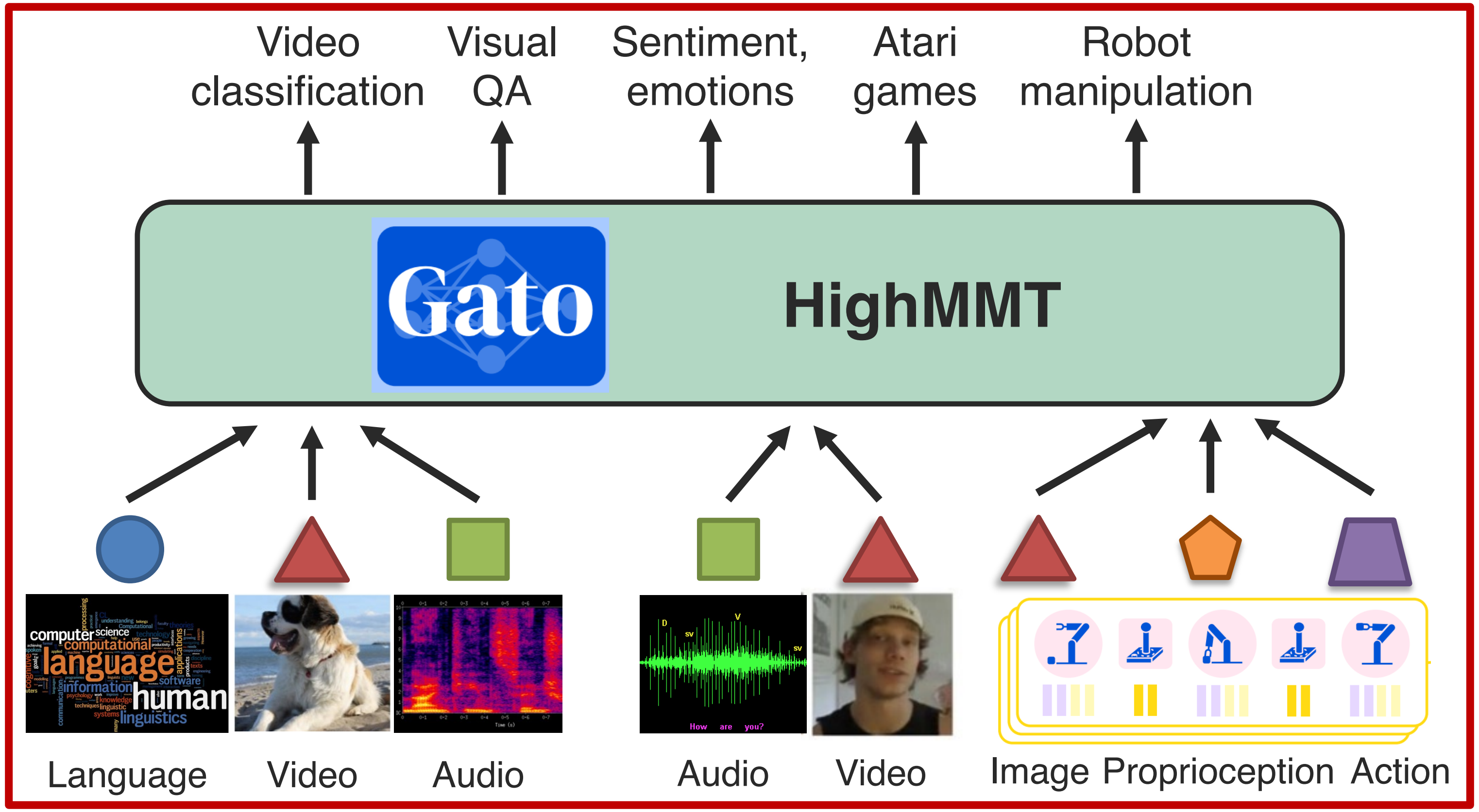
Generalization across modalities and tasks
Important if some tasks are low-resource

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

High-Modality Multimodal Transformers

Transfer across partially observable modalities

Unified model + parameter sharing + multitask and transfer learning



Non-parallel multitask learning

Task-specific classifiers

Same model architecture!

Shared multimodal model

Same parameters!

Modality-specific embeddings

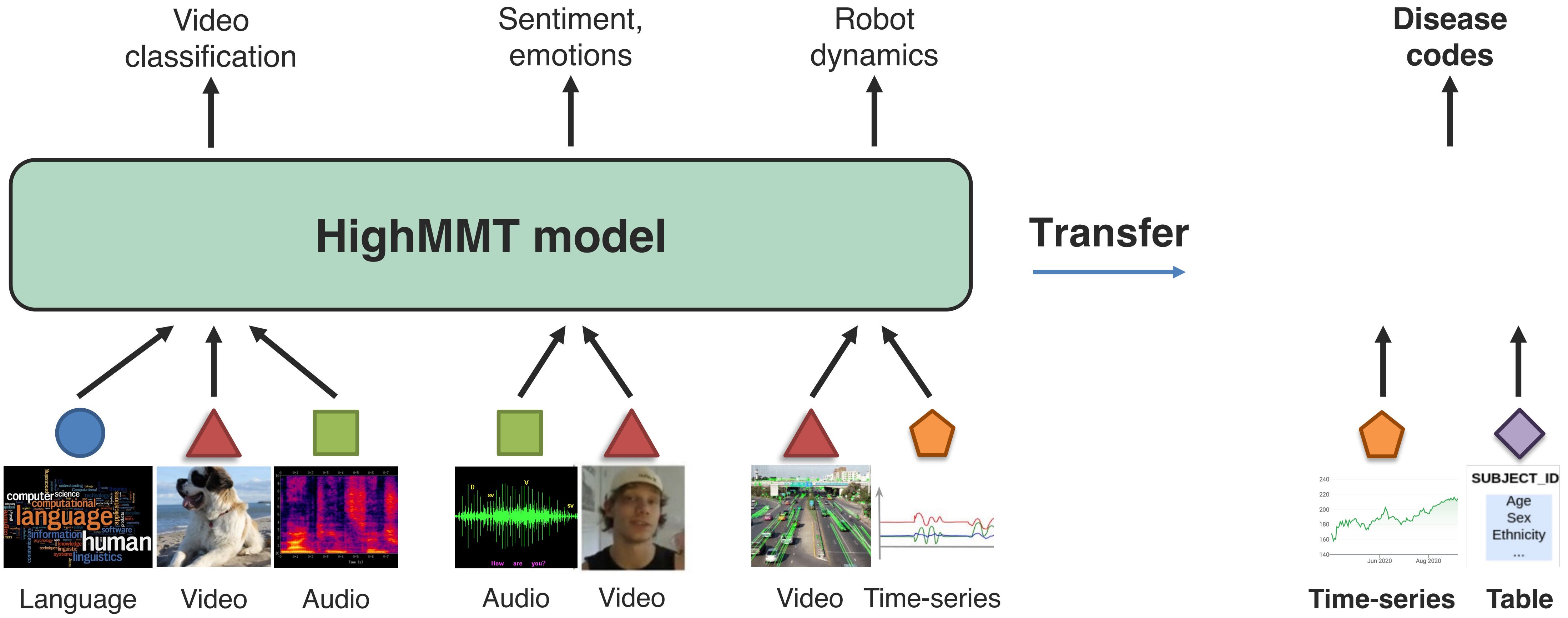
Standardized input sequence

[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

Transfer across partially observable modalities

HighMMT: unified model + parameter sharing + multitask and transfer learning

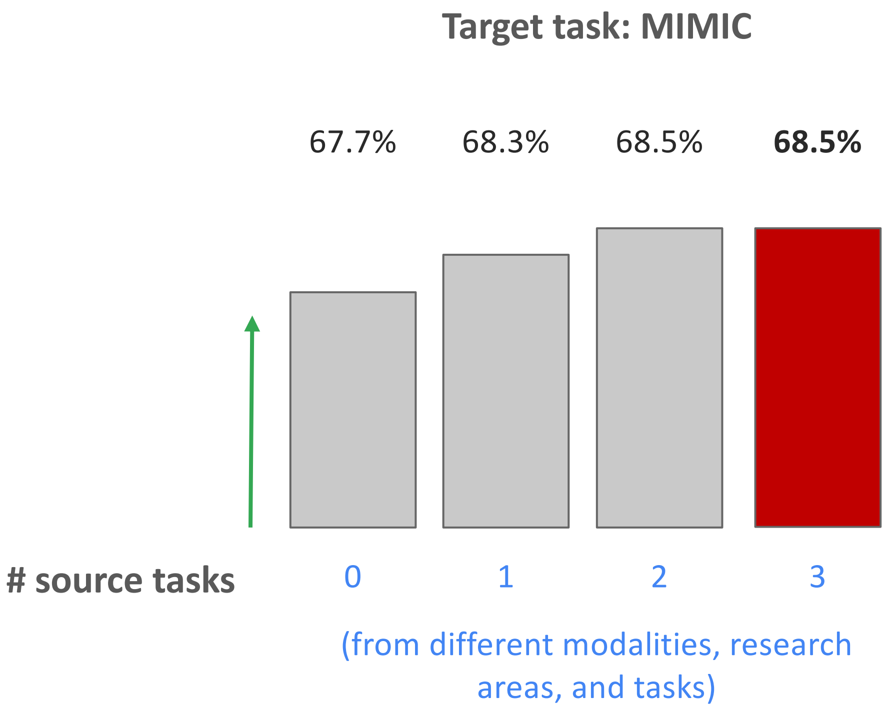


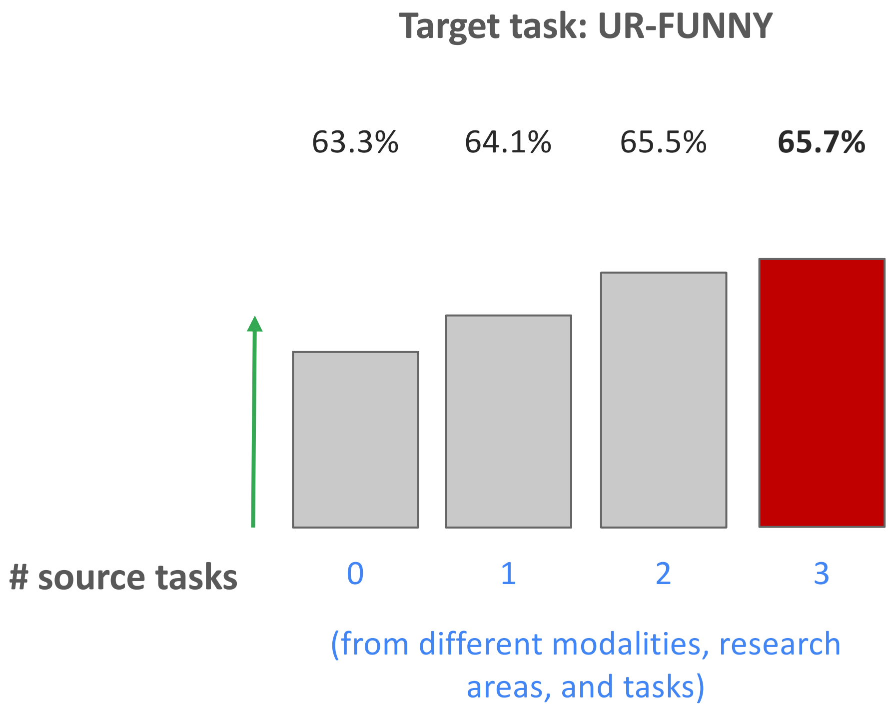
[Liang et al., HighMMT: Towards Modality and Task Generalization for High-Modality Representation Learning. TMLR 2022]

