Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 4.1: Multimodal alignment

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis. Major revision co-developed with Paul Liang. Co-lecturers included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Administrative Stuff

Primary TAs

- Each team has a primary TA
  - Your primary TA should have contacted you to schedule a first meeting together
  - Groups will be created in Piazza for each team
- Some projects may have a secondary TA, with complementary expertise

Schedule a meeting with your Primary TA this week!

First Project Assignment

Due date: Sunday 9/20 at 8pm

Four main sections:

- Introduction
- Related work
- Experimental setup
- Research ideas

Follows ICML paper format

The two main sections are related work and research ideas

teammates = # research ideas

Page limit depends on team size:

- 4 students : 5 pages + references
- 5 students : 5.5 pages + references
- 6 students : 6 pages + references

Proposal Presentations: Team Meetings with Instructor

- No lectures on Tuesday 9/22 and Thursday 9/24
- 15-mins meeting with instructor
  - 5-mins presentations (instruction details will be shared soon)
  - All teammates are required to attend
  - Meetings next week
- Signup form will be shared via Piazza

Attendance using Poll Everywhere

1. Scan QR Code
   or
   https://pe.app/mmml

2. Enter your AndrewID
   (as your “name tag”)

❌ *Do not try to sign-in*

3. Click ”Share location”

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 4.1: Multimodal alignment

Louis-Philippe Morency

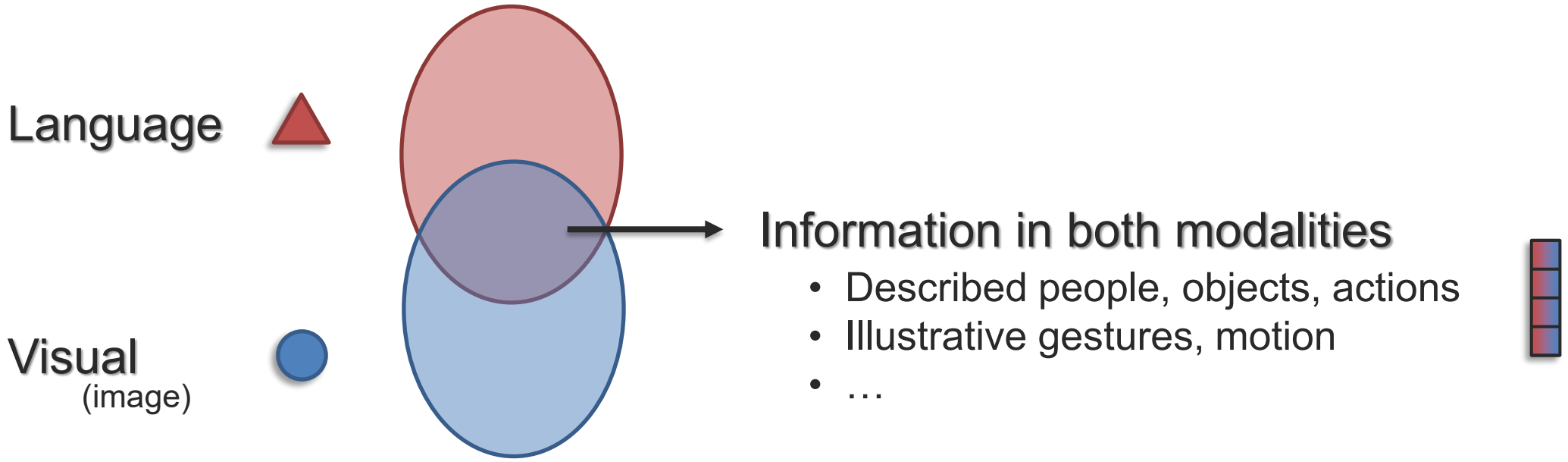
** Original course co-developed with Tadas Baltrusaitis. Major revision co-developed with Paul Liang. Co-lecturers included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Lecture objectives

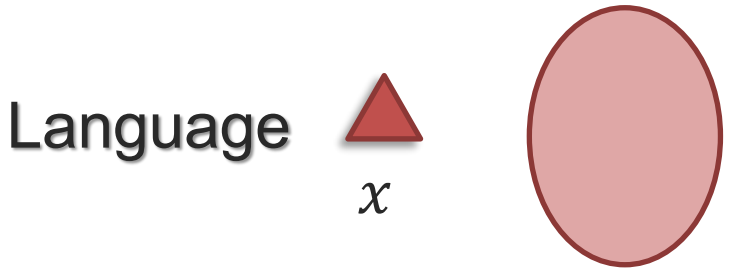
- Representation coordination
  - Information, entropy and mutual information
- Representation fission
  - Factorized multimodal representations
- Discrete alignment
  - Local alignment - Coordinated representations; soft attention
  - Global alignment - Assignment problem and optimal transport
- Continuous alignment
  - Continuous warping- Dynamic time warping
  - Discretization and segmentation

Representation Coordination

Multimodal Coordination – Information Theory



Information and Entropy – Information Theory



How much information in the modality?

