Multimodal AI

Lecture 4.2 – Multimodal Alignment

Paul Liang
Assistant Professor
MIT Media Lab & MIT EECS

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

mit media lab

MIT EECS

Assignments for This Coming Week

I want to meet every group at least once regarding their project ideas. Project mentors will be assigned.

Compute credits: 40 x $50 Kimi credits, 40 x $40 other credits.

HW2 due next Wednesday 3/4.

Today's lecture

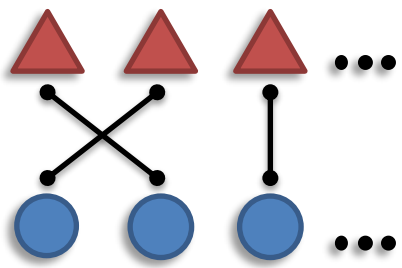
1. Multimodal alignment
2. Explicit alignment and contrastive learning
3. Continuous alignment
4. Implicit (emergent) alignment

Challenge 2: Alignment

Definition: Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure.

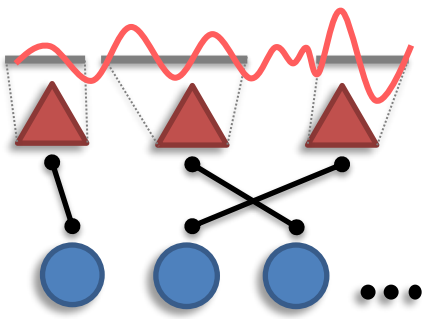
Sub-challenges:

Discrete connections



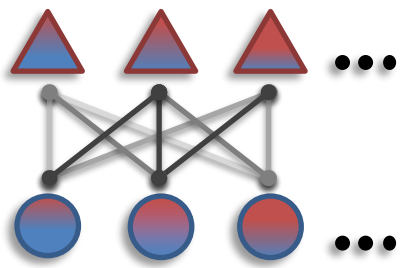
Explicit alignment
(e.g., grounding)

Continuous alignment



Granularity of individual elements

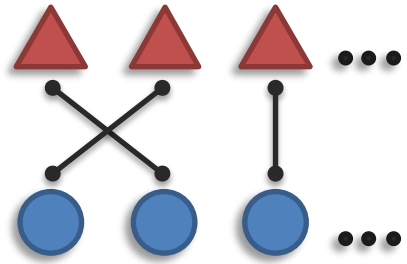
Contextualized representation



Implicit alignment
+ representation

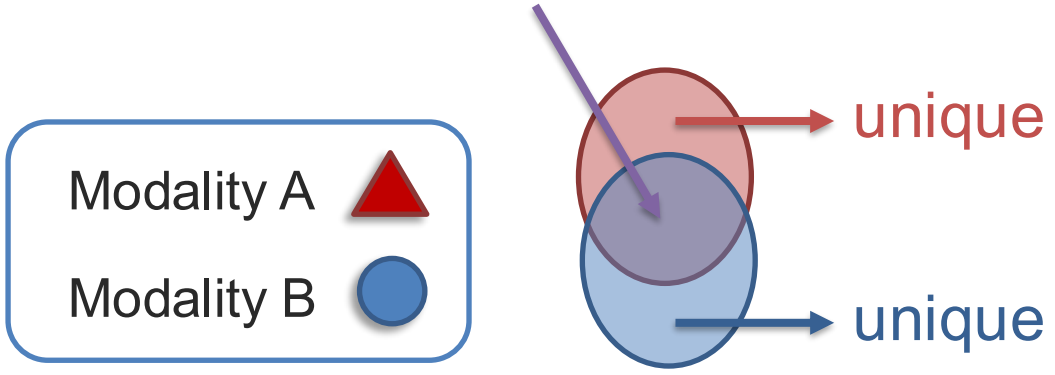
multisensory intelligence

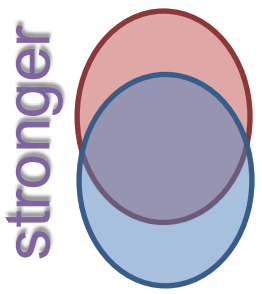
Challenge 2a: Discrete Alignment

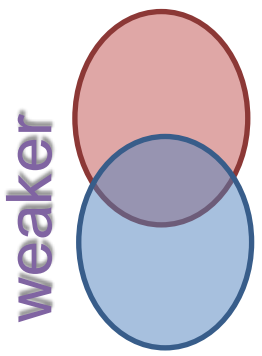


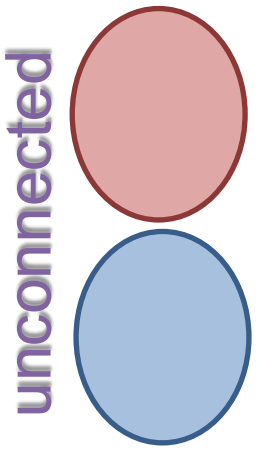
Definition: Identify and model connections between elements of multiple modalities

Shared information that relates modalities



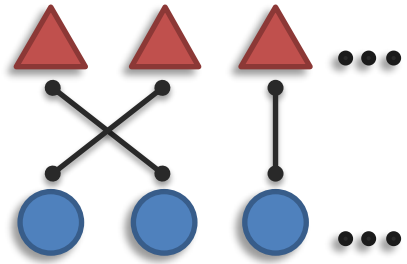






multisensory intelligence

Modality Connections



Definition: Tying language (words, phrases,…) to non-linguistic elements, such as the visual world (objects, people, …)

Statistical

Association

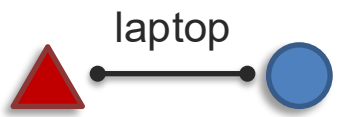
e.g., correlation, co-occurrence

Dependency

e.g., causal, temporal

Semantic

Correspondence

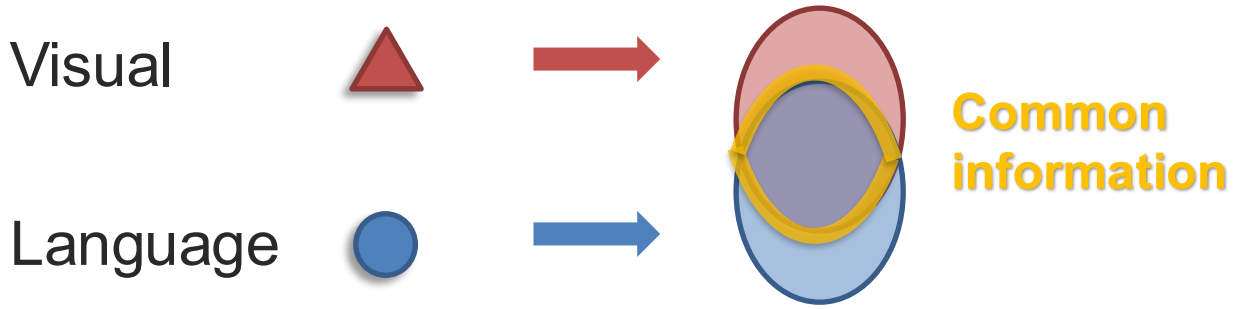


e.g., grounding

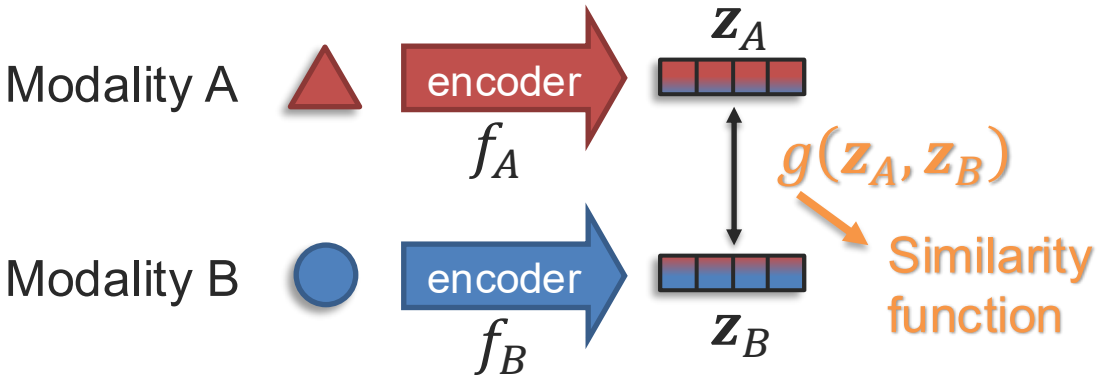
Relationship

e.g., function

Modality Connections

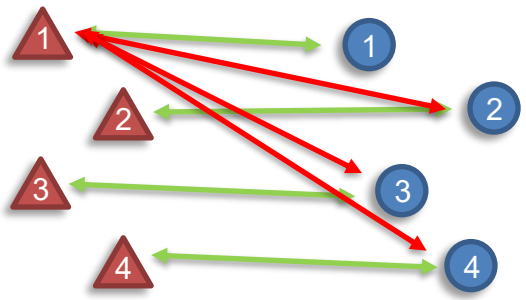


Learning aligned representations:

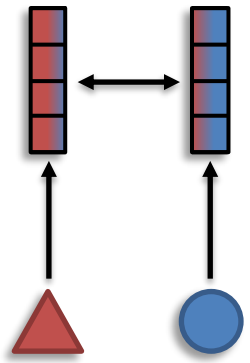


or contrastive learning

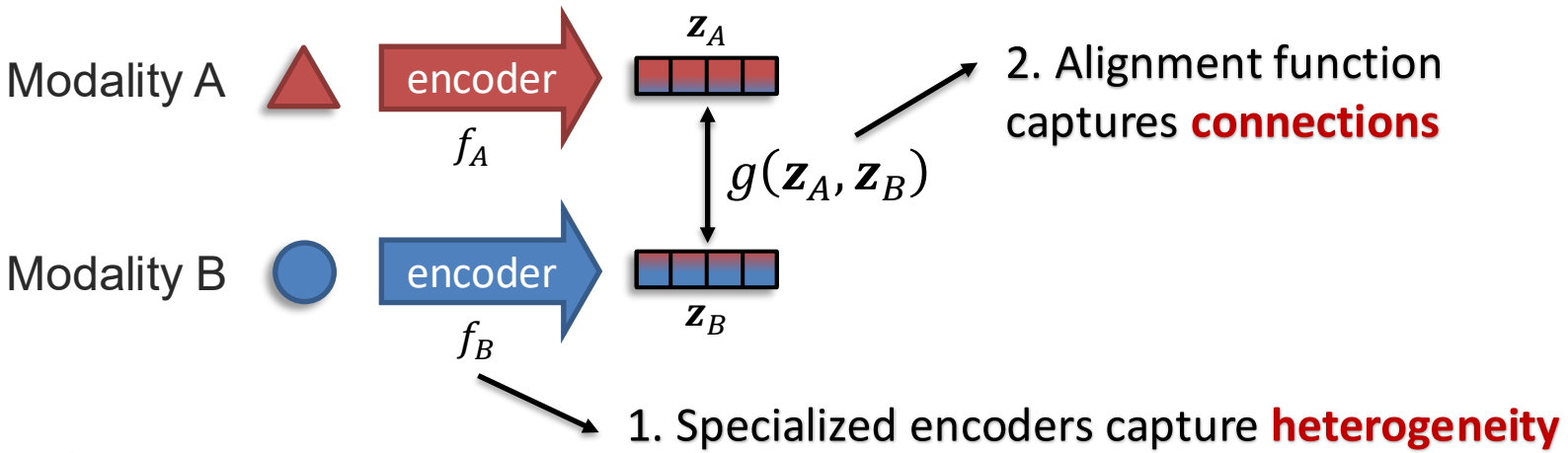
Supervision: Paired data



Aligned Representations



Definition: Learn multimodal representations aligned through their connections.