Multitask and Transfer Learning

Transfer across partially observable modalities

HighMMT: unified model + parameter sharing + multitask and transfer learning





Achieves both multitask and transfer capabilities across modalities and tasks

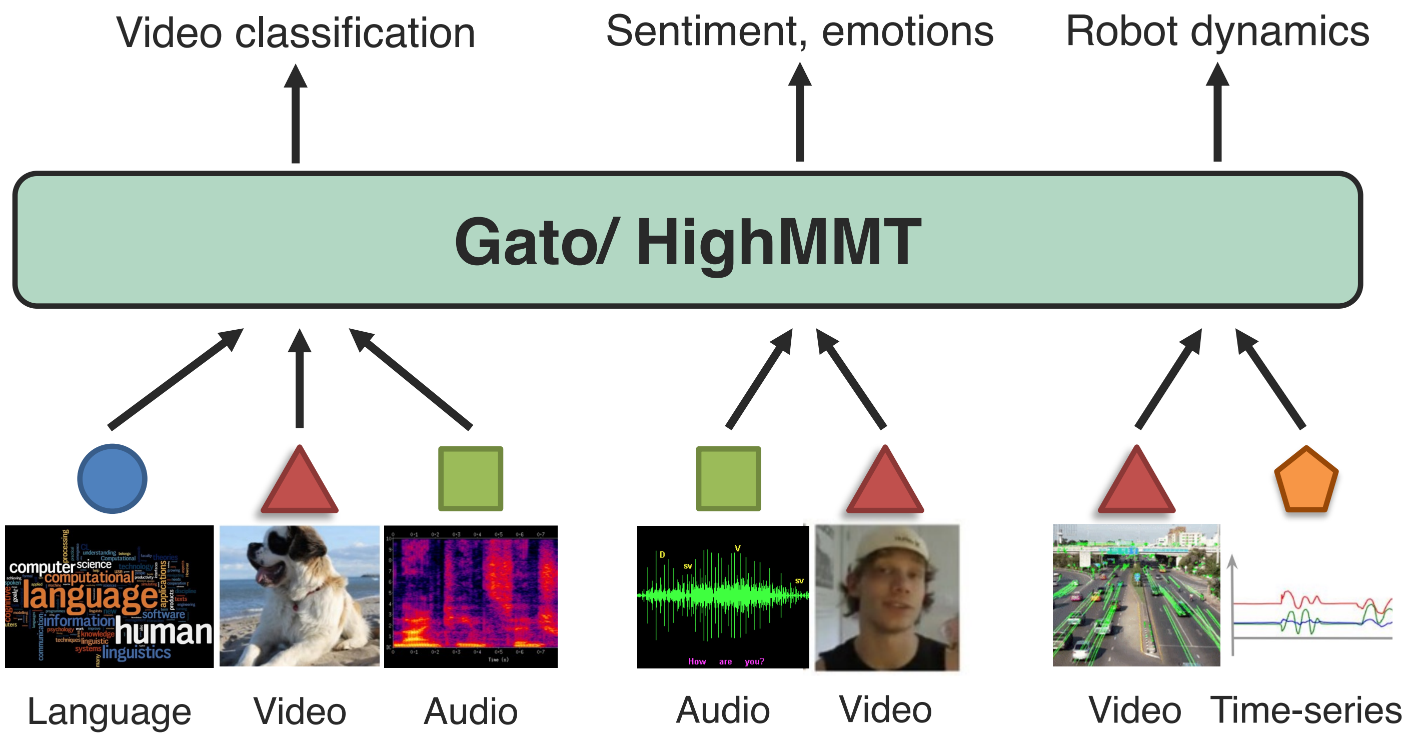
[Liang et al., HighMMT: Towards Modality and Task Generalization for High-Modality Representation Learning. TMLR 2022]

High-Modality Models

Open challenges

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?

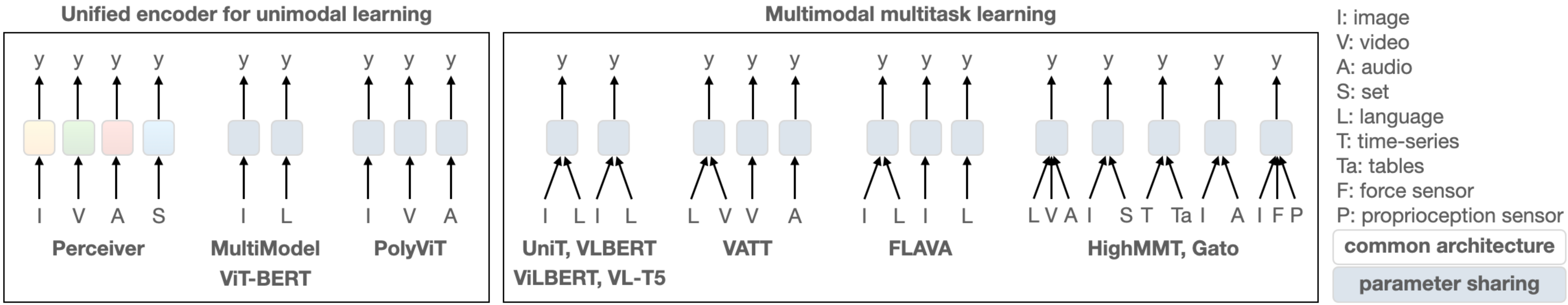
Modality-specific embeddings?

Standardized input sequence?

Multitask and Transfer Learning

Open challenges

Many more dimensions of transfer

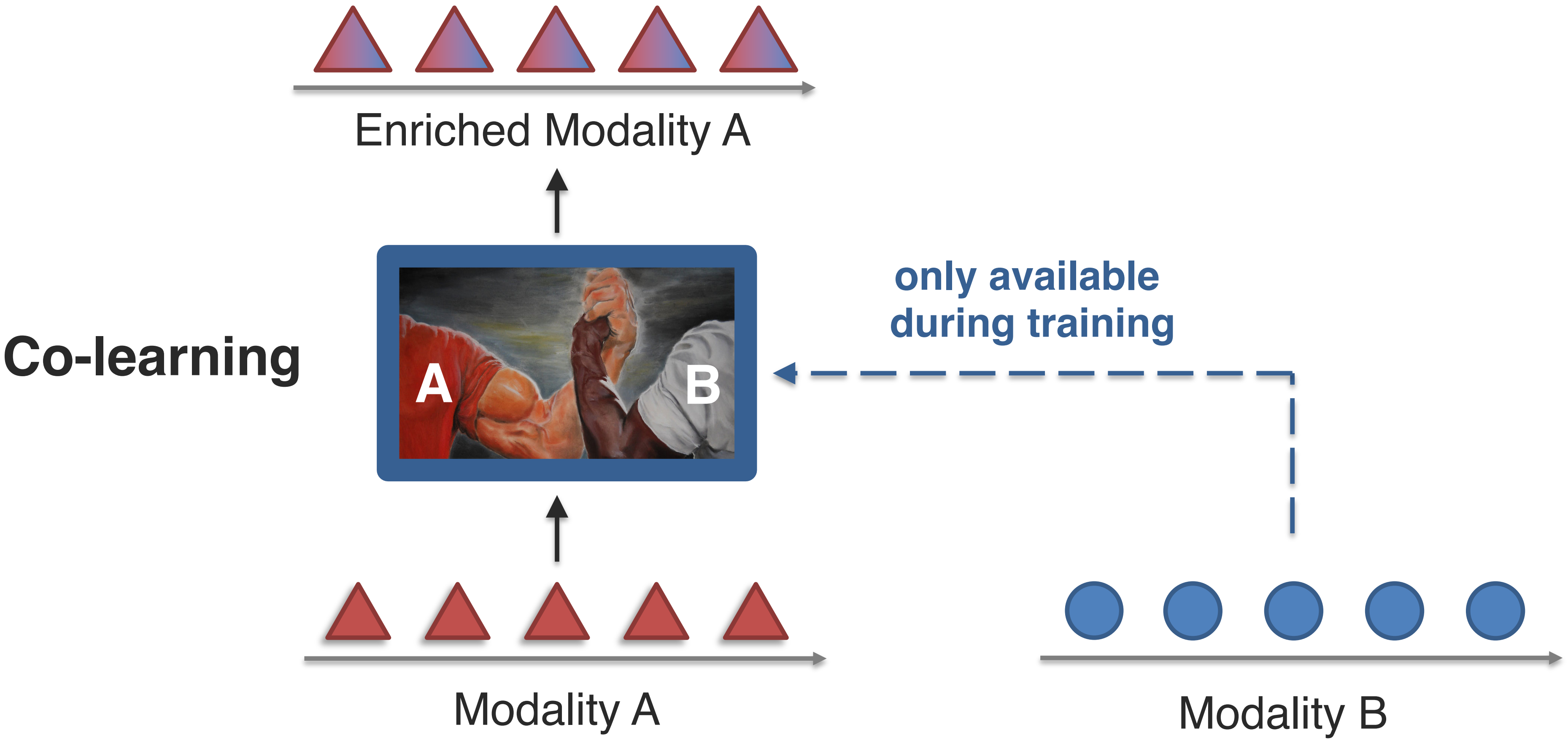


Open challenges:

- 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)