Information Theory (Shannon, 1948)

Main intuition: "Information value" of a communicated message x depends on how random its content is

x: "1,1,1,1,1,1,1,1,1,1,1,1"

➡ Not very random… So, low information

x: "0,1,0,1,0,0,1,1,1,0,0,1"

➡ More random… So, higher information

Information content $I(x)$

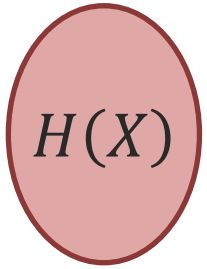
$$I(x) \sim \frac{1}{p(x)}$$

$$I(x) = \log\left(\frac{1}{p(x)}\right) = -\log(p(x))$$

Shannon, A Mathematical Theory of Communication, 1948

Information and Entropy – Information Theory

Language  x



How much information in the modality?

Information Theory (Shannon, 1948)

Information content $I(X) = -\log(p(X))$

- For discrete alphabet $\mathcal{X}$, then X is discrete random variable

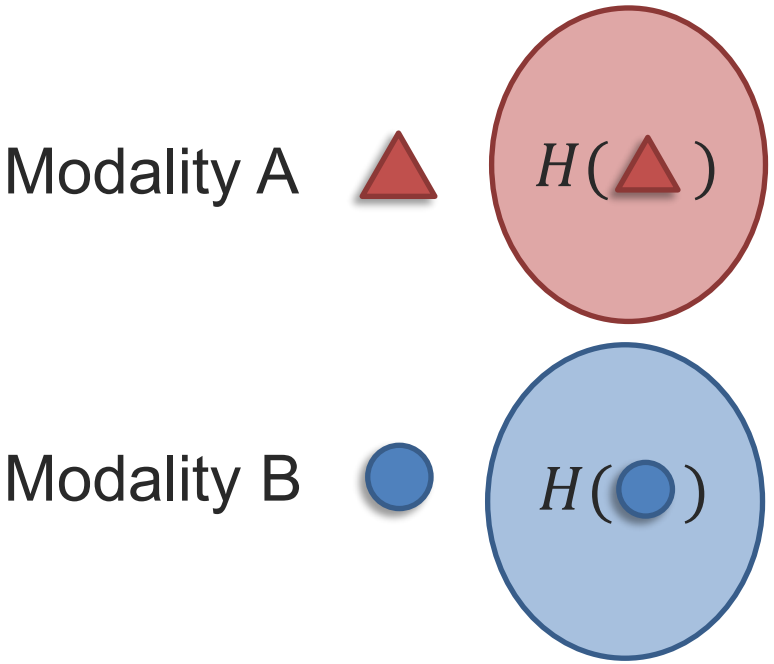
Entropy: weighted average of all possible outcomes from $\mathcal{X}$

$$H(X) = \mathbb{E}[I(X)] = \mathbb{E}[-\log(p(X))] = -\sum_{x \in \mathcal{X}} p(X)\log(p(X))$$

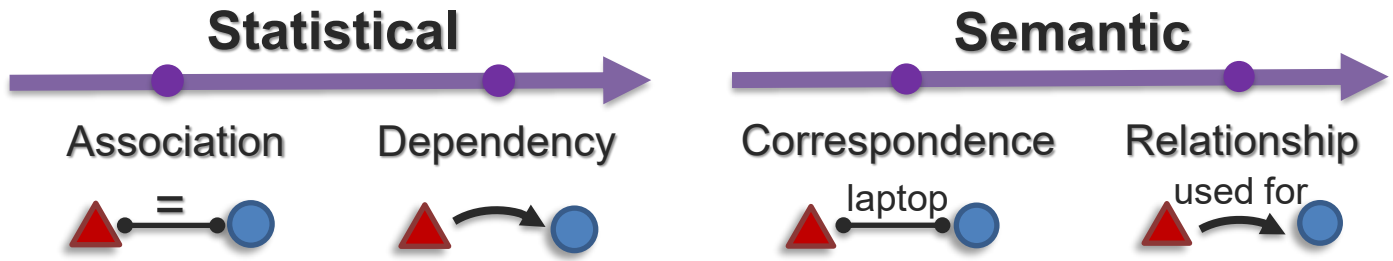
- Entropy can also be defined for continuous random variables