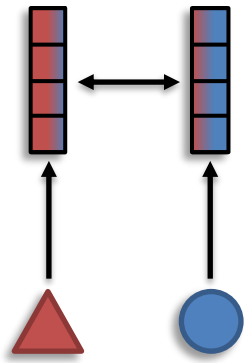
Learning with alignment function:

$$\mathcal{L} = g(f_A(\blacktriangle), f_B(\bullet))$$

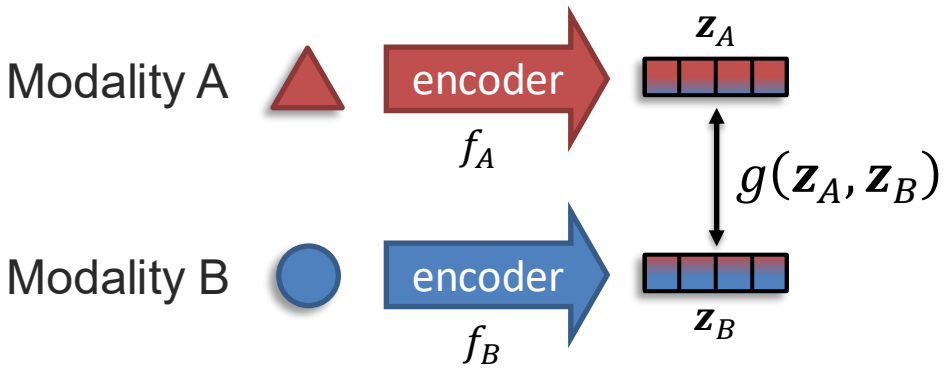
with model parameters θ_g, θ_{f_A} and θ_{f_B}

multisensory intelligence

Aligned Representations



Definition: Learn multimodal representations aligned through their connections.



Learning with alignment function:

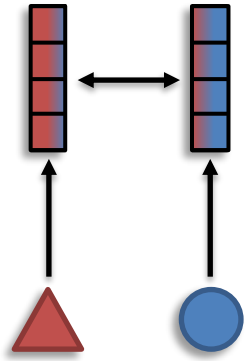
$$\mathcal{L} = g(f_A(\blacktriangle), f_B(\bullet))$$

with model parameters θ_g, θ_{f_A} and θ_{f_B}

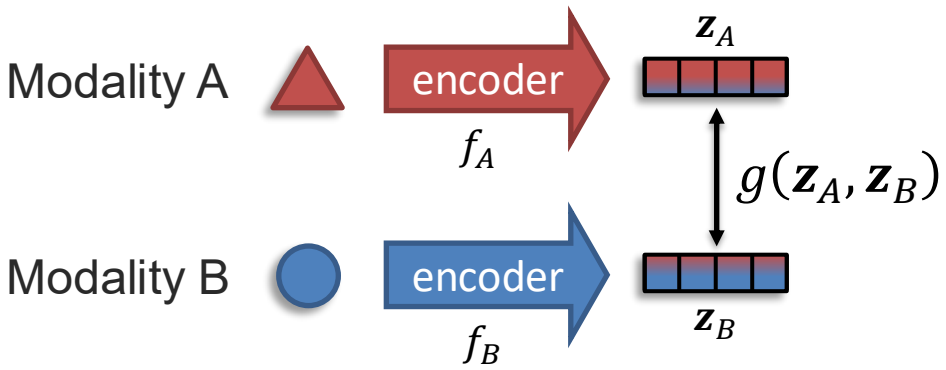
① Cosine similarity: $$g(\boldsymbol{z}_A, \boldsymbol{z}_B) = \frac{\boldsymbol{z}_A \cdot \boldsymbol{z}_B}{\|\boldsymbol{z}_A\| \|\boldsymbol{z}_B\|}$$

multisensory intelligence

Aligned Representations



Definition: Learn multimodal representations aligned through their connections.



Learning with alignment function:

$$\mathcal{L} = g(f_A(\blacktriangle), f_B(\bullet))$$

with model parameters θ_g, θ_{f_A} and θ_{f_B}

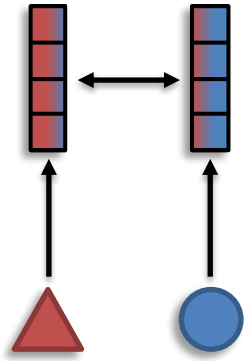
② Kernel similarity functions:

$$g(\boldsymbol{z}_A, \boldsymbol{z}_B) = k(\boldsymbol{z}_A, \boldsymbol{z}_B)$$

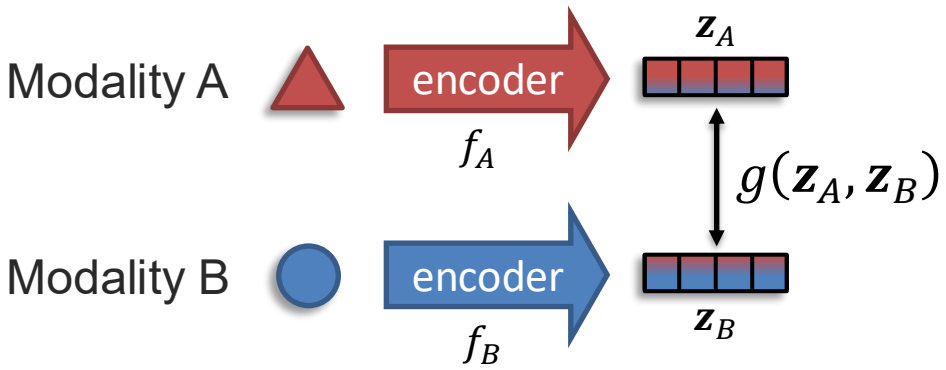
- Linear
- Polynomial
- Exponential
- RBF

multisensory intelligence

Aligned Representations



Definition: Learn multimodal representations aligned through their connections.



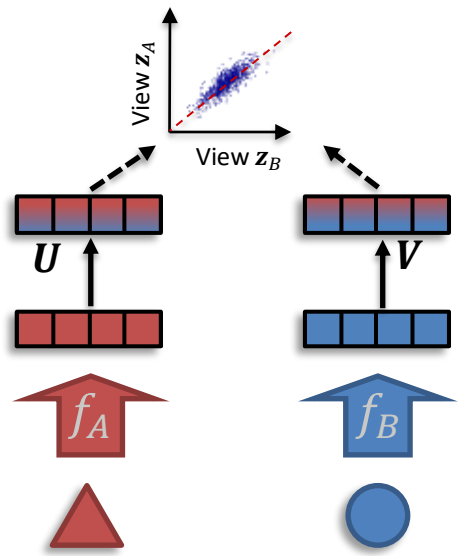
Learning with alignment function:

$$\mathcal{L} = g\big(f_A(\blacktriangle), f_B(\bullet)\big)$$

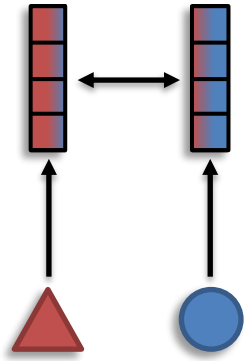
with model parameters θ_g, θ_{f_A} and θ_{f_B}

③ Canonical Correlation Analysis (CCA):

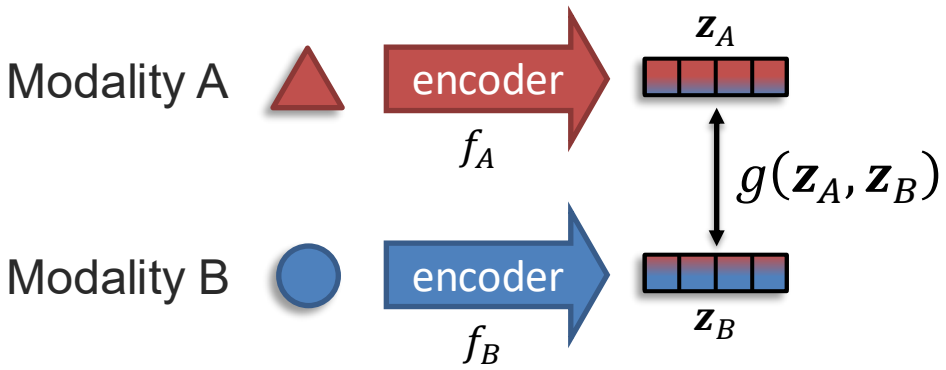
$$\underset{\boldsymbol{V},\boldsymbol{U},f_A,f_B}{\operatorname{argmax}}\, corr(\boldsymbol{z}_A, \boldsymbol{z}_B)$$



Aligned Representations



Definition: Learn multimodal representations aligned through their connections.

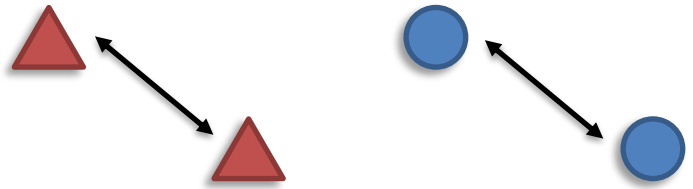


Learning with alignment function:

$$\mathcal{L} = g(f_A(\blacktriangle), f_B(\bullet))$$

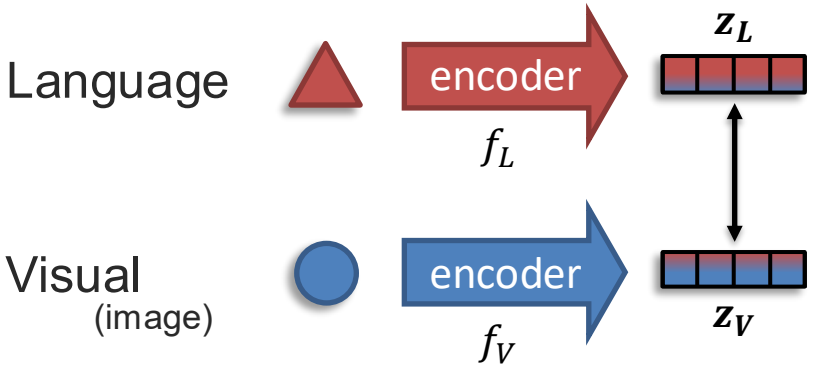
with model parameters θ_g, θ_{f_A} and θ_{f_B}

④ Order, hierarchy, pairwise relationships.

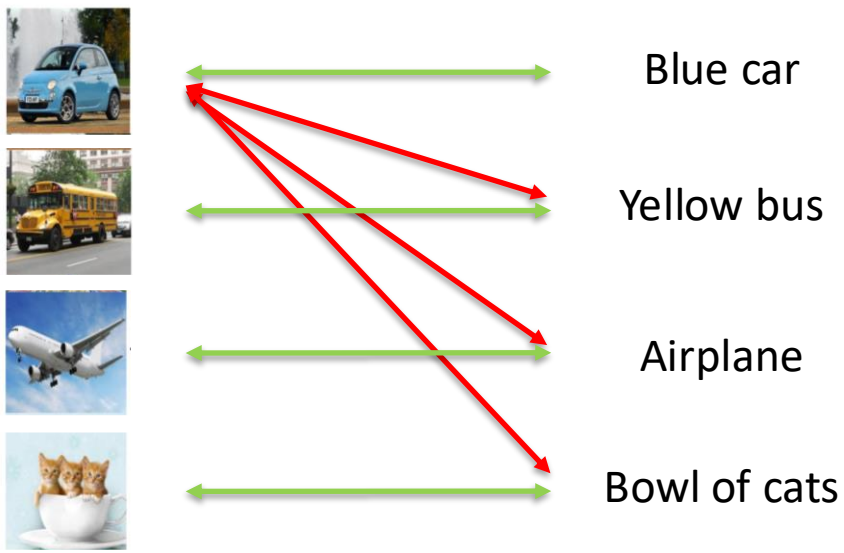


multisensory intelligence

Alignment with Contrastive Learning



Paired data: {▲, ●} (e.g., images and text)



Contrastive loss:

➡ brings positive pairs closer and pushes negative pairs apart

Simple contrastive loss:

$$\max\{0, \alpha + g(\boldsymbol{z}_A, \boldsymbol{z}_B^+) - g(\boldsymbol{z}_A, \boldsymbol{z}_B^-)\}$$