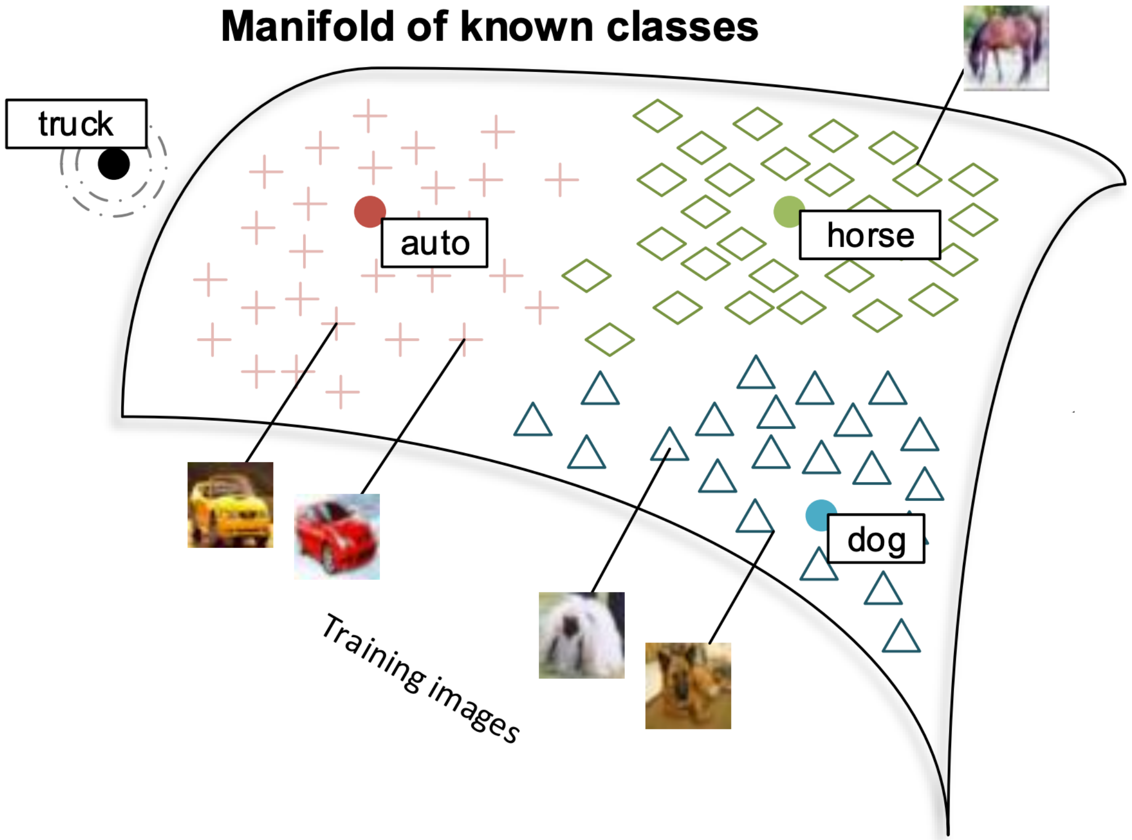
Sub-Challenge 5b: Co-learning

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

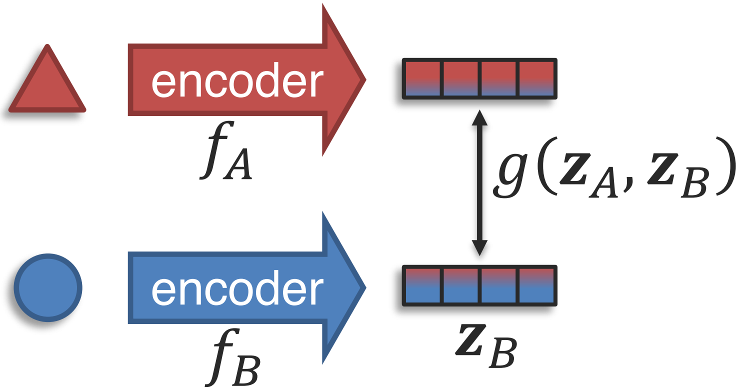


Co-learning via Representation

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



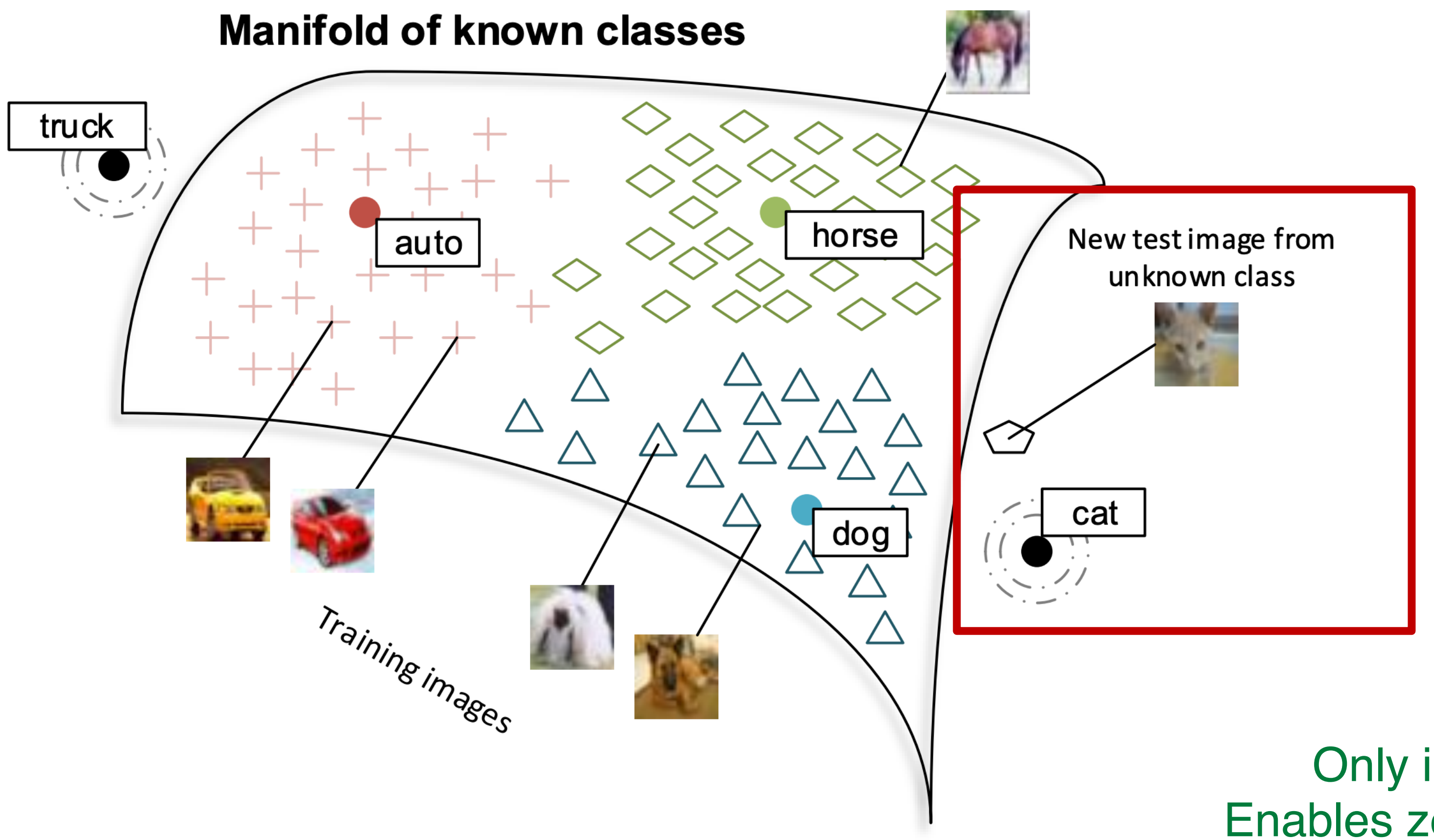
Recall representation coordination!



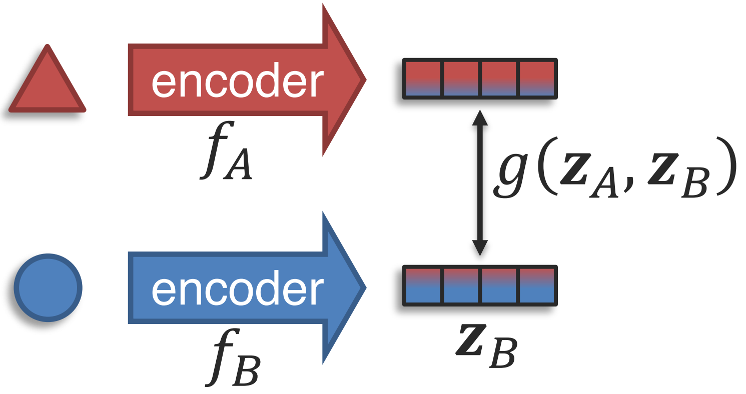
[Socher et al., Zero-Shot Learning Through Cross-Modal Transfer. NeurIPS 2013]

Co-learning via Representation

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



Recall representation coordination!

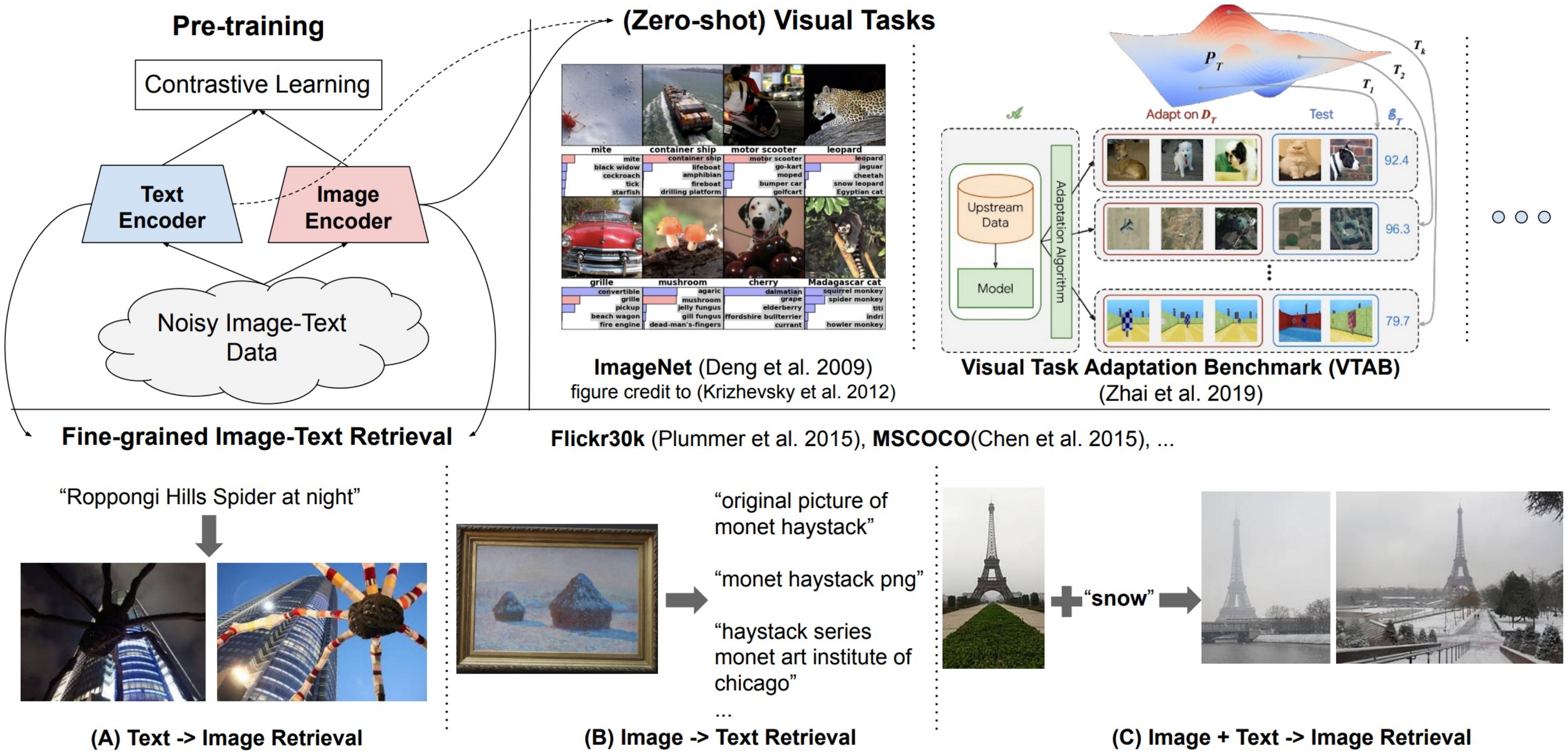


Only images used at test-time
Enables zero-shot image classification

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

Co-learning via Representation

Representation coordination at scale



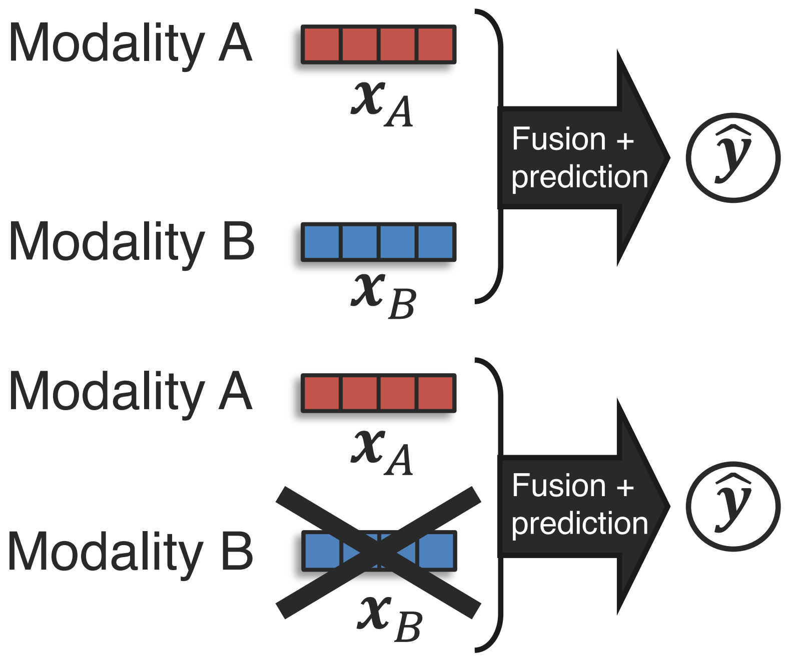
[Jia et al., Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. ICML 2021]