Entropy with Two Modalities

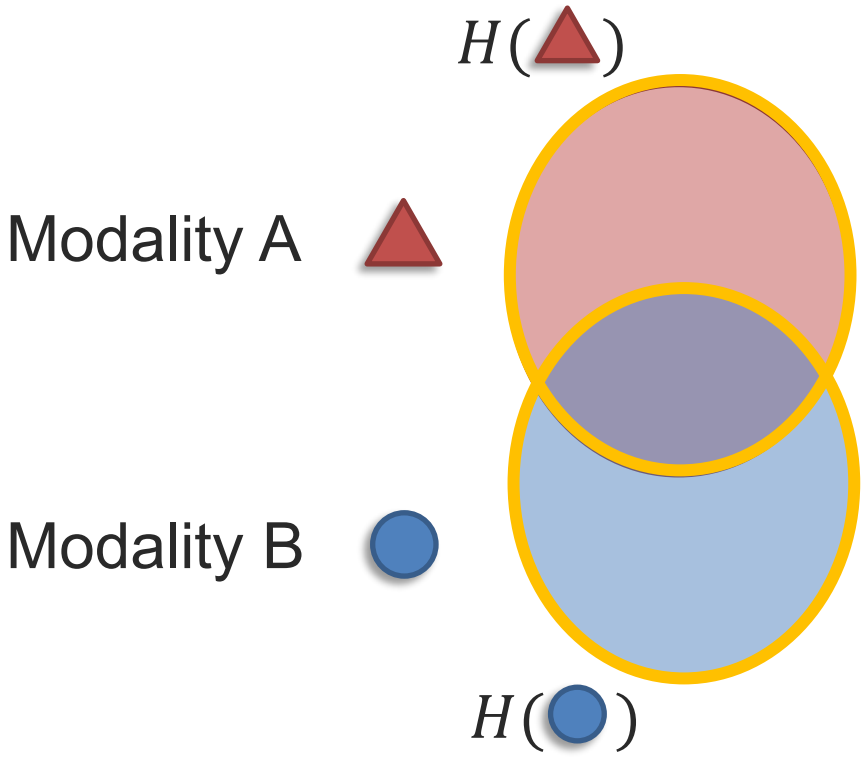
If no overlapping information



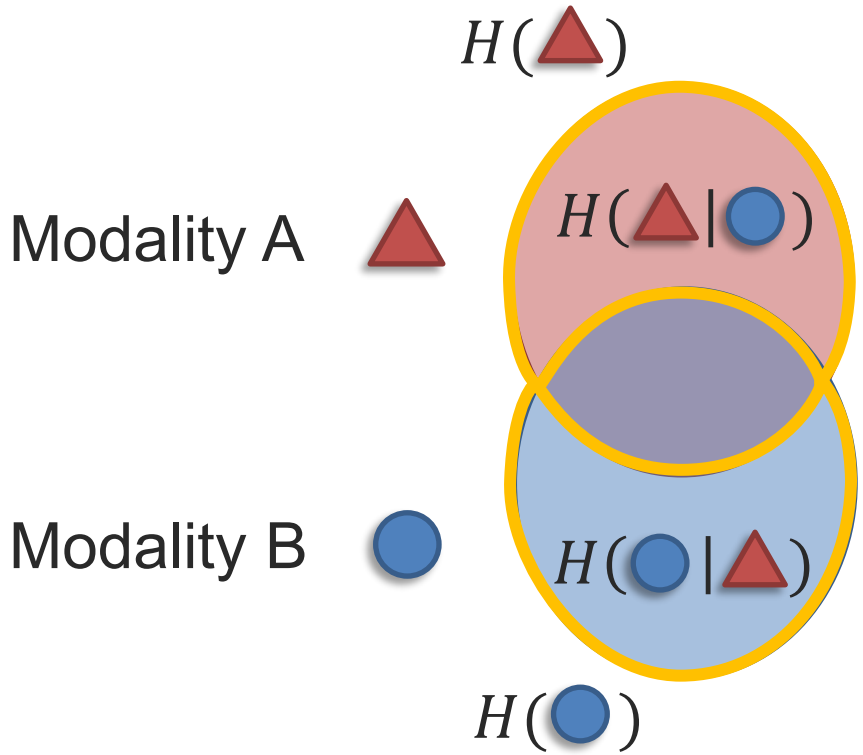
But in most real-world scenarios, modalities are *inter-connected*