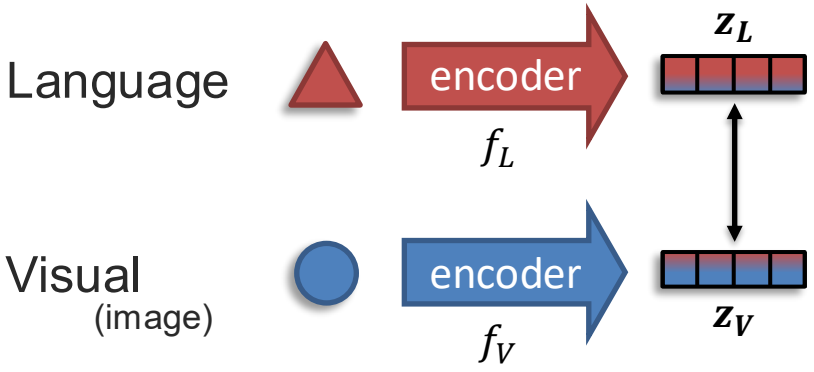
positive pairs

negative pair

Coordination function
(e.g., cosine similarity)

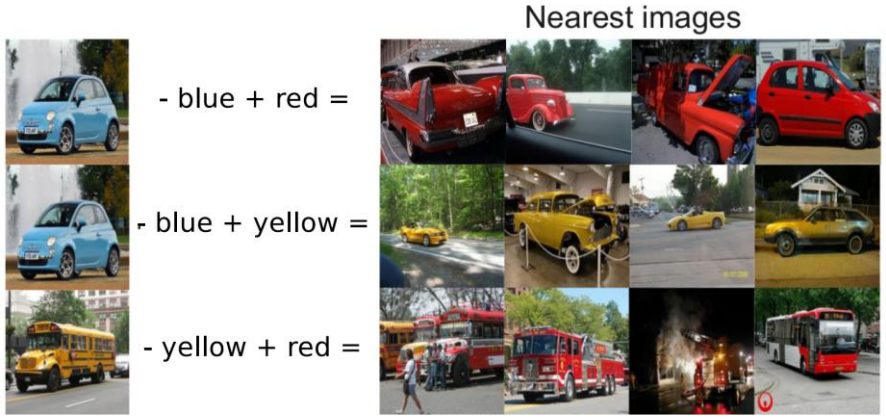
multisensory intelligence

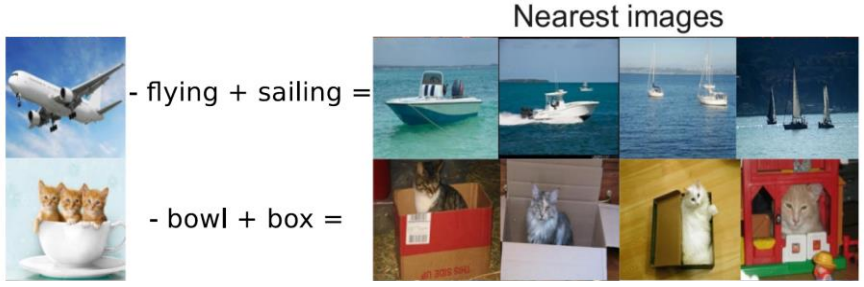
Visual-Semantic Embeddings



Contrastive loss:

➡ brings positive pairs closer and pushes negative pairs apart

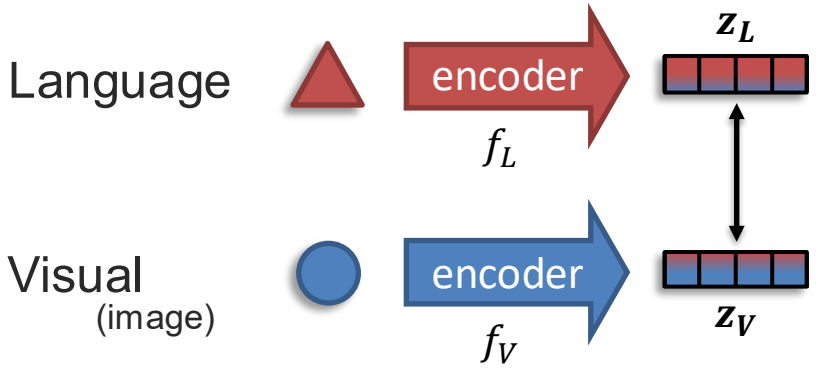




[Kiros et al., Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. NeurIPS 2014]

multisensory intelligence

Contrastive Language Image Pretraining



Popular contrastive loss: InfoNCE

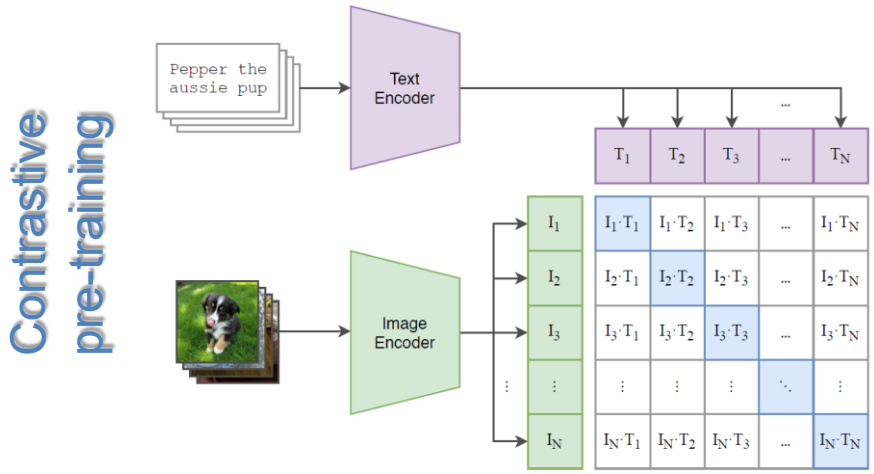
$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\text{sim}(\boldsymbol{z}_A^l, \boldsymbol{z}_B^l)}{\sum_{i=1}^{N} \text{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)}$$

positive pairs

negative pairs and positive pairs

Similarity function can be cosine similarity

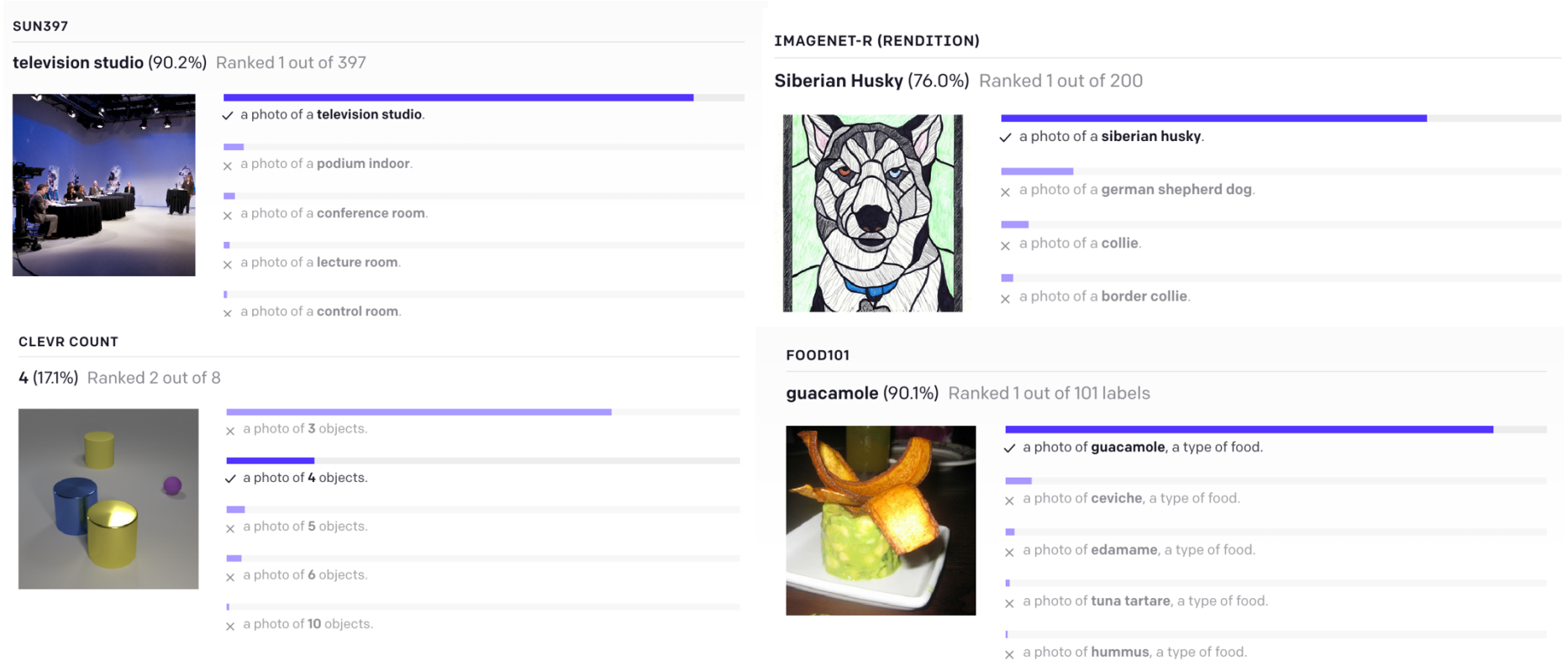
Positive and negative pairs:



- CLIP encoders (f_L and f_V) are great for language-vision tasks
- z_L and z_V are coordinated but not identical representation spaces

[Radford et al., Learning Transferable Visual Models From Natural Language Supervision. ICML 2021]

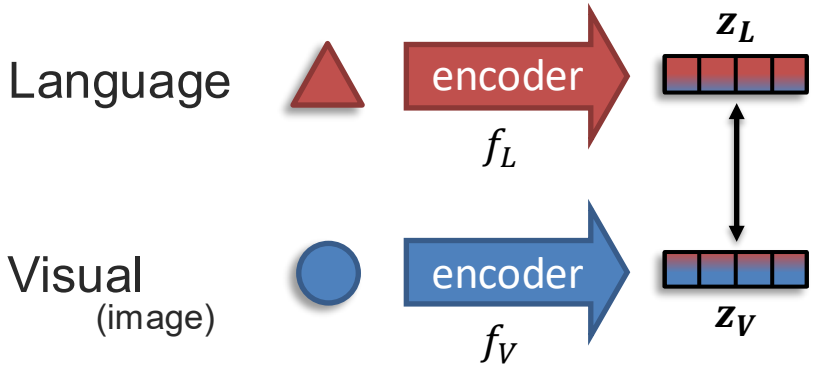
CLIP (Contrastive Language–Image Pre-training)



[Radford et al., Learning Transferable Visual Models From Natural Language Supervision. ICML 2021]

multisensory intelligence

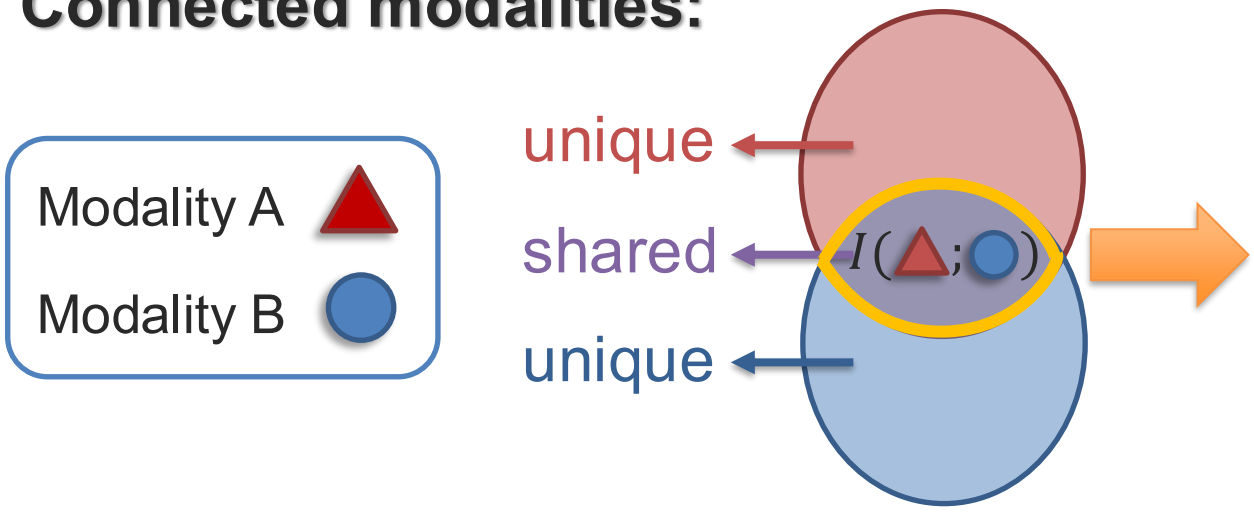
Contrastive Learning and Connected Modalities