Co-learning via Representation

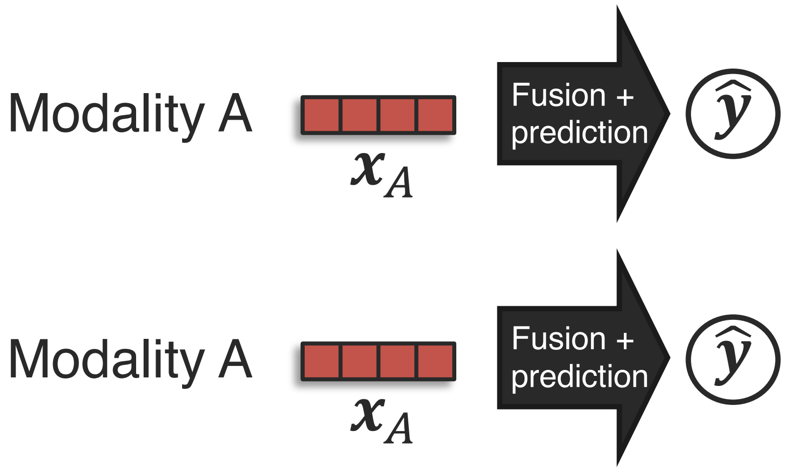
Representation fusion

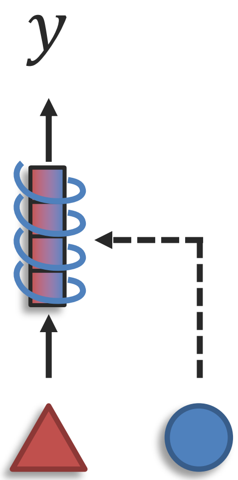
Train	Multimodal data Multimodal model	
Test	Language-only data Language-only model Fill rest by 0s	

Multimodal co-learning



Unimodal learning





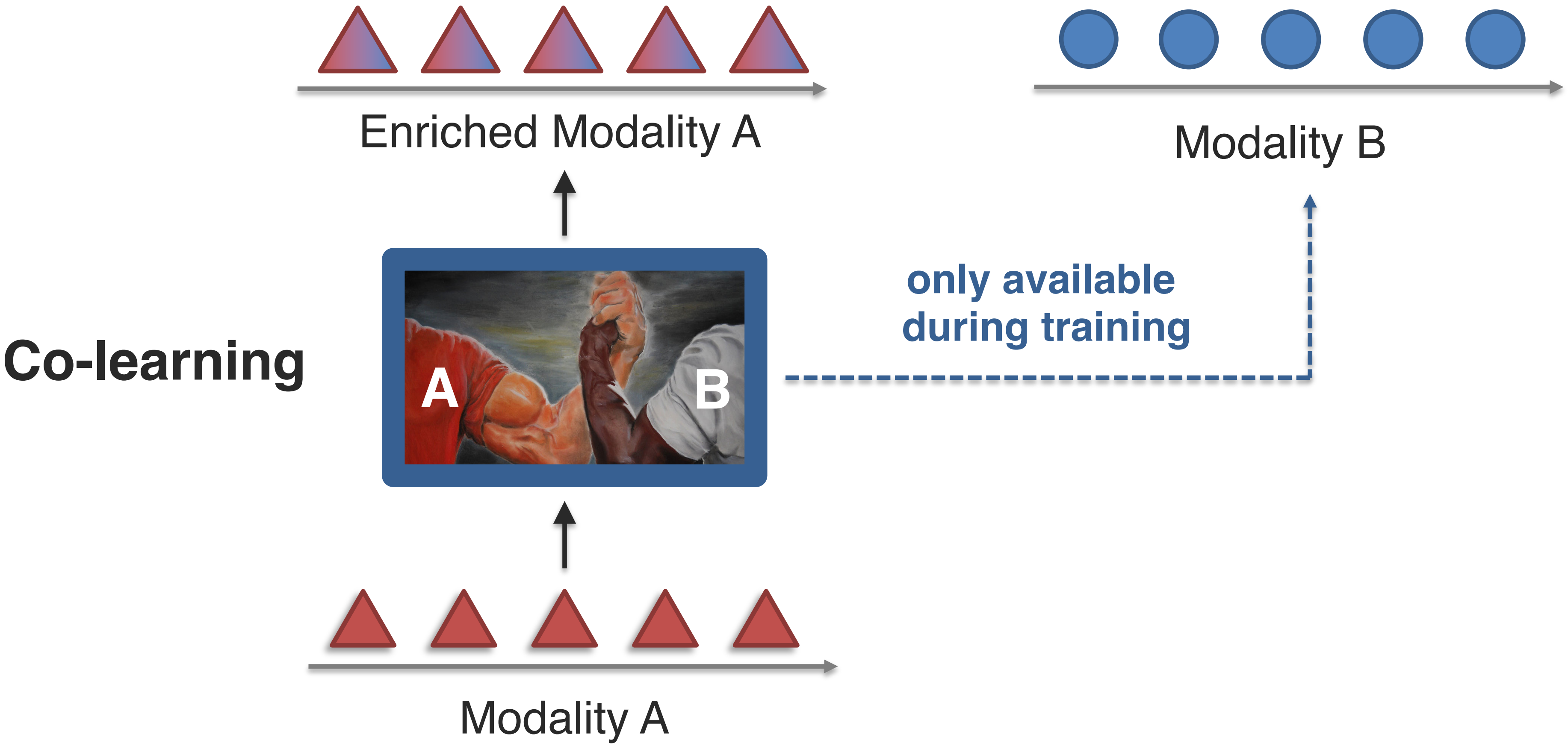
Only text used at test-time

Multimodal co-learning > language-only training

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

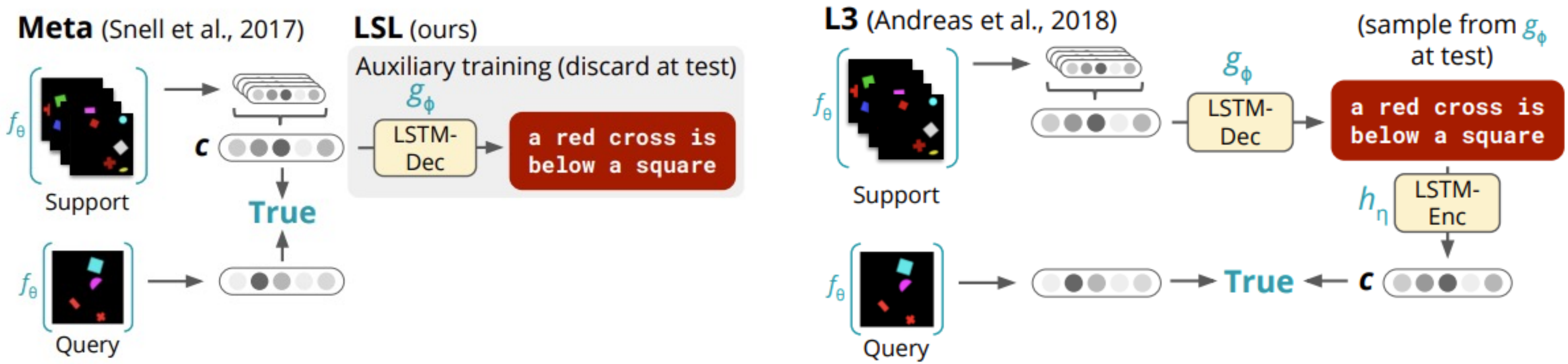
Co-learning via Generation

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



Co-learning via Generation

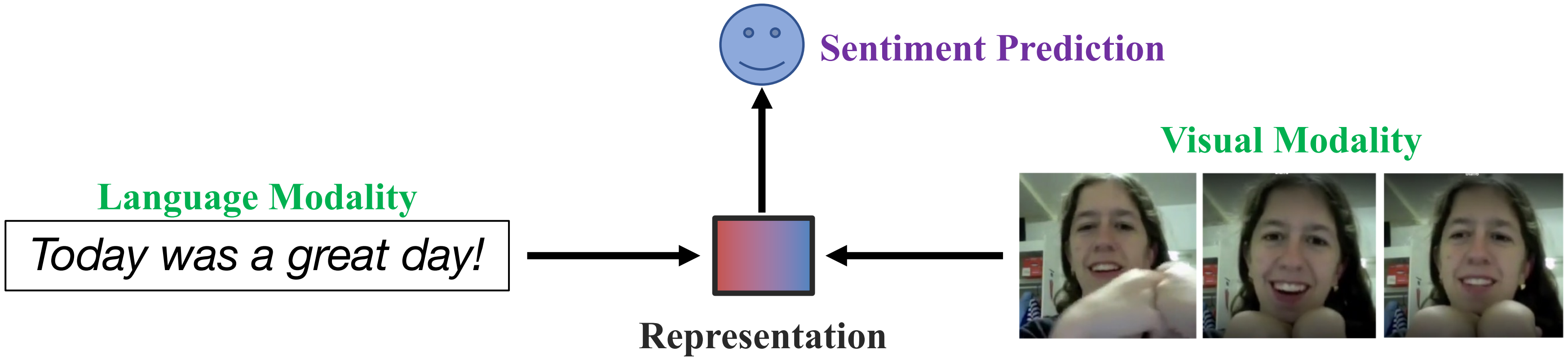
Image to text generation



[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]

Co-learning via Generation

Bimodal translations



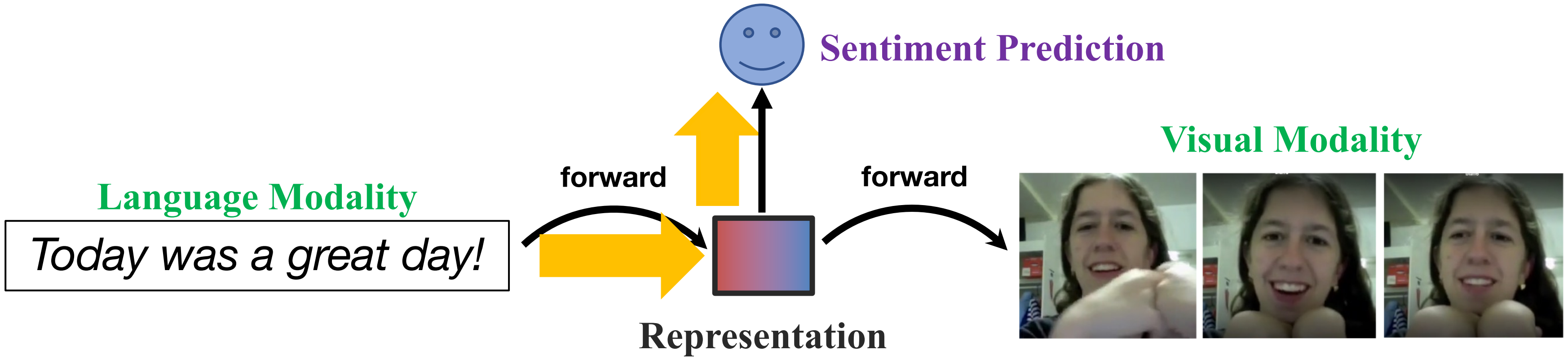
Both modalities required at test time!
Sensitive to noisy/missing visual modality.

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]

Co-learning via Generation

Bimodal translations

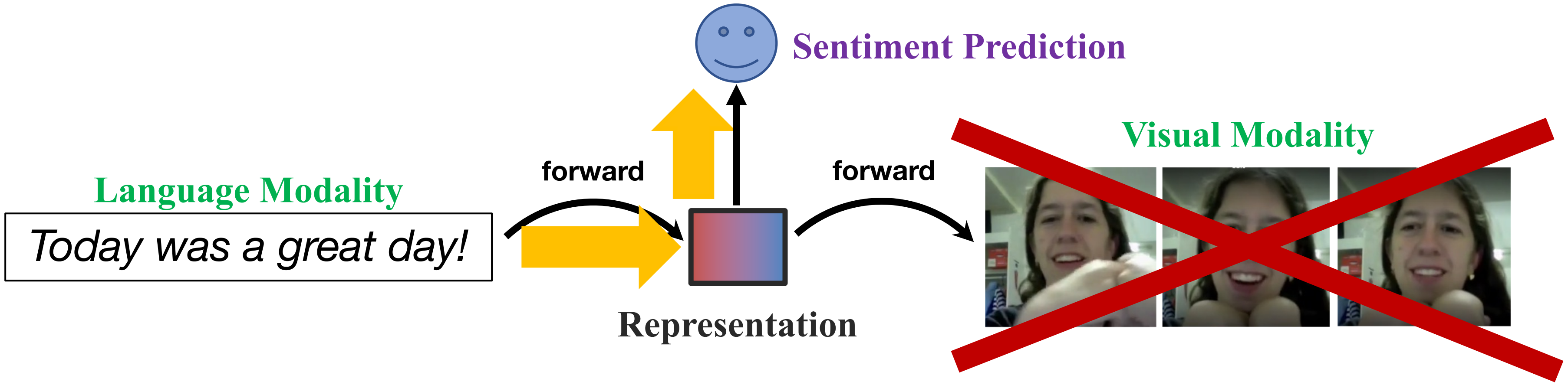


Cross-modal translation during training
Only language modality required at test time!