Entropy with Two Modalities



Entropy with Two Modalities



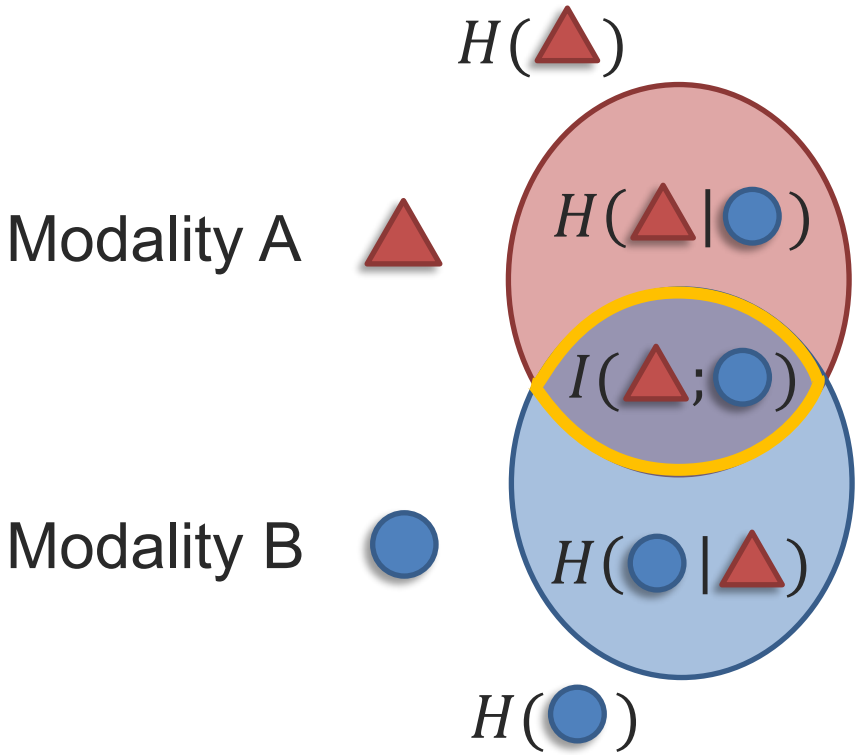
Conditional entropy $H(Y|X)$

$$H(Y|X) = -\mathbb{E}_{X,Y}[\log p(y|x)]$$

$$= -\mathbb{E}_{X,Y}\left[\log \frac{p(x,y)}{p(x)}\right]$$

If X and Y independent, $H(Y|X) = H(Y)$.
If X fully determines Y, then $H(Y|X) = 0$.