Popular contrastive loss: InfoNCE

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\text{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^i)}{\sum_{j=1}^{N}\text{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)}$$

Connected modalities:



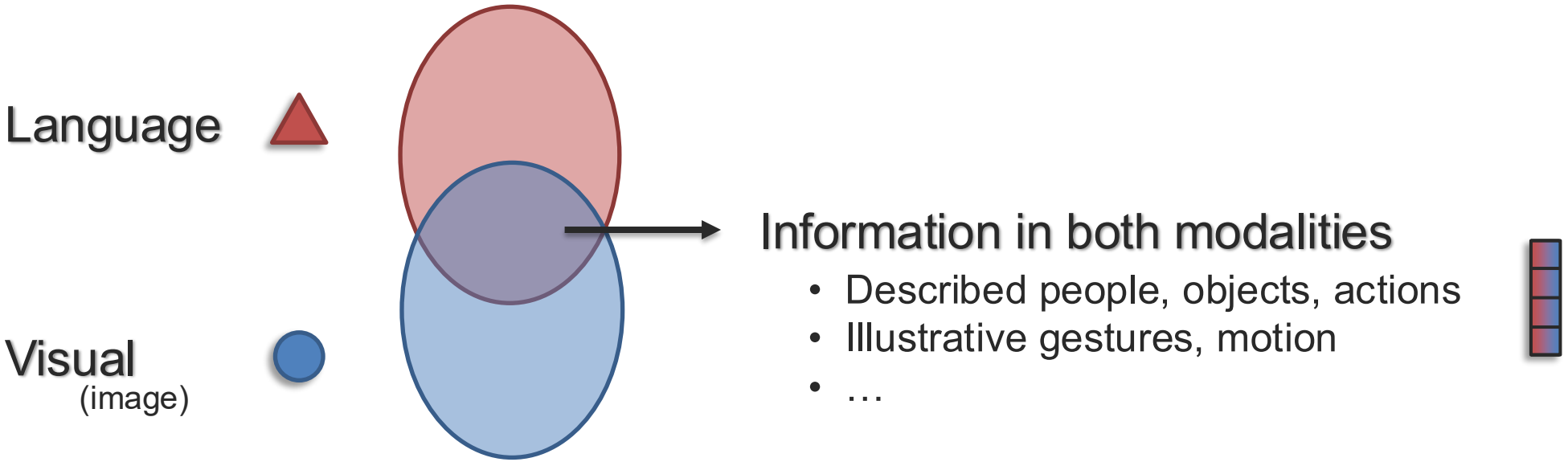
CLIP focuses on shared connections

Mutual information $I(X;Y)$

$$\mathbb{E}_{X,Y}\left[\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\right]$$

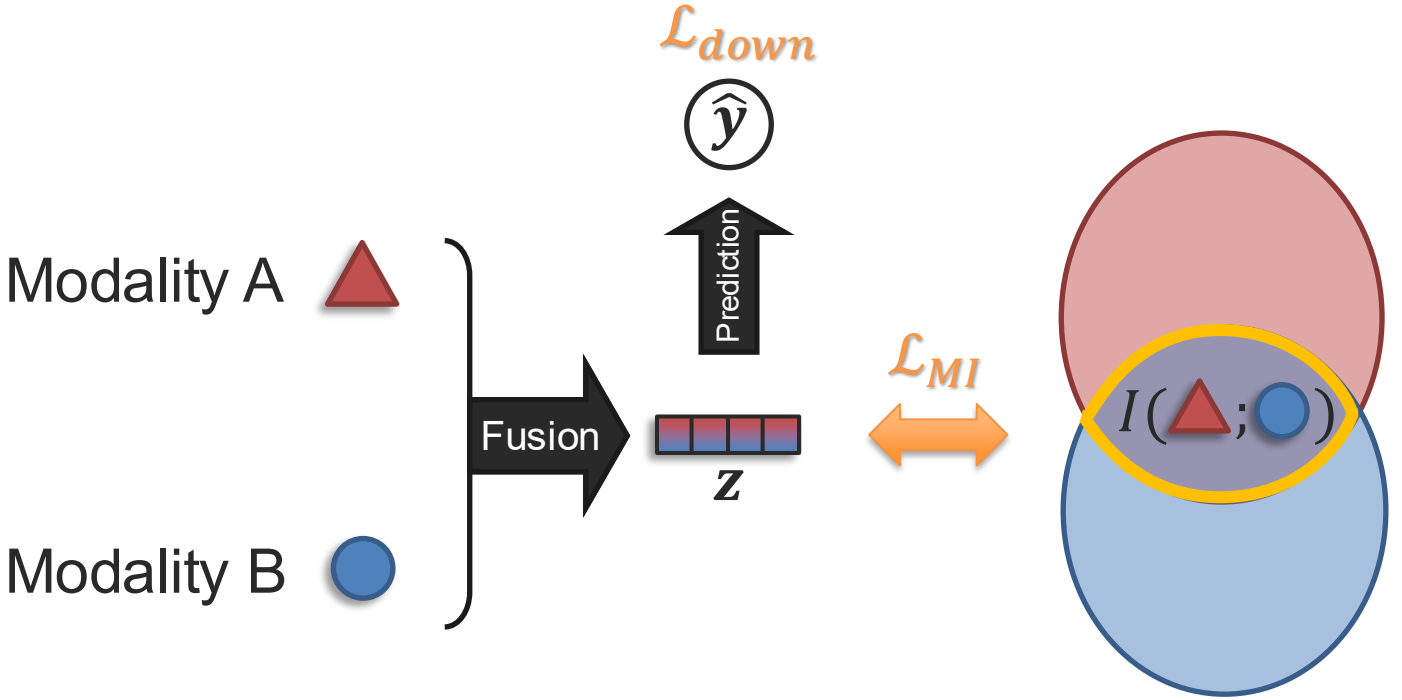
[Radford et al., Learning Transferable Visual Models From Natural Language Supervision. ICML 2021]

Multimodal Coordination – Information Theory



[Radford et al., Learning Transferable Visual Models From Natural Language Supervision. ICML 2021]

Multimodal Fusion with Mutual Information

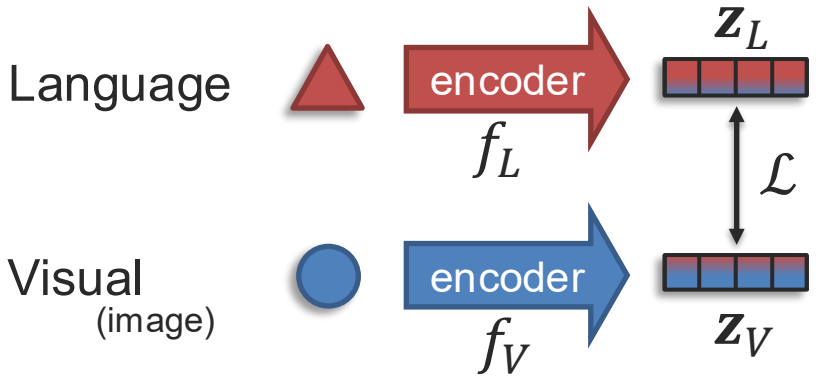


Assumption?

Information present in both modalities is most important for the downstream task

[Colombo et al., Improving Multimodal Fusion via Mutual Dependency Maximization. EMNLP 2021]

Contrastive Learning and Mutual Information



InfoNCE:

$$\mathcal{L} = -\mathbb{E}\left[\log \frac{f(\boldsymbol{x}_A^i, \boldsymbol{x}_B^i)}{\sum_{j=1}^{N} f(\boldsymbol{x}_A^i, \boldsymbol{x}_B^j)}\right]$$

critic function

Critic function f is trained to be a binary classifier distinguishing $\boldsymbol{x}_A, \boldsymbol{x}_B \sim p(\boldsymbol{x}_A, \boldsymbol{x}_B)$ vs $\boldsymbol{x}_A, \boldsymbol{x}_B \sim p(\boldsymbol{x}_A)p(\boldsymbol{x}_B)$

InfoNCE/CL:

- 'Captures' mutual information
- Optimizes a lower bound on mutual information

At optimal loss, $f^*(\boldsymbol{x}_A, \boldsymbol{x}_B) = \frac{p(\boldsymbol{x}_A, \boldsymbol{x}_B)}{p(\boldsymbol{x}_A)p(\boldsymbol{x}_B)}$.

Plugging f^* back into $\mathcal{L}$ gives:

$$\mathcal{L}^* \geq \mathbb{E}\left[\log \frac{p(\boldsymbol{x}_A)p(\boldsymbol{x}_B)}{p(\boldsymbol{x}_A, \boldsymbol{x}_B)} N\right] = -I(X_A, X_B) + \log N$$

In other words:

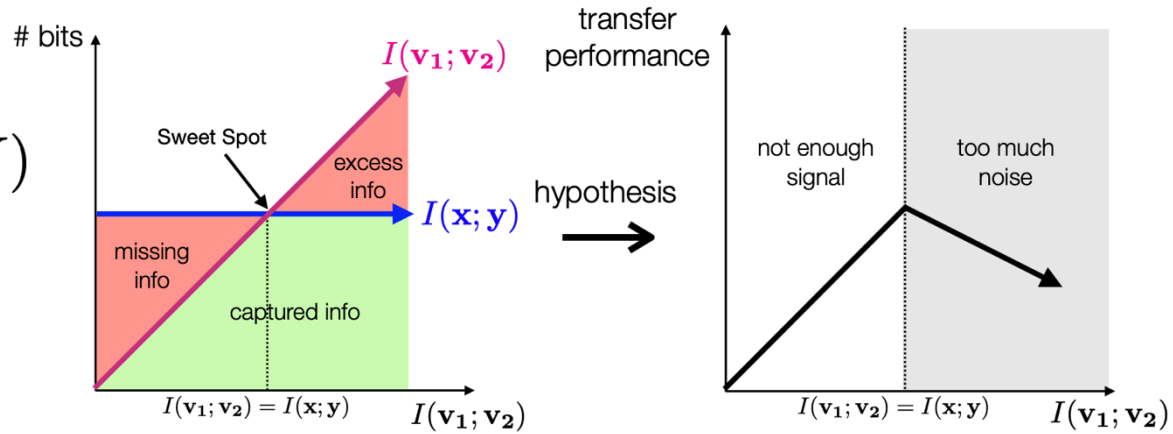
$$I(X_A, X_B) \geq \log N - \mathcal{L}^*$$

[Oord et al., Representation Learning with Contrastive Predictive Coding. 2018]

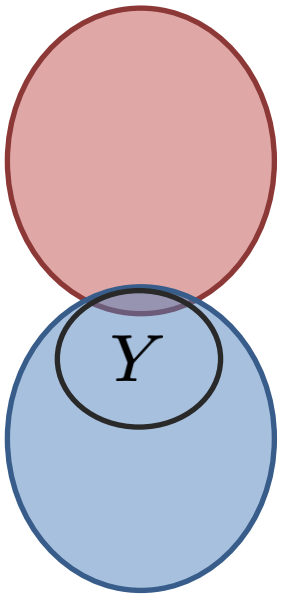
Multiview Redundancy and Contrastive Learning

How much information should be shared?