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

Co-learning via Generation

Bimodal translations

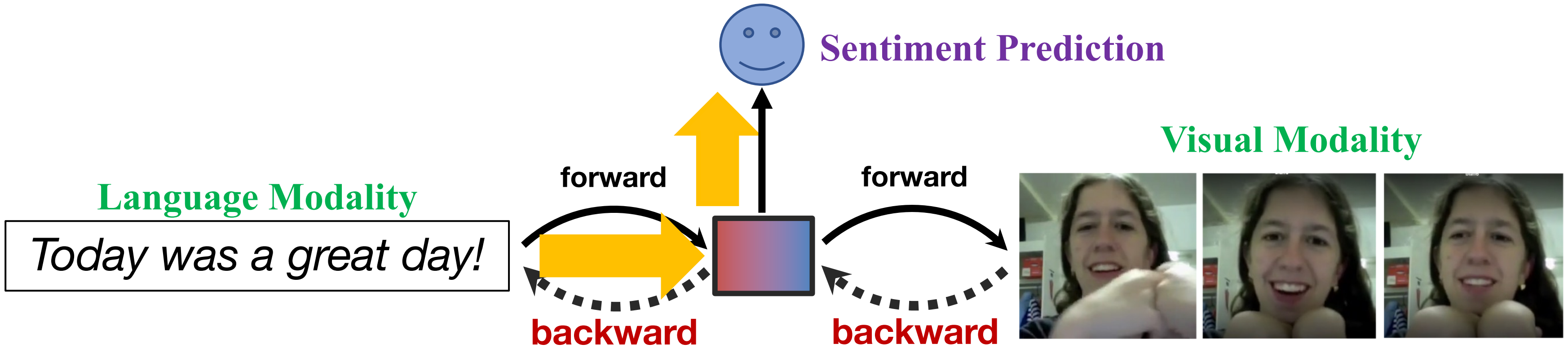


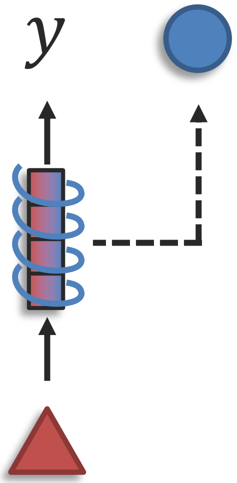
Problem: how do you ensure that both modalities are being used?

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

Co-learning via Generation

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 via Generation

Predicting images from corresponding language

Voken (visual token) classification

Masked language modeling

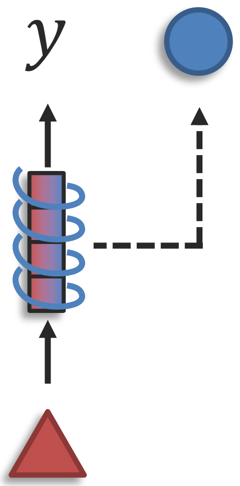
[learn] *[listening]*

BERT language model

Humans [mask] language by [mask] speaking

Only text used at test-time

Multimodal co-learning > language-only training

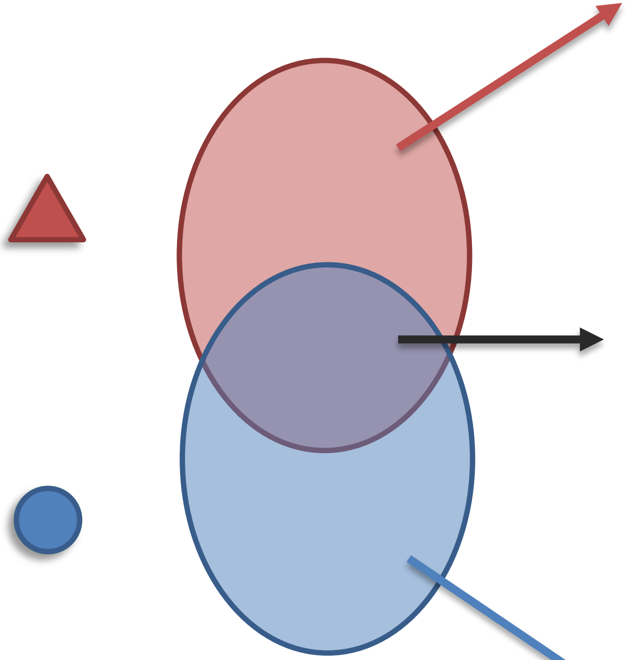


[Tan and Bansal, Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. EMNLP 2020]

Co-learning via Generation

Language

Visual
(image)



Information primarily in language modality

- Syntactic structure
- Vocabulary, morphology
- …

Information in both modalities

- Described people, objects, actions
- Illustrative gestures, motion
- …

Information primarily in visual modality

- Texture, visual appearance
- Depth, perspective, motion
- …

Co-learning may not always work…

Open challenges

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]

Co-learning may not always work…

Open challenges

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

"Remove gloss from furniture."

↓

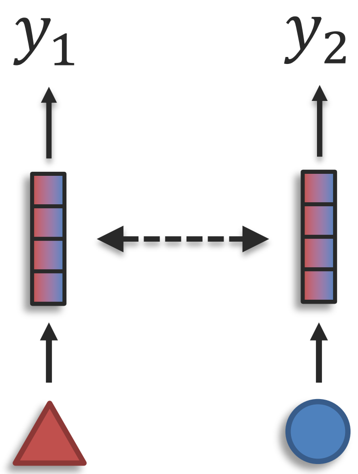
"Rub furniture with steel wool/cotton ball"

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

Sub-challenge 5c: Model Induction

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

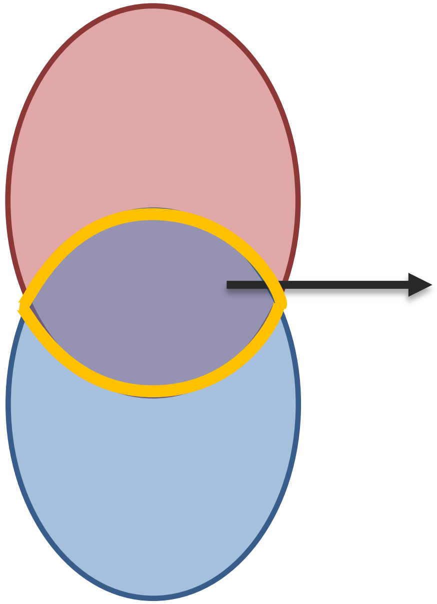
Model Induction



Sub-challenge 5c: Model Induction

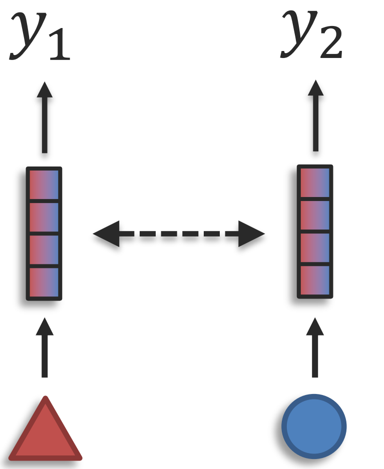
Language

Visual
(image)



Information in both modalities = Y

- Described people, objects, actions
- Illustrative gestures, motion



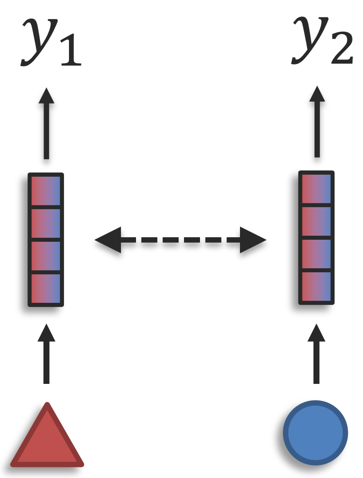
Ideally: $X_1 \perp X_2 \mid Y$.
Or equivalently: $I(X_1; X_2 \mid Y) = 0$.

Multi-view redundancy assumption

Co-training

Setup

Ideally: $X_1 \perp X_2 \mid Y$.
Or equivalently: $I(X_1; X_2 | Y) = 0$.



Multi-view redundancy assumption

1. X_1 = text on the web page.
2. X_2 = text on hyperlinks pointing into the web page.
3. Y = category of web page: academic, sports, news, music etc.

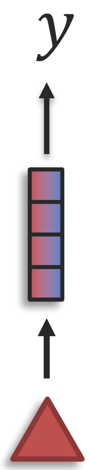
Sufficiency assumption

$X_1 \rightarrow Y$ is learnable given enough data
$X_2 \rightarrow Y$ is learnable given enough data

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

Self-training

Warmup: a single view – Self-training



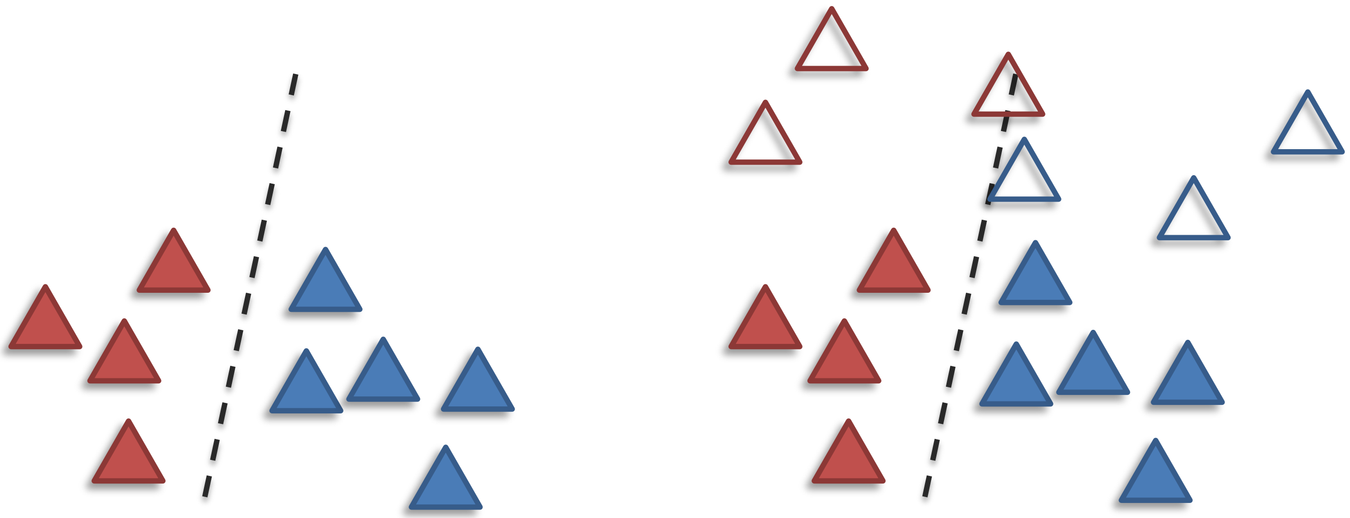
Assume:
1. Labeled data $\{X_1^L, Y\}$.
2. Unlabeled data $\{X_1^U\}$.

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 $f_1(X_1)$.