Entropy with Two Modalities



Mutual information $I(X;Y)$

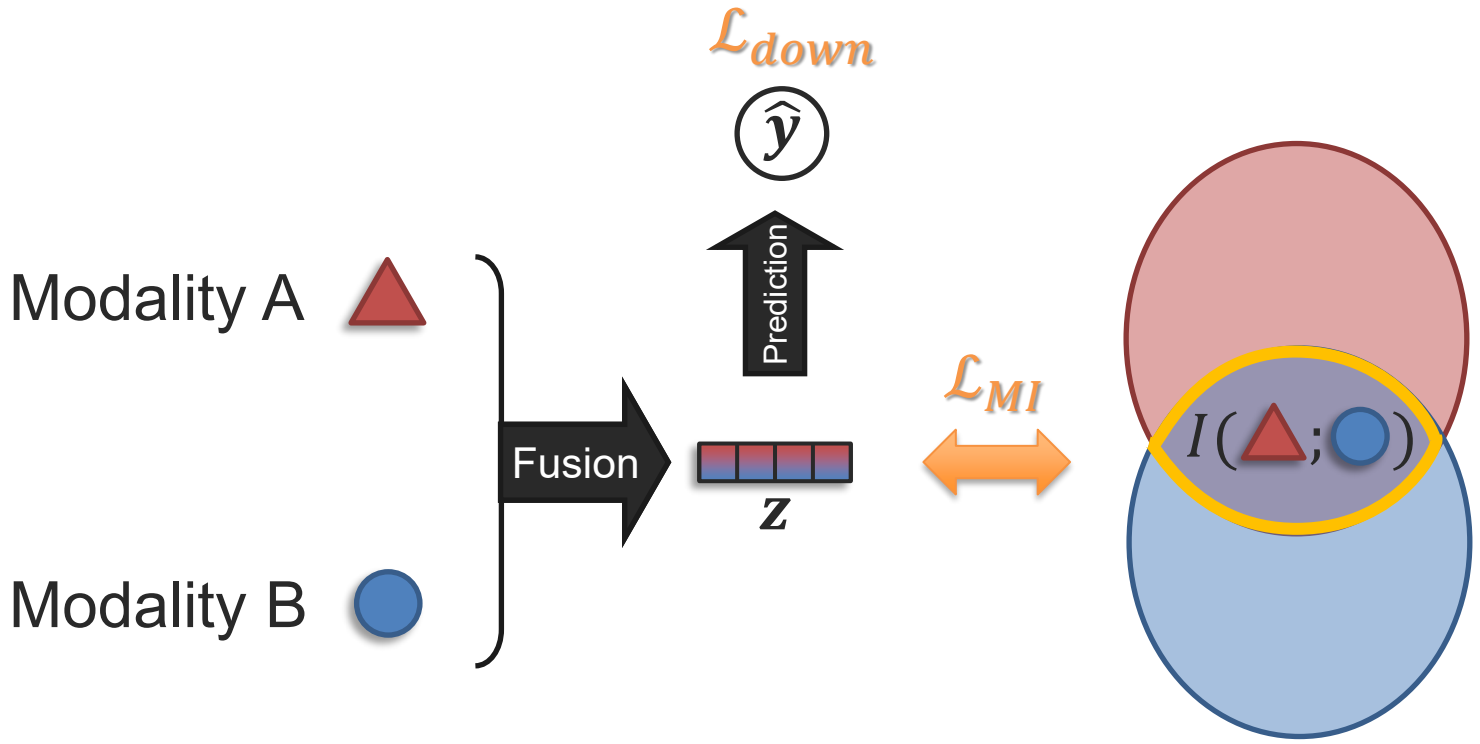
$$I(X;Y) = H(X) - H(X|Y)$$

$$= \mathbb{E}_{X,Y}\left[\log\frac{1}{P_X(x)} + \log\frac{P_{XY}(x,y)}{P_Y(y)}\right]$$

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

using KL-divergence ⟸ $I(X;Y) = D_{KL}(P_{XY}(x,y) \parallel P_X(x)P_Y(y))$

Multimodal Fusion with Mutual Information

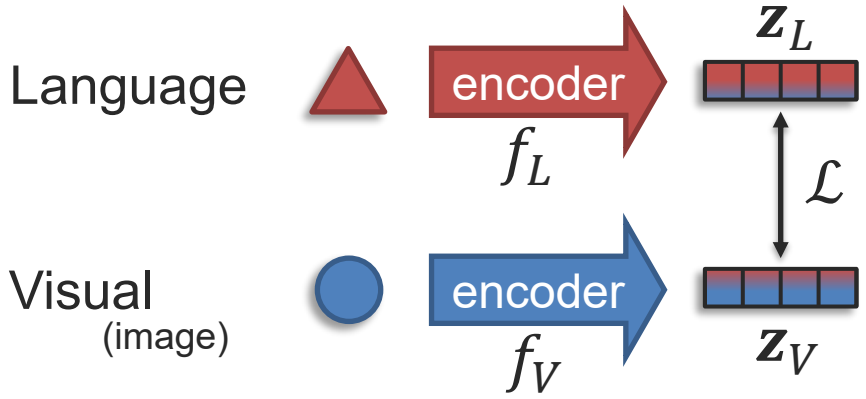


Assumption?

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

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

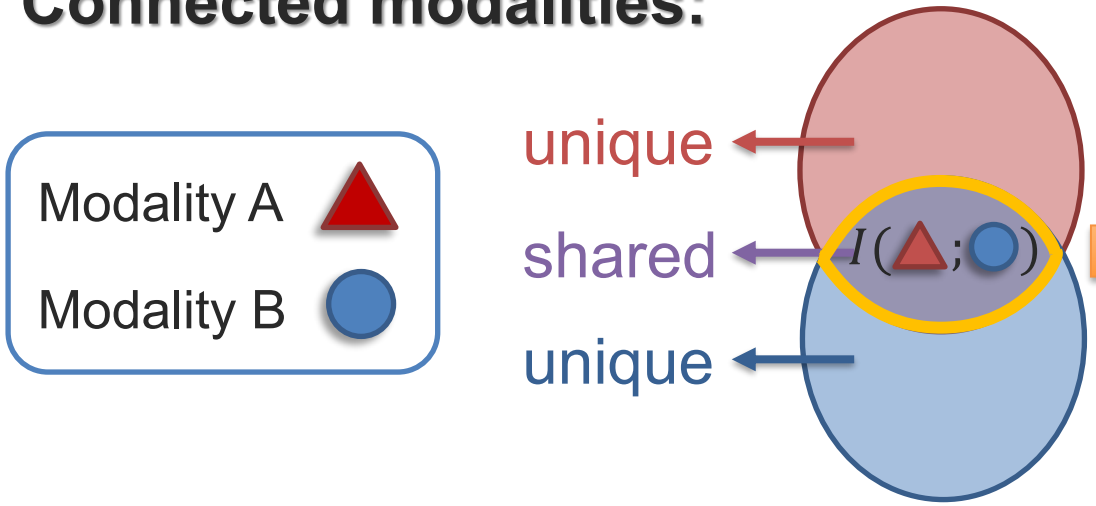
Contrastive Learning and Connected Modalities