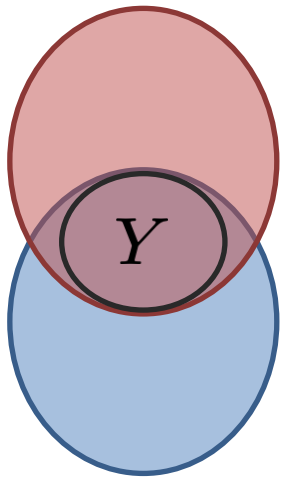
Multi-view redundancy: $I(X_1; X_2) = I(X_1; Y)$



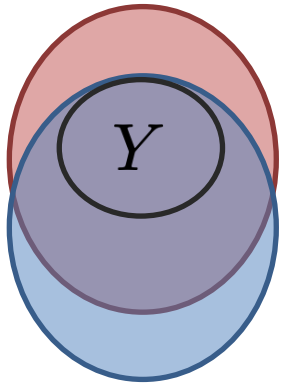
Not enough signal



Just right



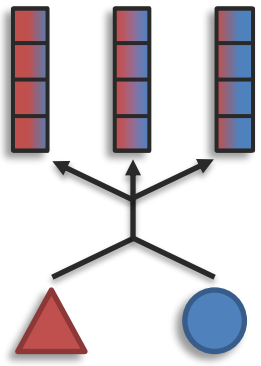
Too much noise



Multi-view redundancy may not hold for multimodal problems!

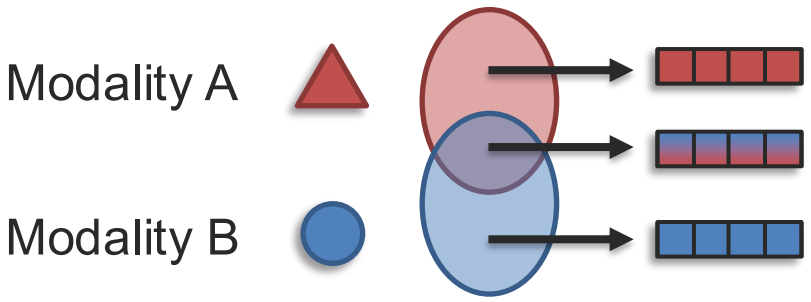
[Tian et al., What makes for Good Views for Contrastive Learning? NeurIPS 2020]
[Tosh et al., Contrastive Learning, Multi-view Redundancy, and Linear models. ALT 2021]

Representation Fission

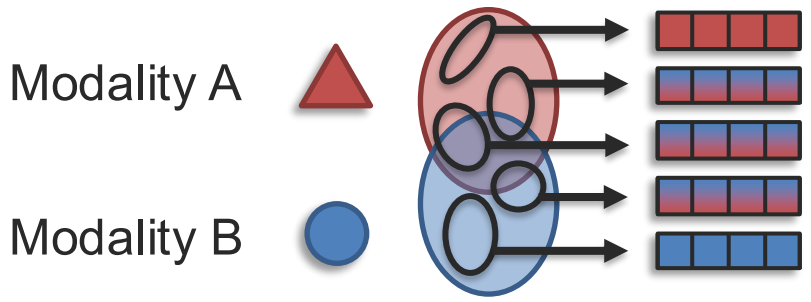


Definition: Learning a new set of representations that reflects multimodal internal structure such as data factorization or clustering

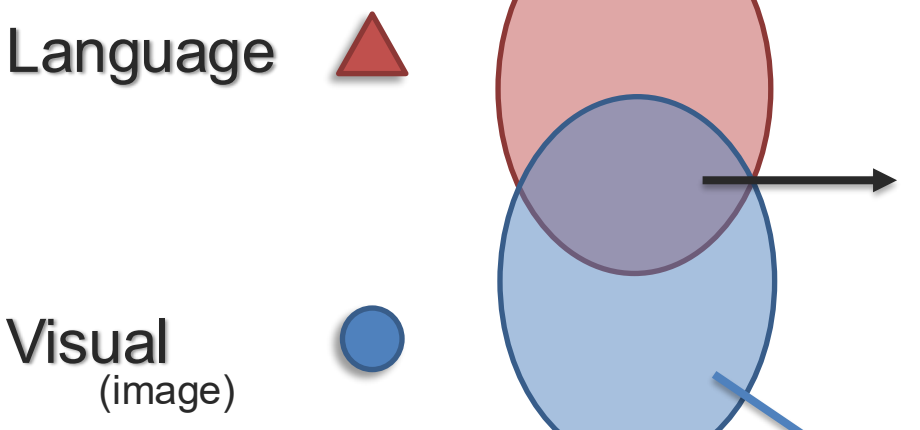
Modality-level fission:



Fine-grained fission:



Modality-level Fission



Information primarily in language modality

- Syntactic structure
- Vocabulary, morphology
- …

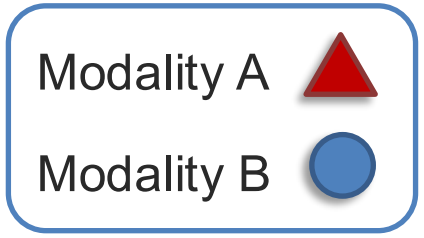
Information in both modalities

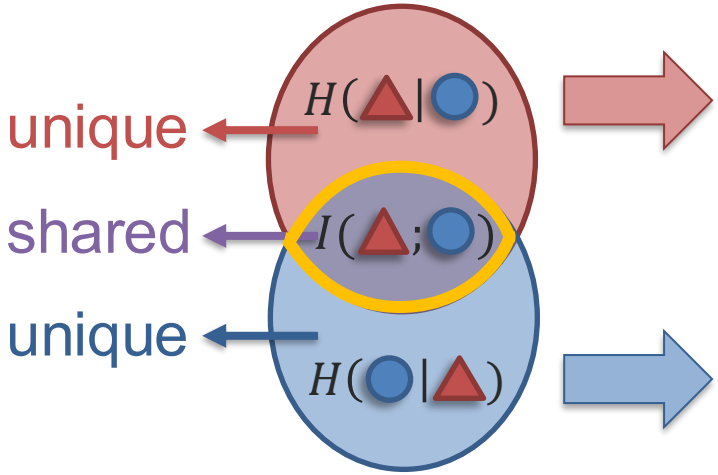
- Described people, objects, actions
- Illustrative gestures, motion
- …

Information primarily in visual modality

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

Factorized Learning of Shared + Unique Information





1 Maximize the mutual information

$I(\boldsymbol{z};\bullet)$ and $I(\boldsymbol{z};\blacktriangle)$

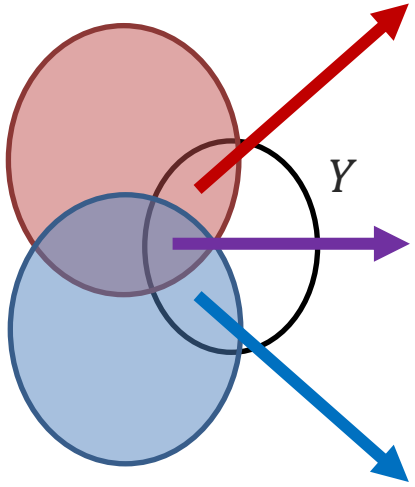
2 Minimize the conditional entropy

$H(\boldsymbol{z}|\bullet)$ and $H(\boldsymbol{z}|\blacktriangle)$

[Tsai et al., Learning Factorized Multimodal Representations. ICLR 2021]
[Wang et al., Rethinking Minimal Sufficient Representation in Contrastive Learning. CVPR 2022]

multisensory intelligence

Factorized Contrastive Learning



Can you please pass the cow?

Self CL

Cross-modal CL

Self CL

Can you please pass the cow?

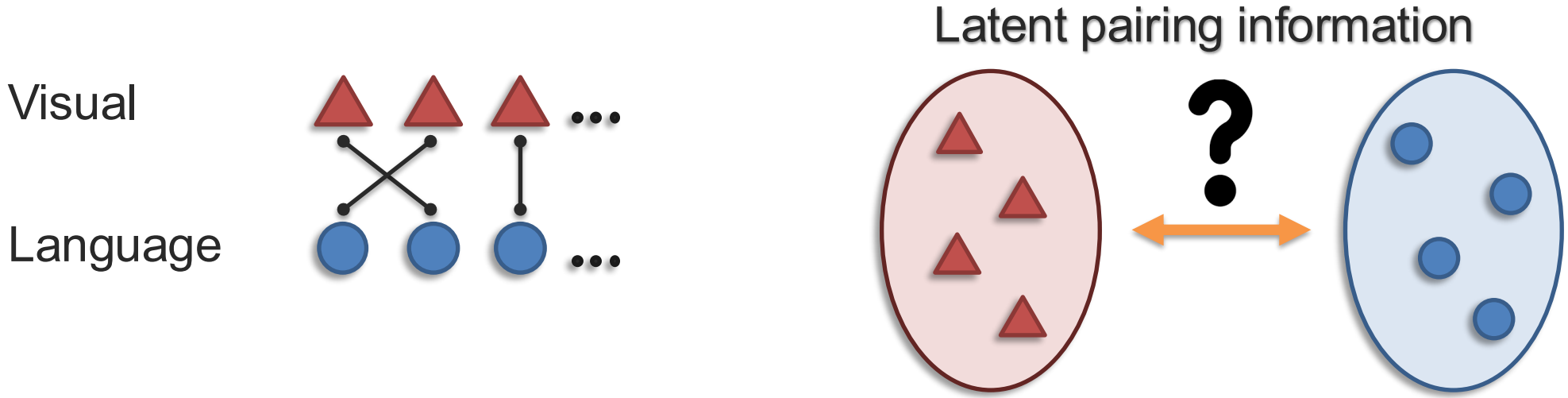
Can you kindly pass the cow?

Learns both shared and unique information.

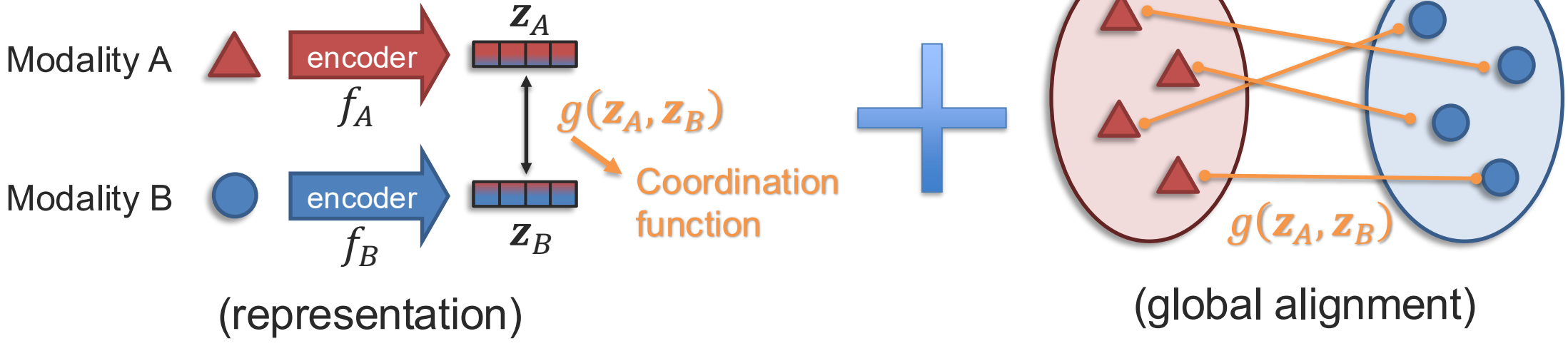
[Liang et al., Factorized Contrastive Learning: Going Beyond Multi-view Redundancy. NeurIPS 2023]

multisensory intelligence

Extensions: Global Alignment



Jointly optimize representation + global alignment:

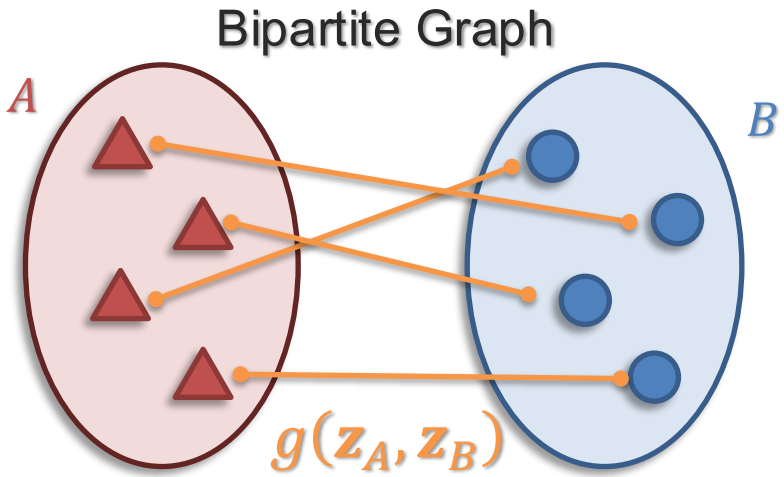


(representation)

(global alignment)

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

Assignment Problem



Initial assumptions:

- Same number of elements in A and B modalities
- 1-to-1 "hard" alignment between elements
- All elements assigned (aka "perfect matching")

➡ How to solve?