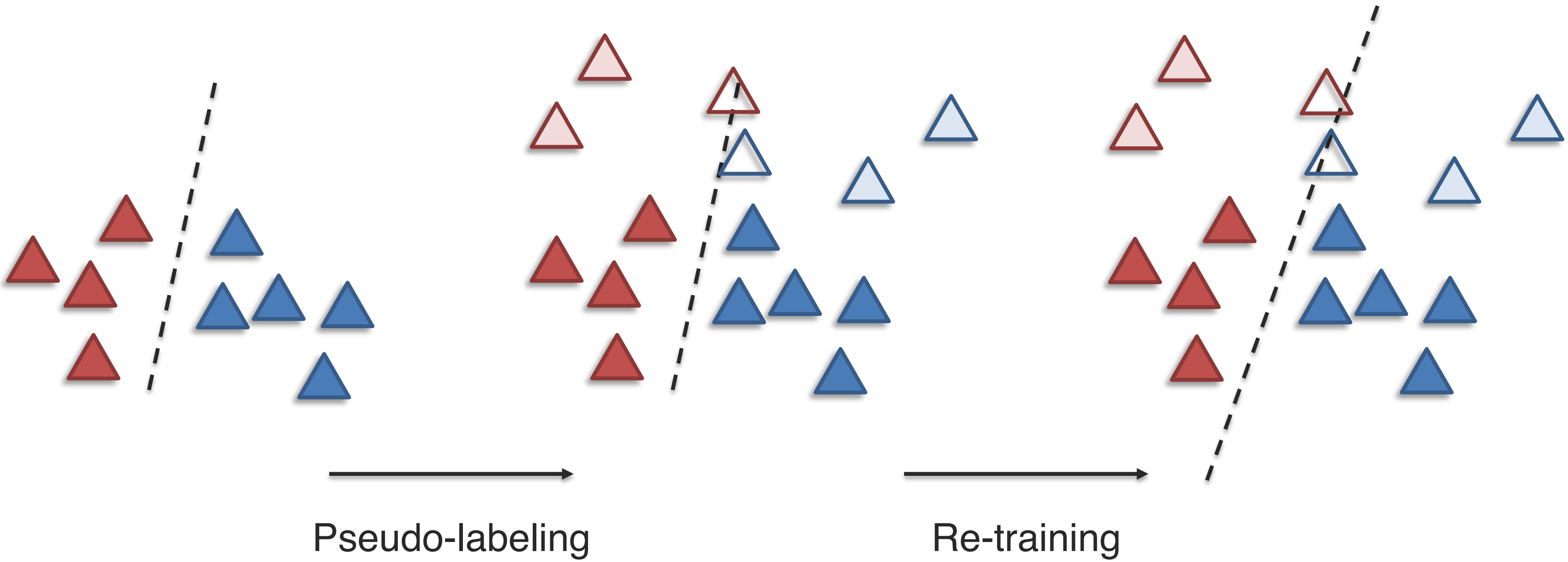
Self-training

Warmup: a single view – Self-training



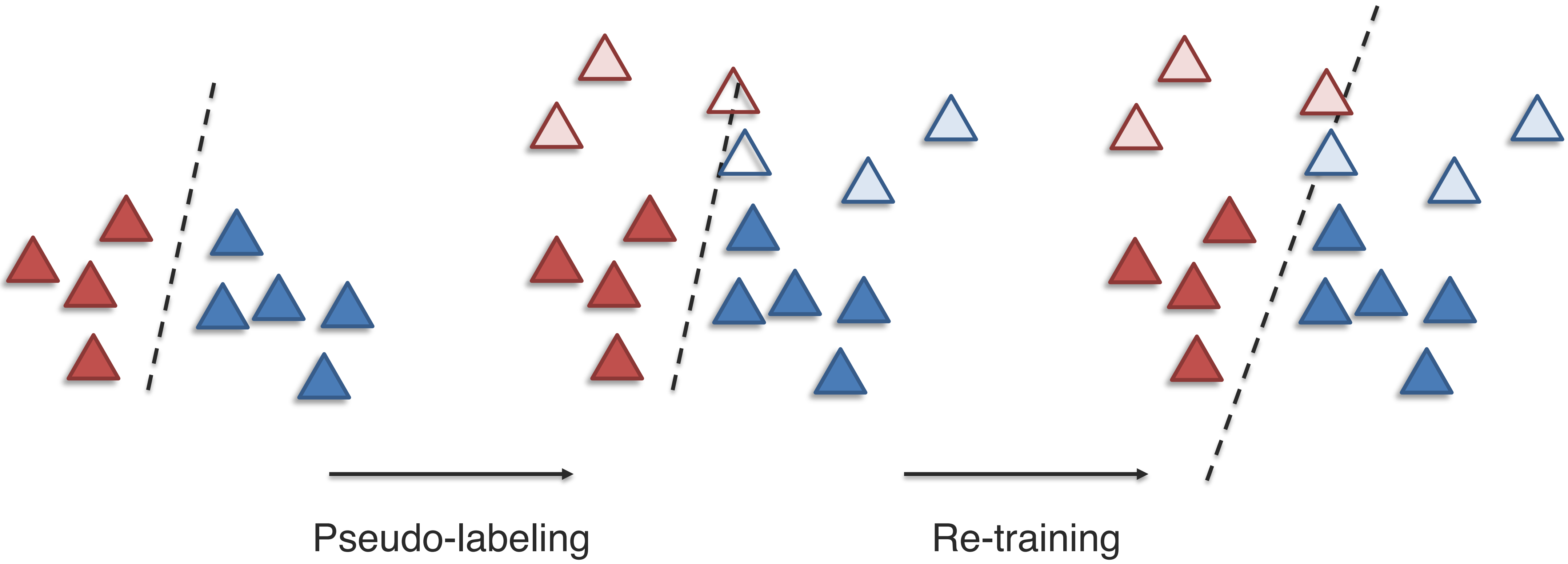
Self-training

Warmup: a single view – Self-training



Self-training

Warmup: a single view – Self-training

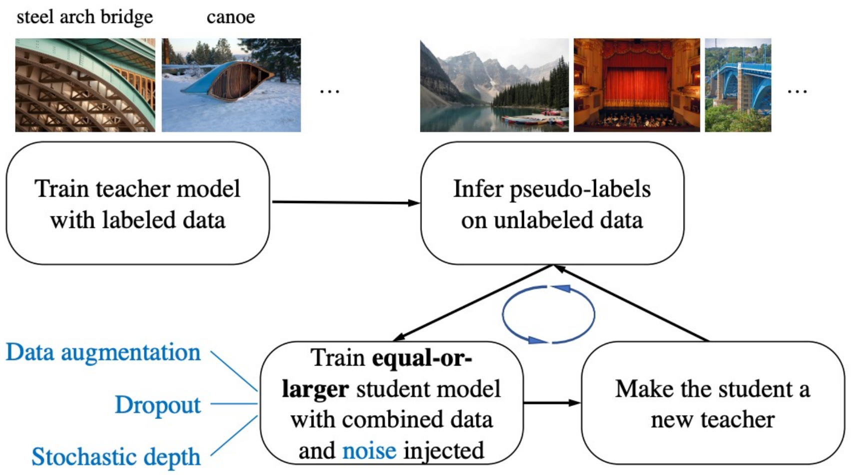


Self-training

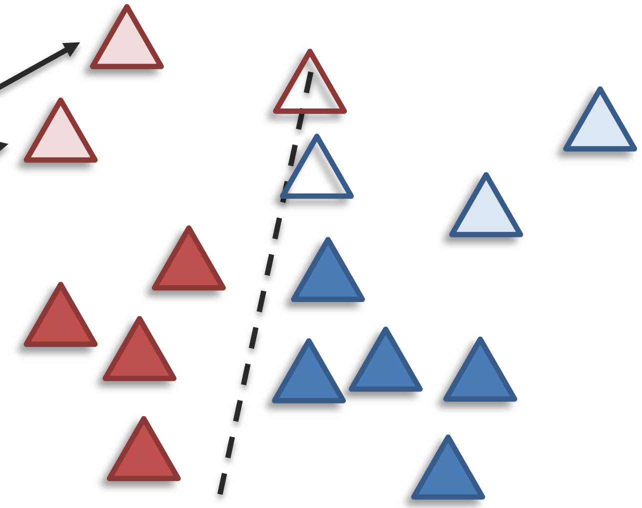
Key-words: semi-supervised learning, label propagation, domain adaptation/shift

Critical:

1. Can't label all unlabeled data in one step, or you recover original classifier just trained on labeled data.
2. Sequence of pseudo-labeling is important to gradually shift classification boundary.
3. Input consistency regularization: shape of data space is important – implicit assumption that similar datapoints have similar labels (i.e., label consistency)



Input consistency:
- Data augmentation
- Adding noise



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

Co-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)

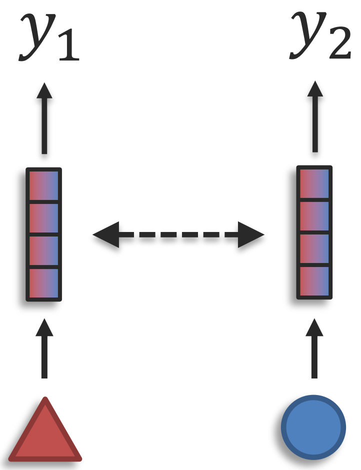
Assumptions:

1. Either view is sufficient to predict the label alone, with enough data
2. Views should be as independent as possible: examples where f_1 has high confidence but not f_2 and vice-versa.

[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 $\{X_2^U, Y = f_1(X_1^U)\}$.
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:

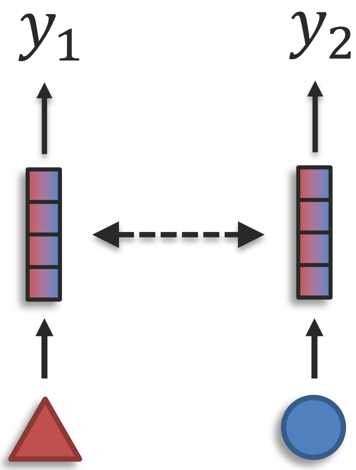
1. For a new unlabeled sample $\{X_1, X_2\}$, ensemble $f_1(X_1)$ and $f_2(X_2)$.

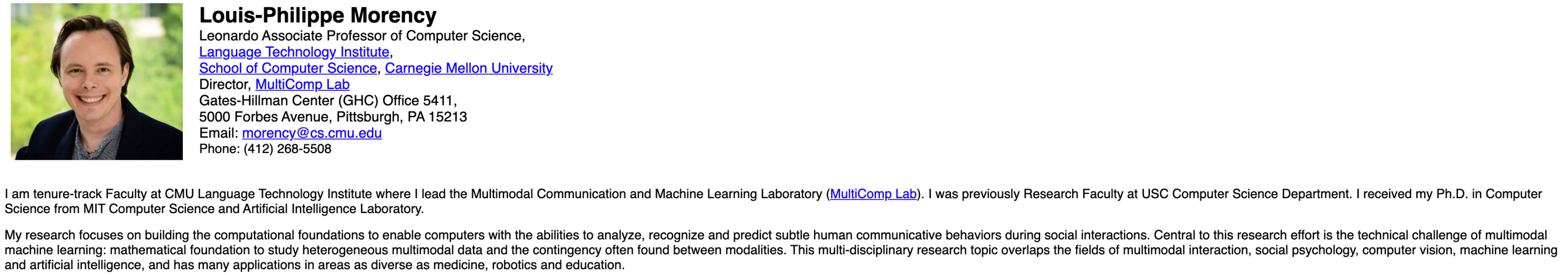
[Blum and Mitchell, Combining Labeled and Unlabeled Data with Co-Training. COLT 1998]

Co-training

Co-training

1. X_1 = text on the web page, X_2 = text on hyperlinks pointing into the web page.
2. Y = category of web page: academic, sports, news, music etc.