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

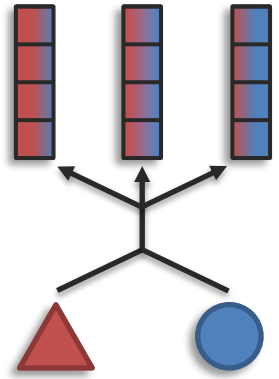
Mutual information $I(X;Y)$

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

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

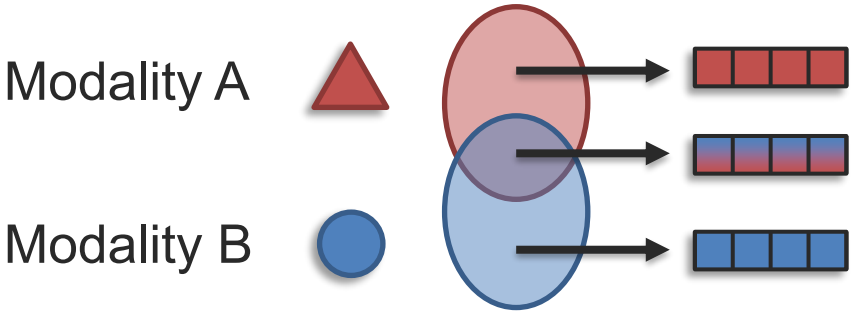
Representation Fission

Sub-Challenge 1c: Representation Fission

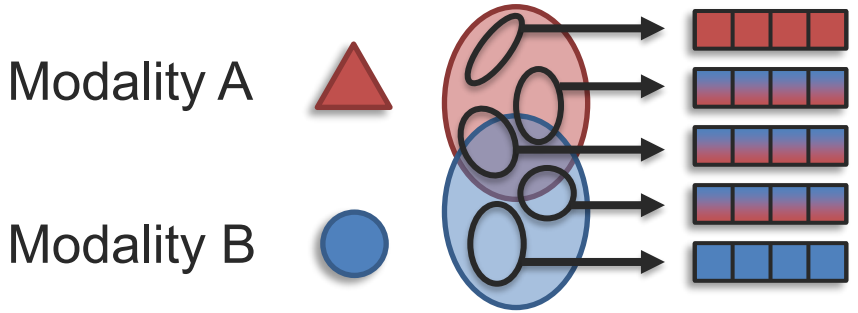


Definition: Learning a new set of representations that reflects multimodal internal structure such as data factorization or clustering

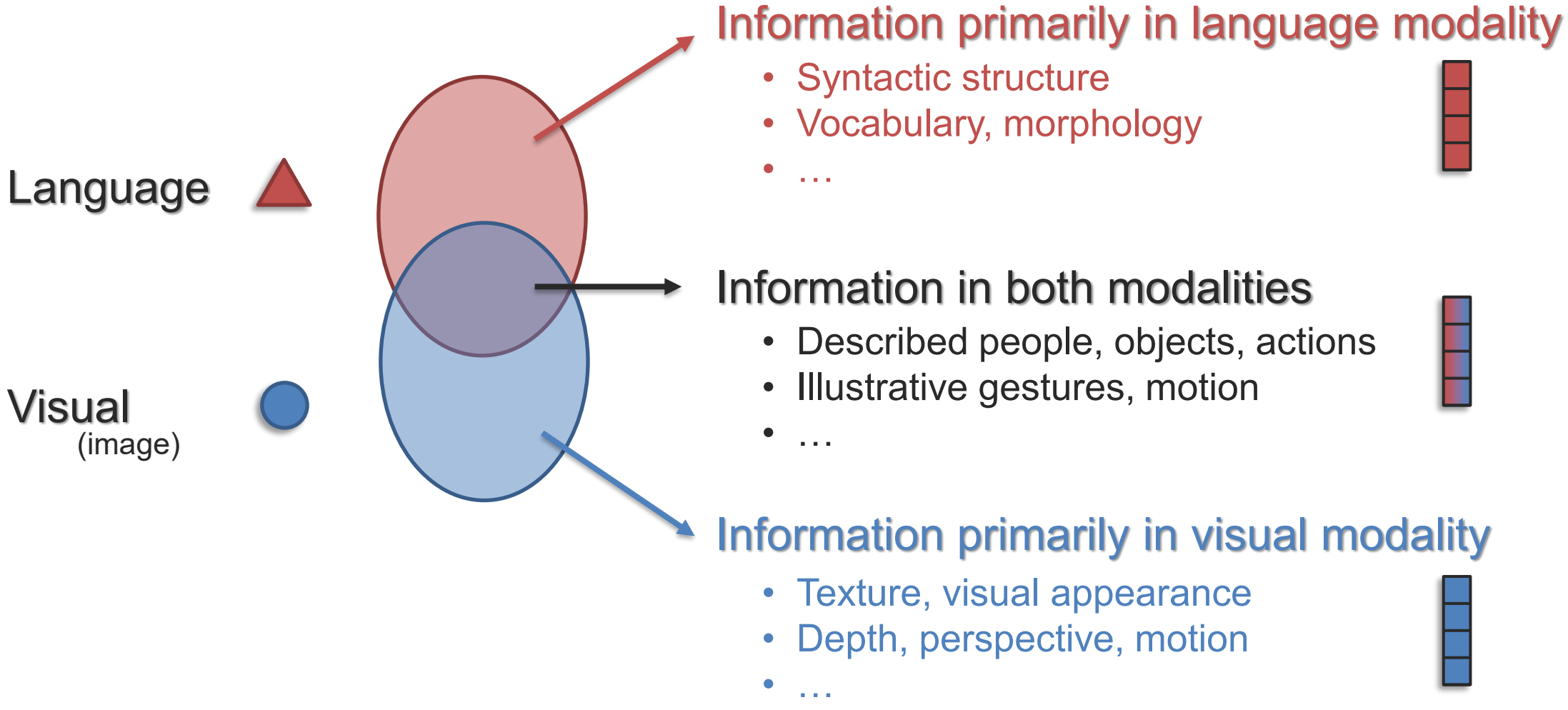
Modality-level fission:



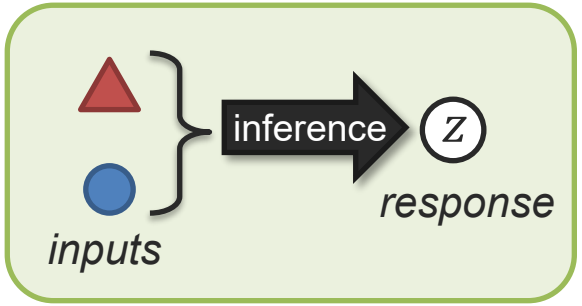
Fine-grained fission:



Modality-Level Fission

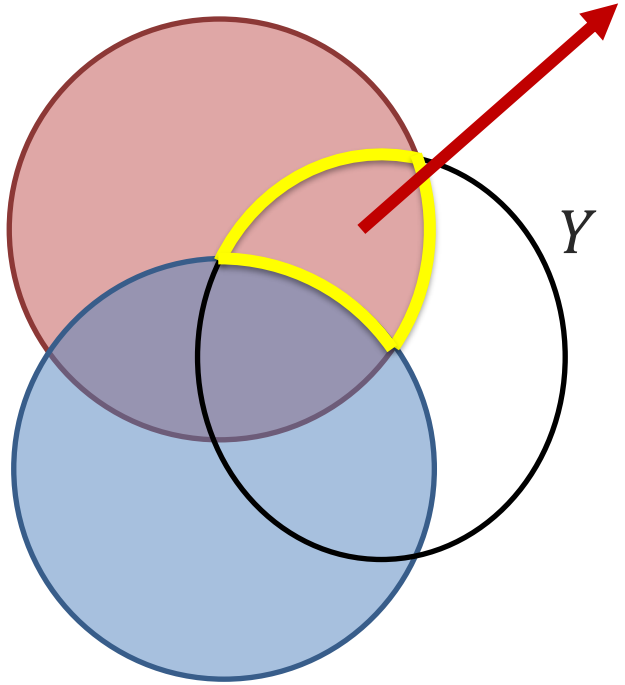


Modeling Multimodal Interactions



Modality A

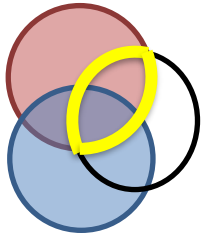
Modality B



Task-relevant **unique** information

$I(\triangle; Y | \bigcirc)$

Mutual information…



$I(\triangle; Y)$

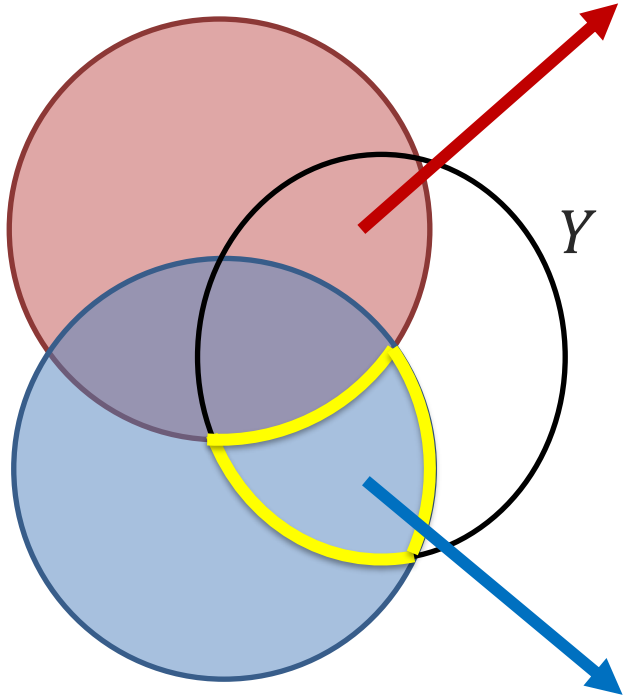
…but without ●

Information theory as a framework for modeling multimodal interactions between input modalities and responses

Modeling Multimodal Interactions

Modality A

Modality B



Task-relevant **unique** information