Naive solution: check all assignments

Better solution: Linear Programming

Assignment: (vector of indices)

~~$f: A \rightarrow B$~~

$x_{ij} = 1$ when matching connection, otherwise 0

Similarity weights:

~~$w_{(i,f(i))} = g(\boldsymbol{z}_A^i, \boldsymbol{z}_B^{f(i)})$~~

$w_{(i,j)} = g(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)$

Maximize:

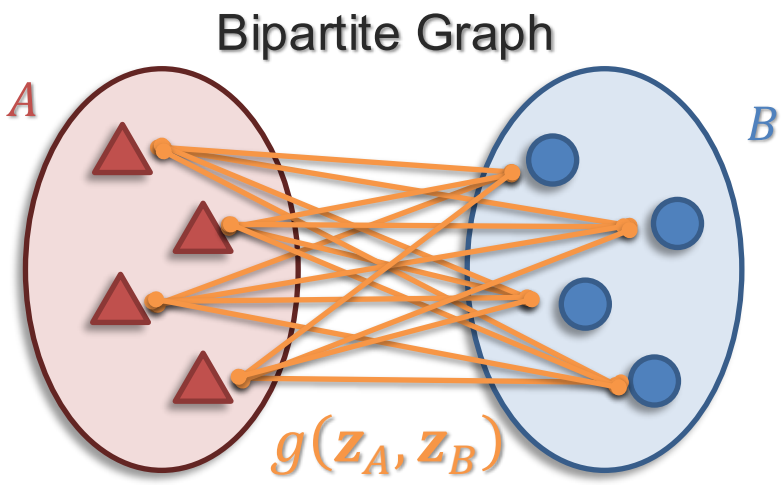
~~$\max_{f \in \mathrm{Perm}(N)} \sum_{i=1}^{N} w_{i,f(i)}$~~

$\max_{\{x_{ij}\}} \sum_{(i,j) \in A \times B} w_{i,j} x_{ij}$

➡ Can be solved with simplex algorithm

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

Optimal Transport



New assumptions:

- Different number of elements in A and B modalities
- Many-to-many "soft" alignment between elements

➡ It can be seen as "transporting" elements from modality A to modality B (and vice-versa)

Assignments: $x_{(i,j)}$: soft alignment between $\boldsymbol{z}_A^i$ and $\boldsymbol{z}_B^j$

Similarity weights: $w_{(i,j)} = g(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)$

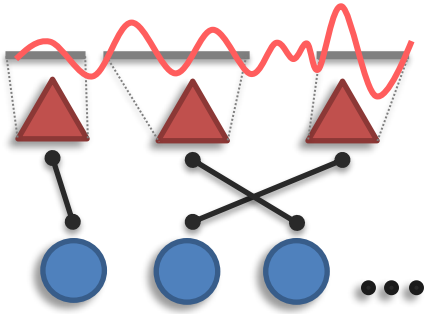
Maximize: $\max_{\{x_{ij}\}} \sum_{(i,j)\in A\times B} w_{i,j} x_{ij}$

➡ Wassertein distance give optimal transport

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

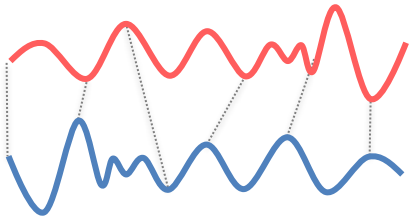
multisensory intelligence

Extensions: Continuous Alignment

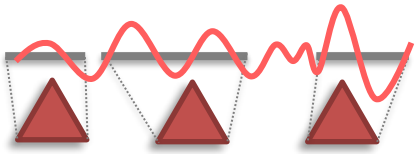


Definition: Model alignment between modalities with continuous signals and no explicit elements

Continuous warping



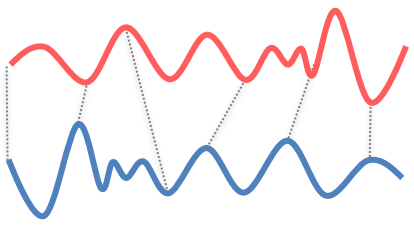
Discretization (segmentation)



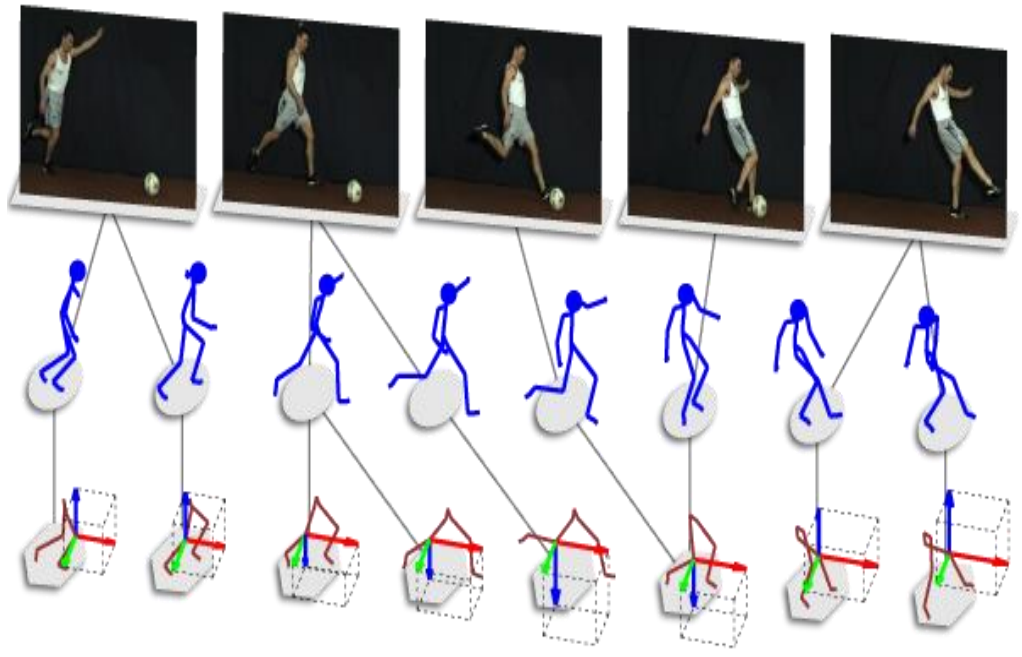
[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

Continuous Alignment

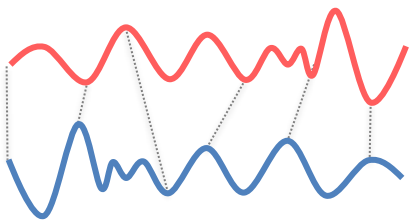


Aligning video sequences

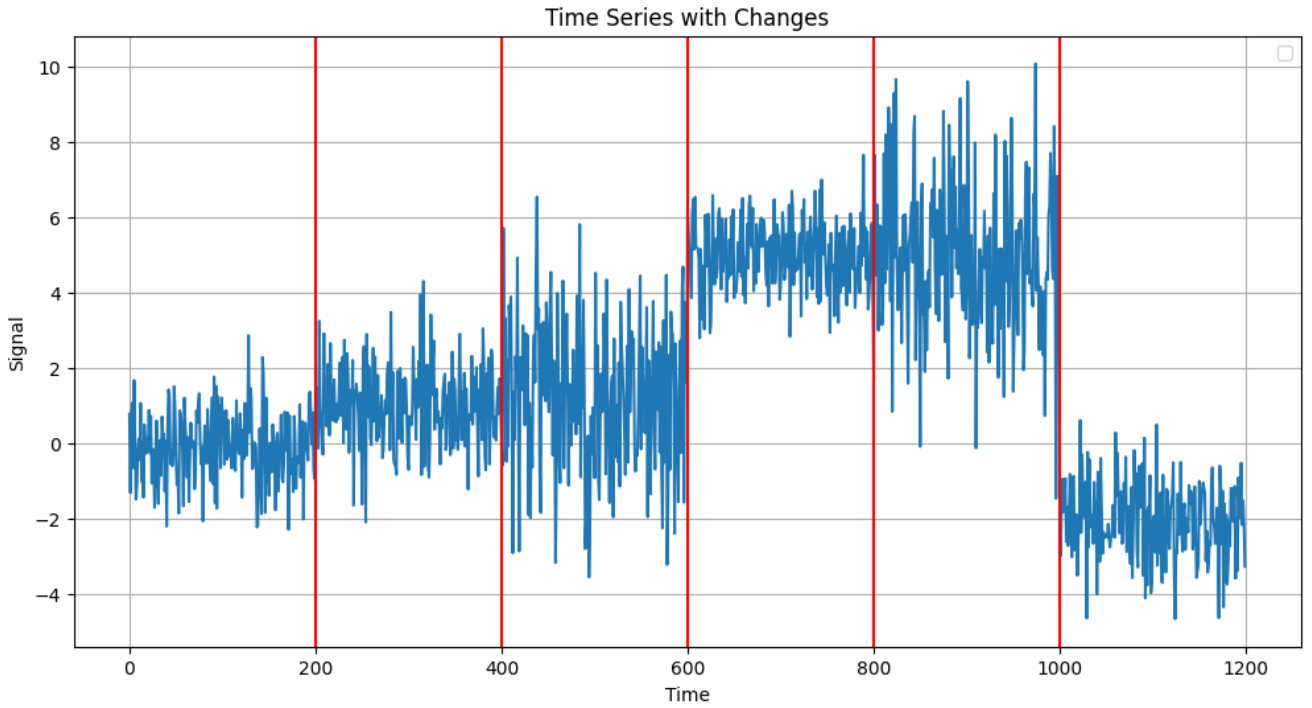


[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]
[Zhou and De la Torre. Canonical Time Warping for Alignment of Human Behavior. NeurIPS 2009]

Continuous Alignment

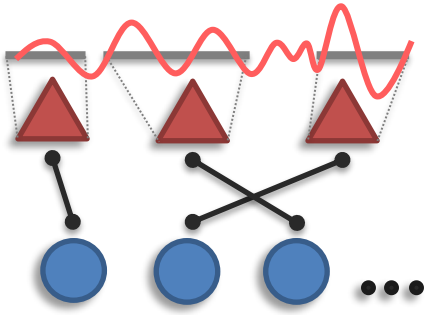


Changepoint detection



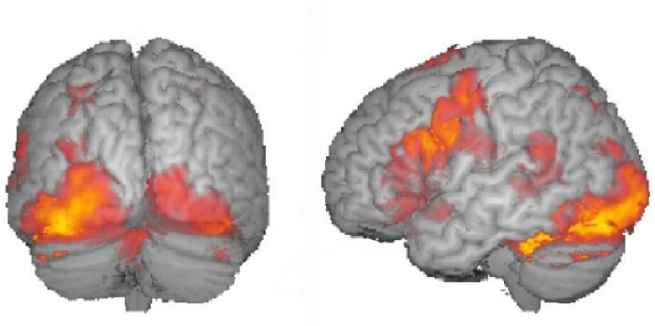
multisensory intelligence

Discretization (aka Segmentation)