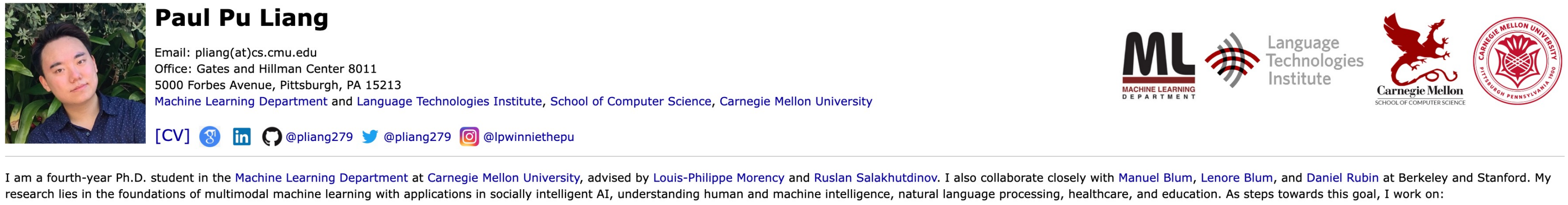


Louis-Philippe Morency
Leonardo Associate Professor of Computer Science,
Language Technology Institute,
School of Computer Science, Carnegie Mellon University
Director, MultiComp Lab
Gates-Hillman Center (GHC) Office 5411,
5000 Forbes Avenue, Pittsburgh, PA 15213
Email: morency@cs.cmu.edu
Phone: (412) 268-5508

I am tenure-track Faculty at CMU Language Technology Institute where I lead the Multimodal Communication and Machine Learning Laboratory (MultiComp Lab). I was previously Research Faculty at USC Computer Science Department. I received my Ph.D. in Computer Science from MIT Computer Science and Artificial Intelligence Laboratory.

My research focuses on building the computational foundations to enable computers with the abilities to analyze, recognize and predict subtle human communicative behaviors during social interactions. Central to this research effort is the technical challenge of multimodal machine learning: mathematical foundation to study heterogeneous multimodal data and the contingency often found between modalities. This multi-disciplinary research topic overlaps the fields of multimodal interaction, social psychology, computer vision, machine learning and artificial intelligence, and has many applications in areas as diverse as medicine, robotics and education.

Labeled, learn that 'X_1(LP) = CMU -> academic' and 'X_2(Paul -> LP) = advised by -> academic'



Paul Pu Liang
Email: pliang(at)cs.cmu.edu
Office: Gates and Hillman Center 8011
5000 Forbes Avenue, Pittsburgh, PA 15213
Machine Learning Department and Language Technologies Institute, School of Computer Science, Carnegie Mellon University

[CV] @pliang279 @pliang279 @lpwinniethepu

I am a fourth-year Ph.D. student in the Machine Learning Department at Carnegie Mellon University, advised by Louis-Philippe Morency and Ruslan Salakhutdinov. I also collaborate closely with Manuel Blum, Lenore Blum, and Daniel Rubin at Berkeley and Stanford. My research lies in the foundations of multimodal machine learning with applications in socially intelligent AI, understanding human and machine intelligence, natural language processing, healthcare, and education. As steps towards this goal, I work on:

Unlabeled, label using 'f_1: X_1(Paul) = CMU -> academic' and learn that 'X_2(MLD -> Paul) = PhD program -> academic'

Another student -> **Unlabeled**, label using 'f_2: X_2(Berkeley CS -> student) = 'PhD program -> academic'
and learn that 'X_1(student) = robotics -> academic'

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

Co-training

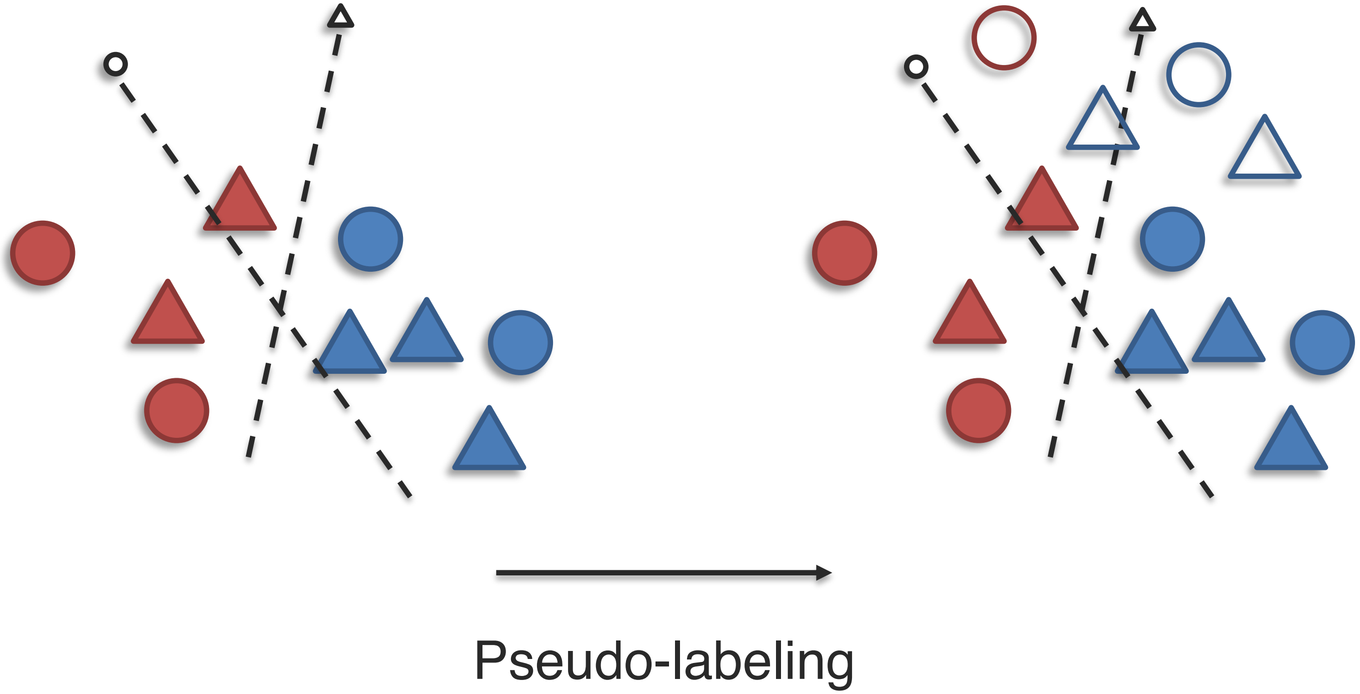
From self-training to co-training

Assumptions:
1. Either view is sufficient to predict the label alone.
2. Views should be as independent as possible: examples where f_1 has high confidence but not f_2 and vice-versa.

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

Co-training

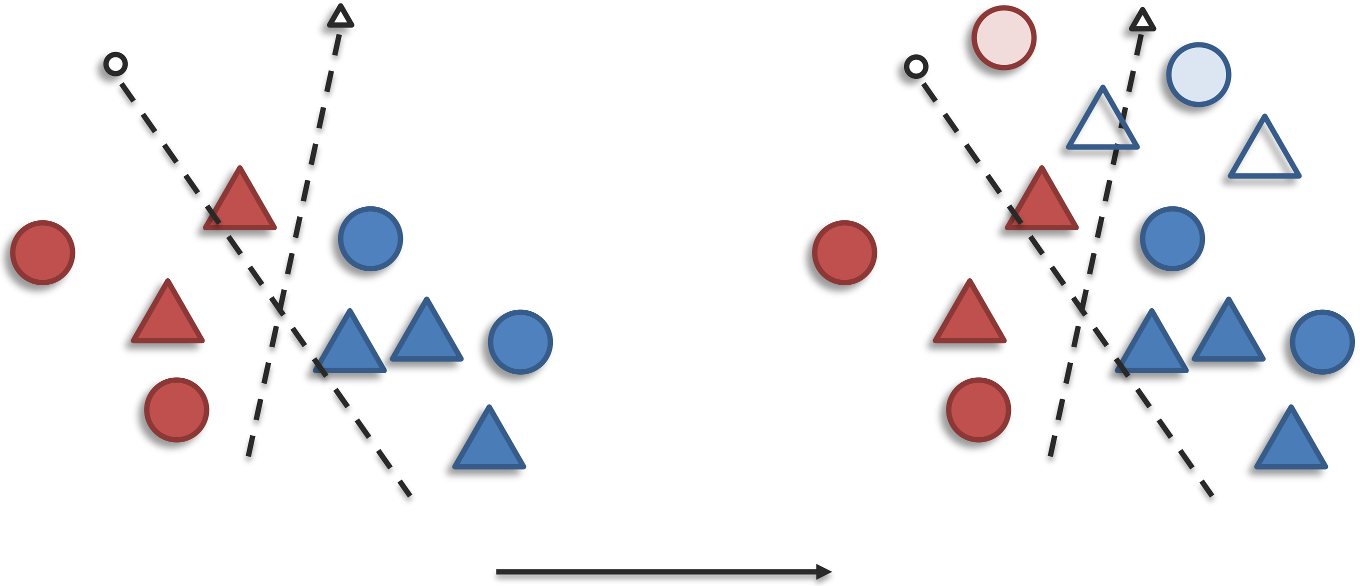
From self-training to co-training



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

Co-training

From self-training to co-training

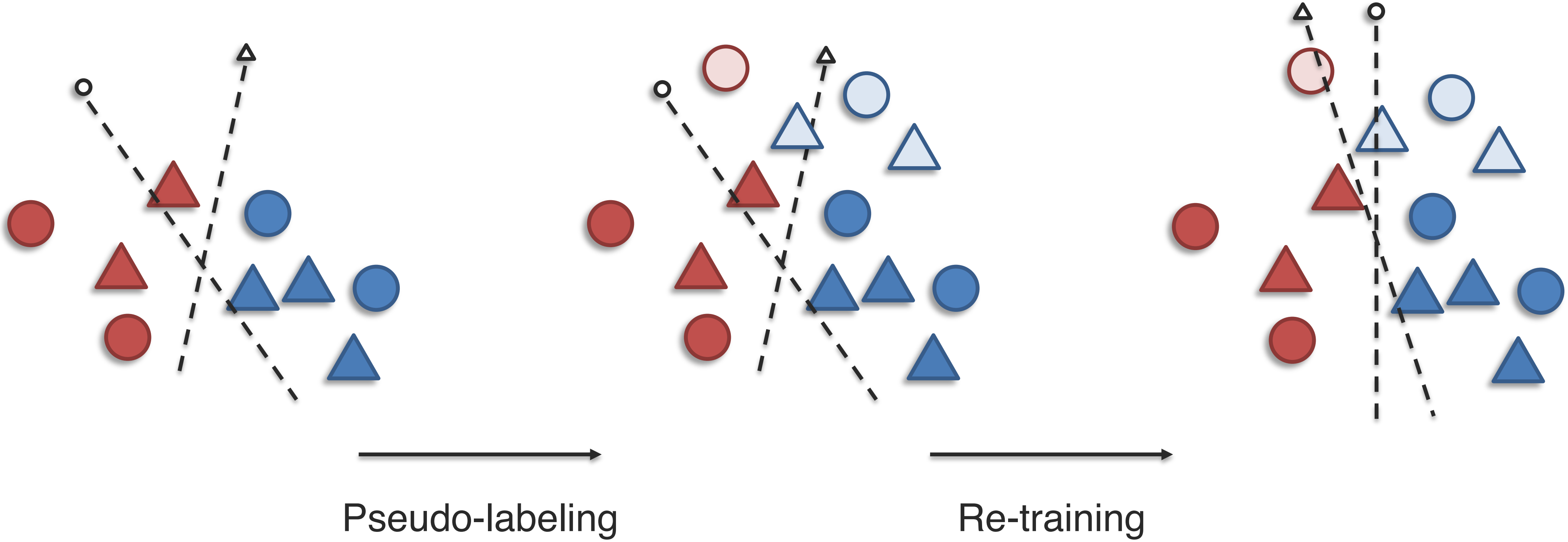


Pseudo-labeling

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

Co-training

From self-training to co-training
Key idea: functions on both views must be compatible and agree



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

Co-training

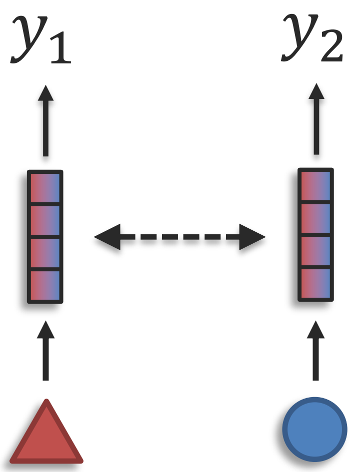
From self-training to co-training
Key idea: functions on both views must be compatible and agree

Intuitions:
1. Either view is sufficient to predict the label alone.
2. Views should be as independent as possible: examples where f_1 has high confidence but not f_2 and vice-versa.
3. Input consistency regularization: shape of data space is important – implicit assumption that similar datapoints have similar labels (i.e., label consistency).
→ In co-training, data from another view help us to supplement the label space!
→ Both views must agree = input consistency which enables cross-view pseudo-labeling.
4. Eventually, will converge on 2 classifiers that agree and each separate both views.