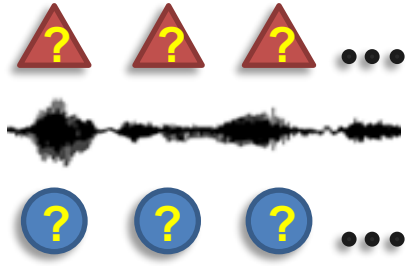


Common assumptions: ① Segmented elements

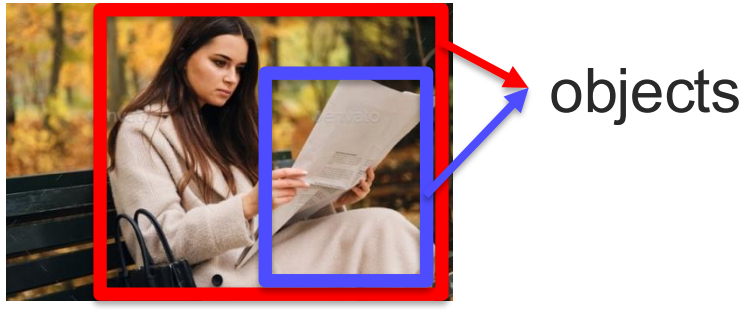
Examples:



Medical imaging



Signals

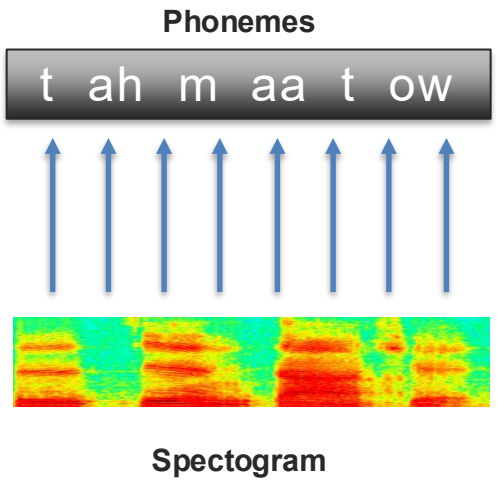


Images

multisensory intelligence

Discretization – Example

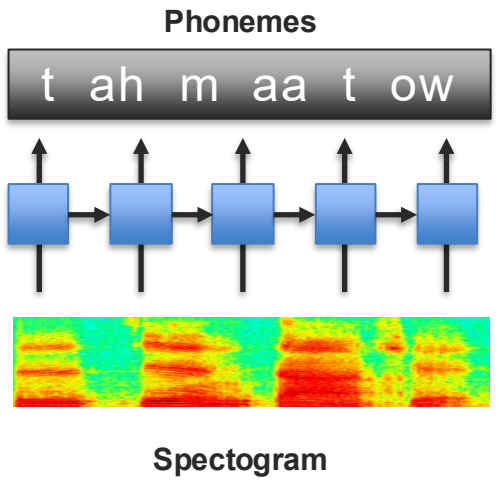
Sequence Labeling and Alignment



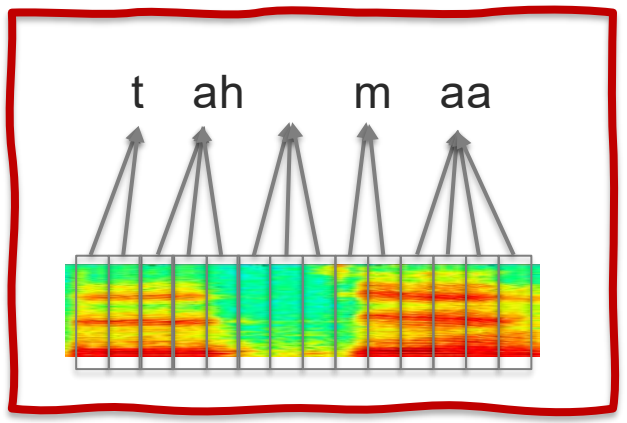
How can we predict the sequence of phoneme labels?

Discretization – Example

Sequence Labeling and Alignment



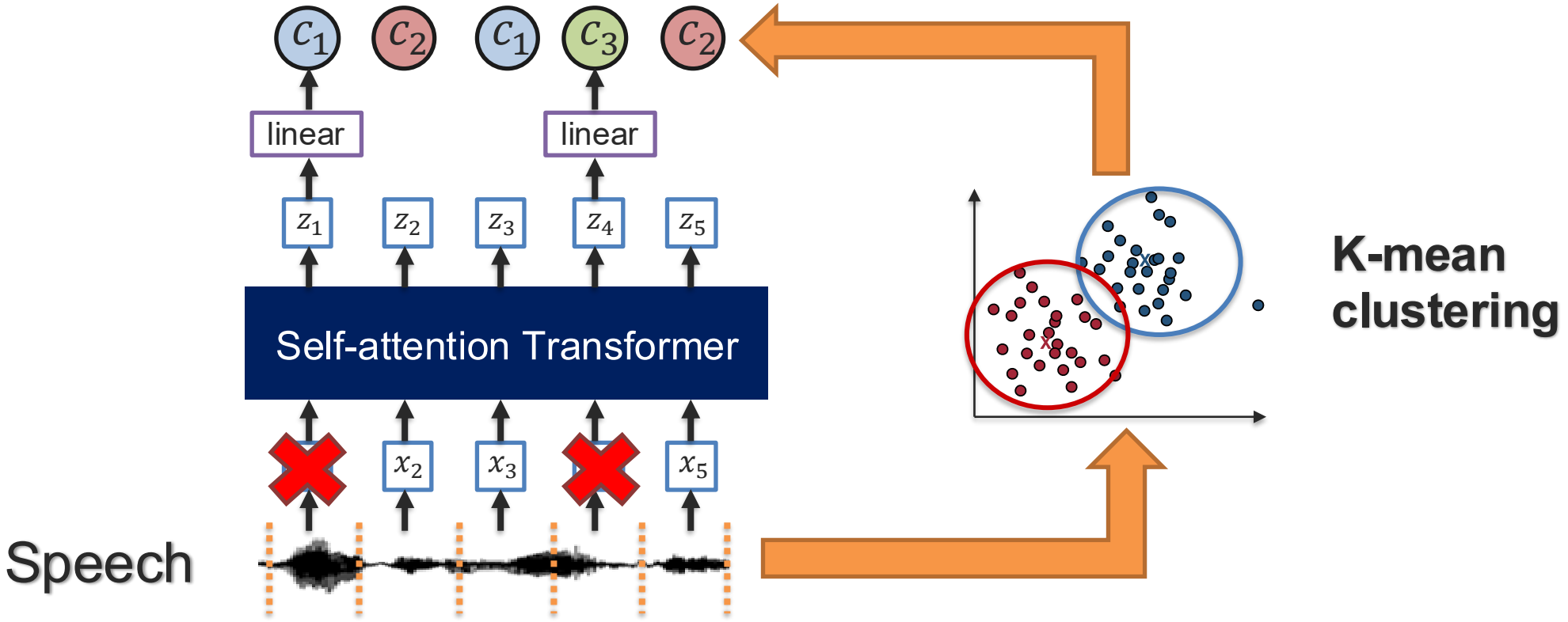
Challenge: many-to-1 alignment



How can we predict the sequence of phoneme labels?

Hidden-Unit Pre-training

HUBERT: Hidden-Unit BERT



[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

[Hsu et al., HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE TASLP 2021]

multisensory intelligence

VQ-VAE

Using a discrete variational autoencoder to learn discrete visual tokens

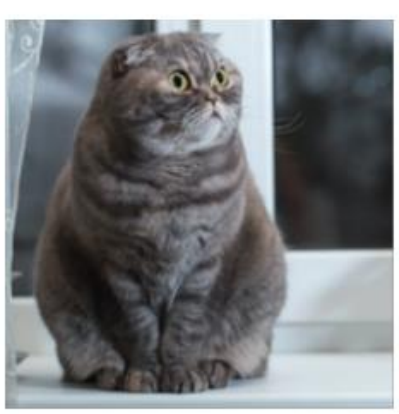


Image encoder

[98] [3] [990]

Image decoder

List of digits, [0… 8192]
Each digit is a “visual token”

Image encoder

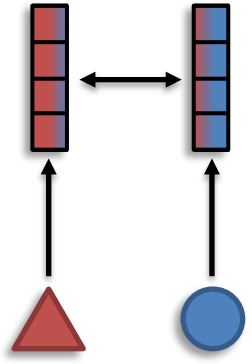
[46] [390] [6]

Image decoder

[van den Oord et al., Neural Discrete Representation Learning, NeurIPS 2017]

Extension: Implicit (Emergent) Alignment

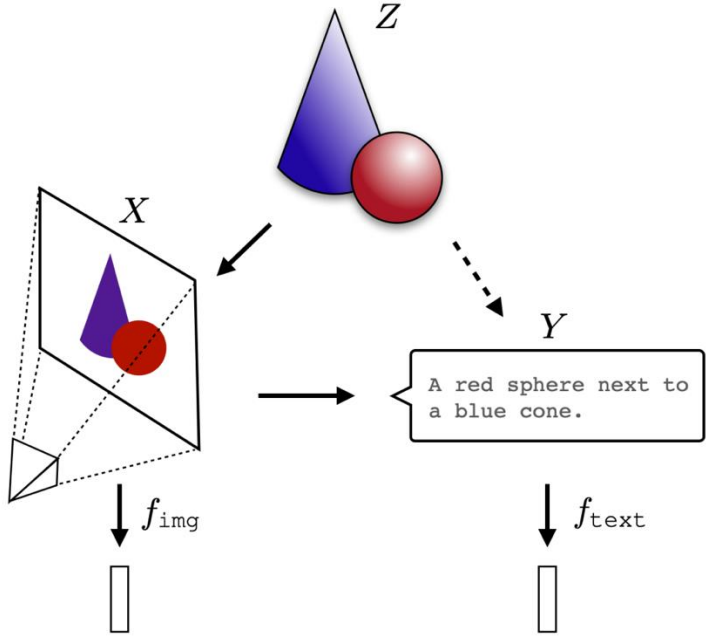
Explicit alignment



Implicit alignment

The Platonic Representation Hypothesis

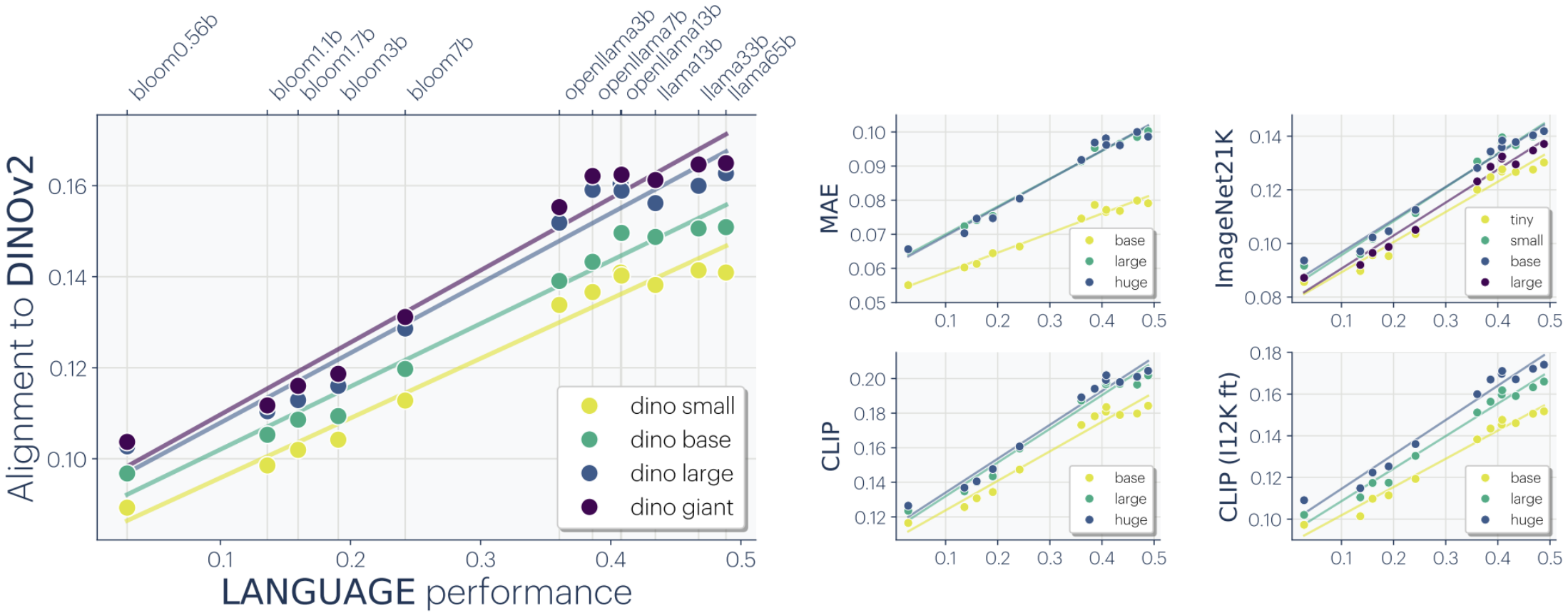
Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.



[Huh et al., The Platonic Representation Hypothesis. ICML 2024]

multisensory intelligence

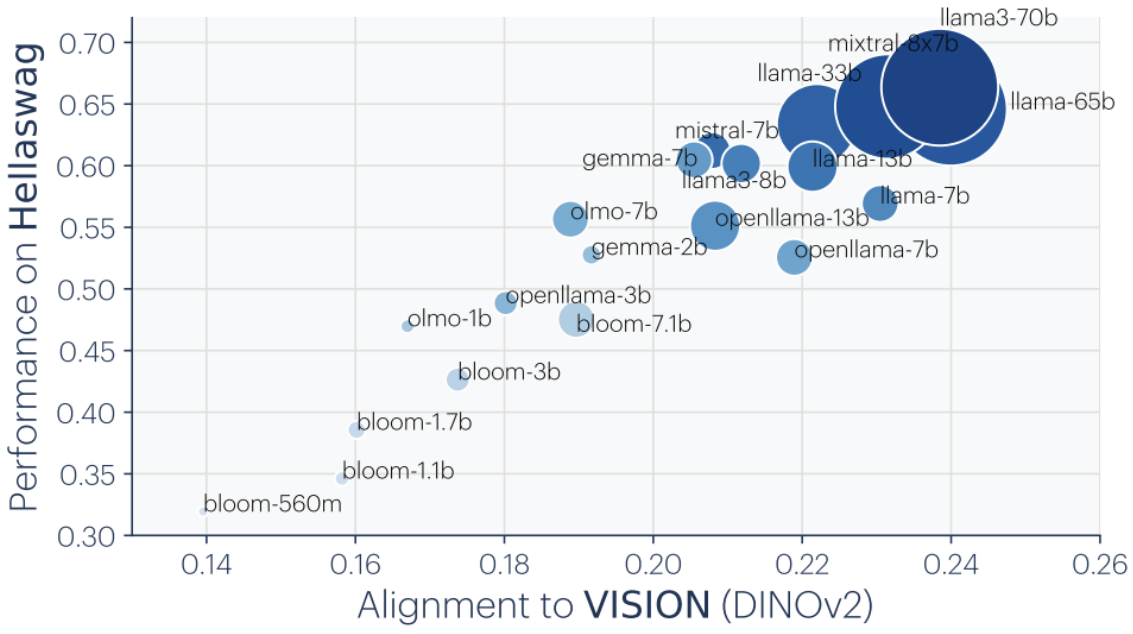
Emergence of Alignment

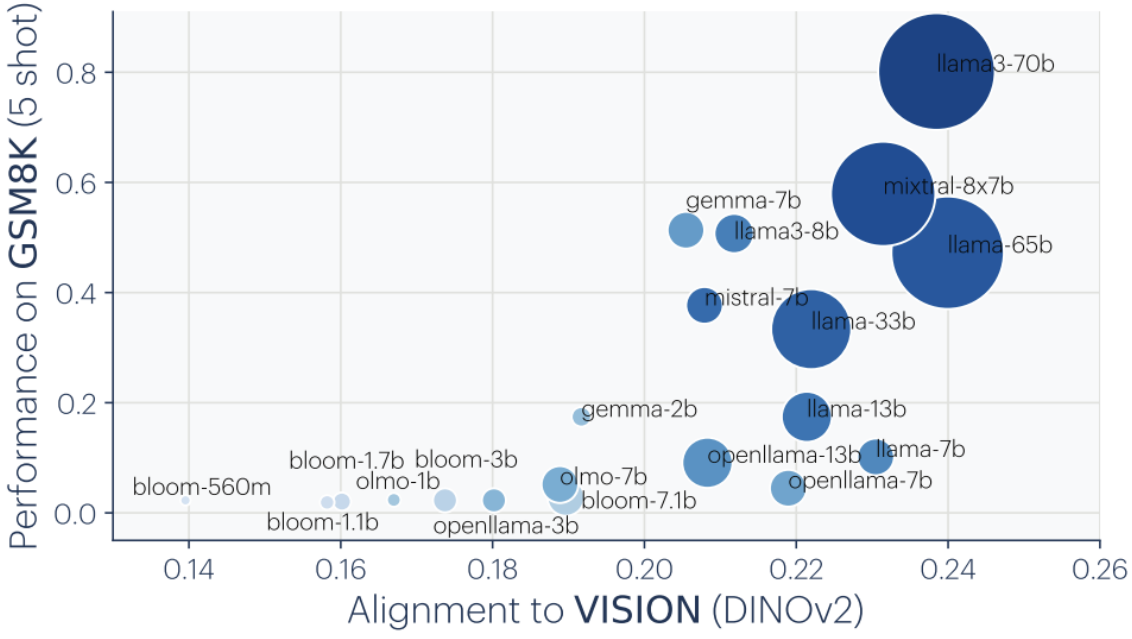


[Huh et al., The Platonic Representation Hypothesis. ICML 2024]

multisensory intelligence

Emergence of Alignment



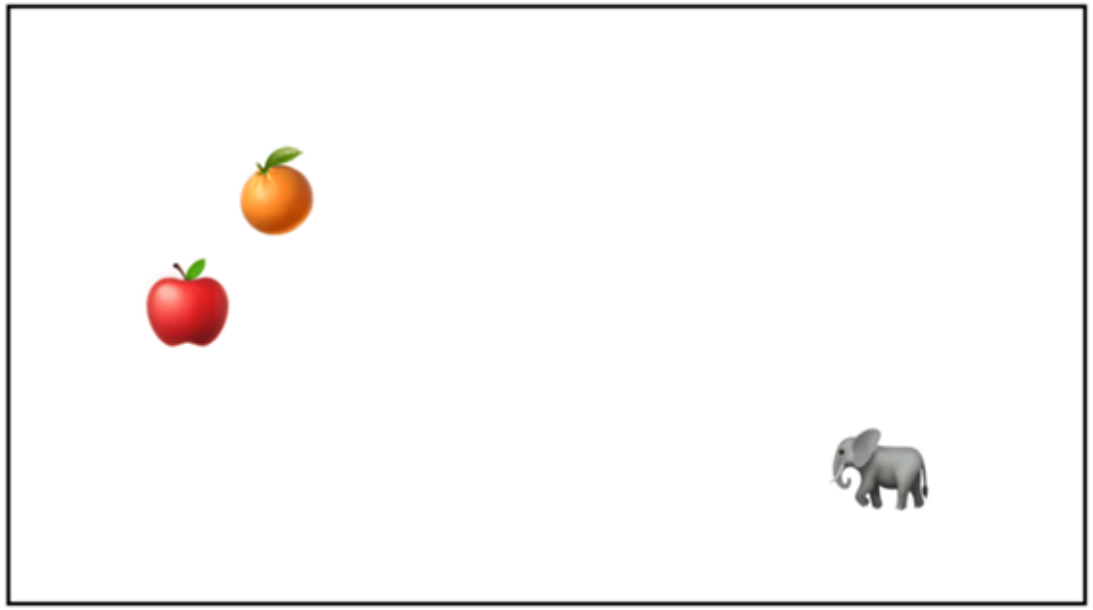


[Huh et al., The Platonic Representation Hypothesis. ICML 2024]

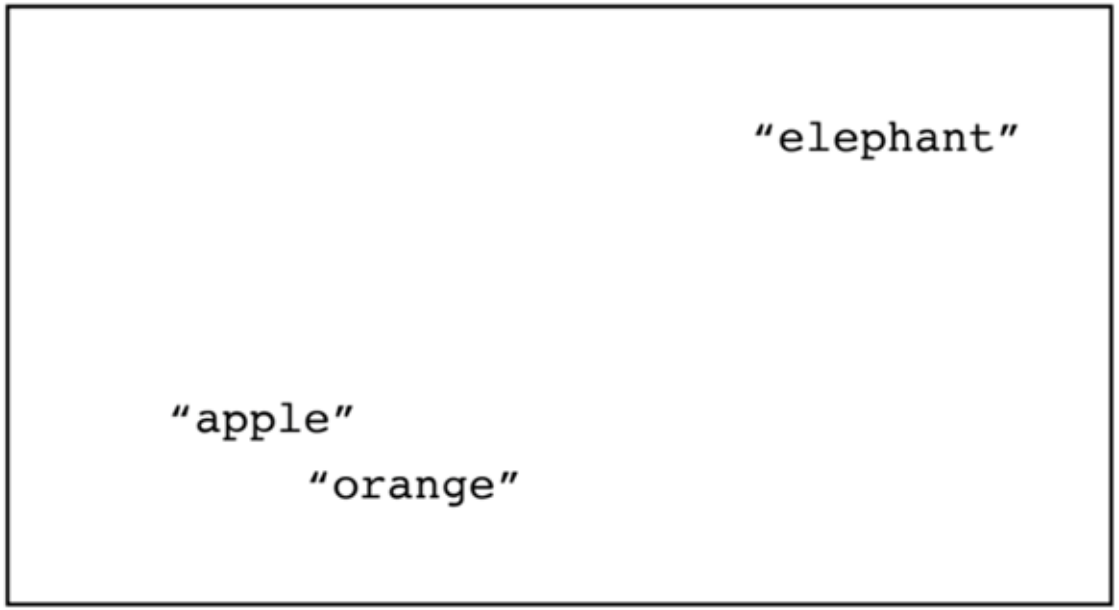
multisensory intelligence

Alignment via Kernel Similarity

Image embeddings



Text embeddings



$$K_{\text{img}}(i,j) = \langle E_{\text{img}}(x_i), E_{\text{img}}(x_j) \rangle$$

$$K_{\text{text}}(i,j) = \langle E_{\text{text}}(y_i), E_{\text{text}}(y_j) \rangle$$

$$K_{\text{img}}(i,j) \approx K_{\text{text}}(i,j) \quad \forall i,j$$

[Huh et al., The Platonic Representation Hypothesis. ICML 2024]

Limitations of Emergence?

Summary

1. Multimodal alignment
2. Explicit alignment and contrastive learning
3. Continuous alignment
4. Implicit (emergent) alignment

Assignments for This Coming Week

I want to meet every group at least once regarding their project ideas.

Compute credits: 40 x $50 Kimi credits, 40 x $40 other credits.

HW2 due next Wednesday 3/4.