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

Co-training

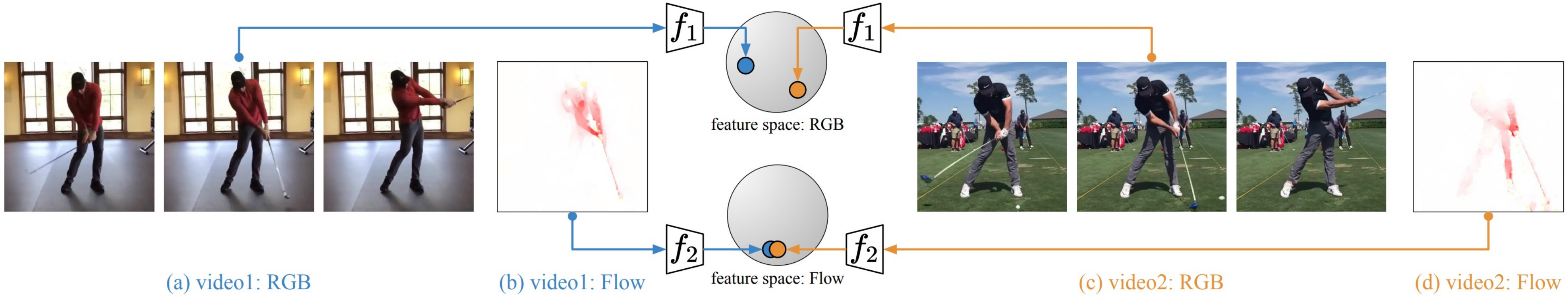
Recent applications of co-training



Self-supervised learning with positive and negative samples

→ 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.

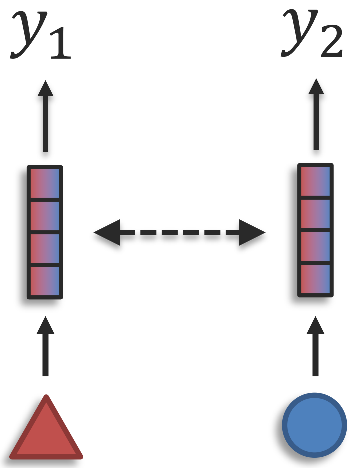


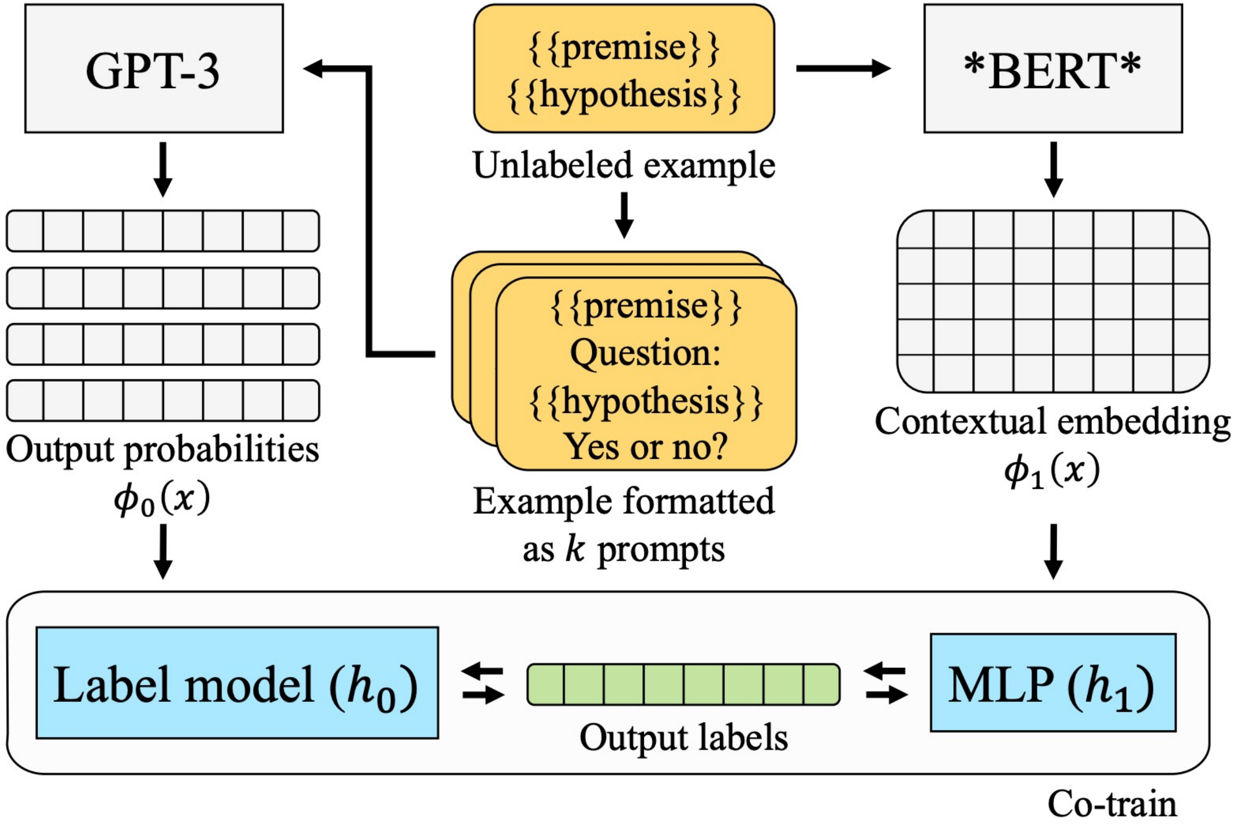
[Han et al., Self-supervised Co-training for Video Representation Learning. NeurIPS 2020]

Co-training

Recent applications of co-training

Language-model prompting

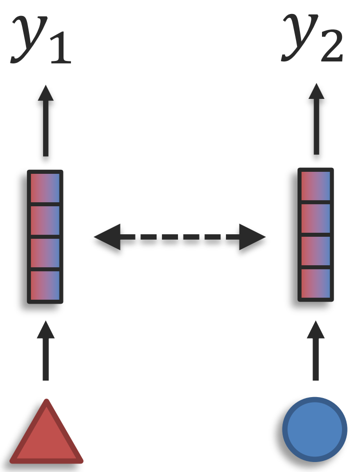




[Lang et al., Co-training Improves Prompt-based Learning for Large Language Models. ICML 2022]

Co-Regularization

Co-regularization



Add a loss term to ensure both model predictions are similar:

$$L = (f_1(X_1) - f_2(X_2))^2$$

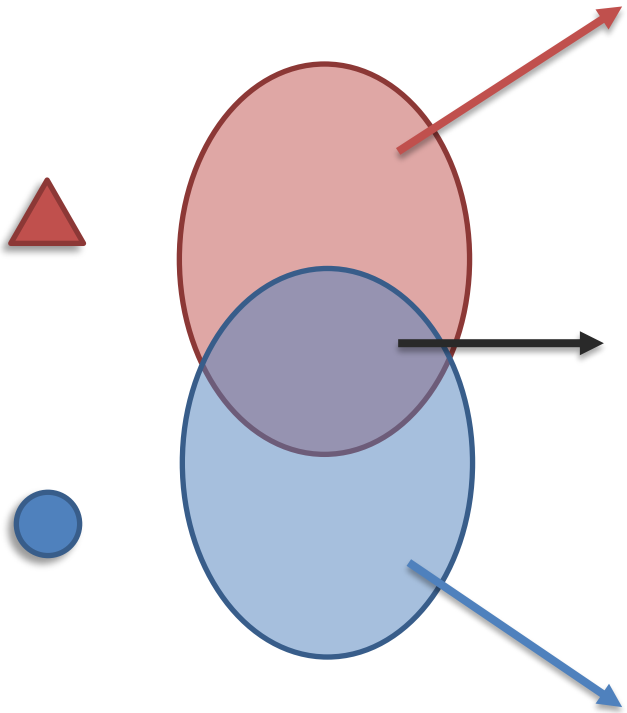
Recall representation coordination.

[Sridharan and Kakade, An Information Theoretic Framework for Multi-view Learning. COLT 2008]

Sub-challenge 5c: Model Induction

Language

Visual
(image)



Information primarily in language modality

- Syntactic structure
- Vocabulary, morphology
- …

Information in both modalities

- Described people, objects, actions
- Illustrative gestures, motion
- …

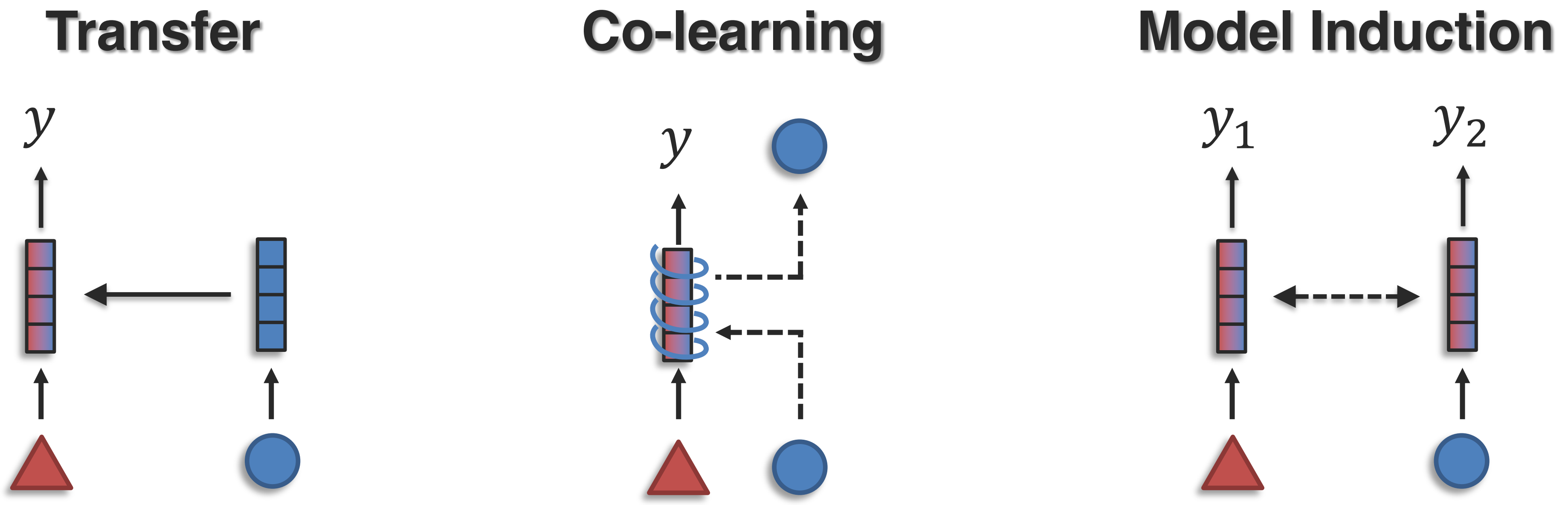
Information primarily in visual modality

- Texture, visual appearance
- Depth, perspective, motion
- …

Summary: Transference

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

Sub-challenges:



More Transference

Open challenges

Many more dimensions of transfer:
→ Multimodal {multitask, transfer, few-shot, meta} learning.
→ Domain adaptation, domain shift, label shift.
→ Core: representation, alignment, reasoning!

Open challenges:

- Low-resource: little downstream data, lack of paired data, robustness (next section).
- Settings where SOTA unimodal encoders are not deep learning e.g., tabular data.
- Evaluating reasoning and robustness and large models.
- Limits of transfer beyond redundancy/joint information.
- Interpretability (next section).