$I(\blacktriangle; Y | \bullet)$

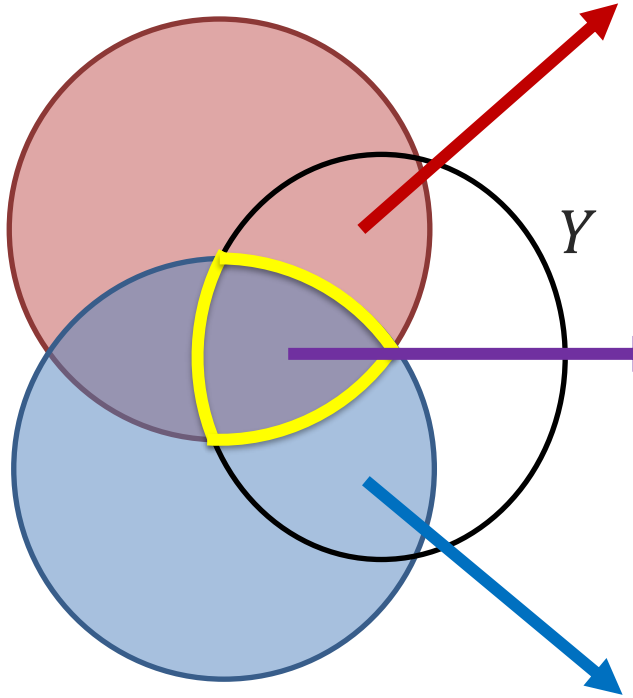
Task-relevant **unique** information

$I(\bullet; Y | \blacktriangle)$

Modeling Multimodal Interactions

Modality A

Modality B



Task-relevant **unique** information

$I(\blacktriangle; Y | \bullet)$

Task-relevant **shared** information

$I(\blacktriangle; \bullet; Y)$

…but how to compute this 3-way mutual information?

Task-relevant **unique** information

$I(\bullet; Y | \blacktriangle)$

Modeling Multimodal Interactions

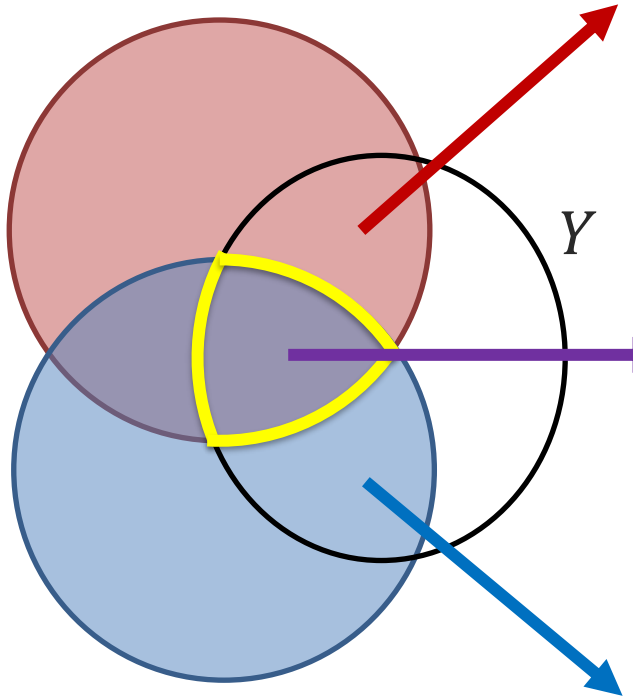
Partial Information Decomposition (PID)

$I(\triangle;\bigcirc;Y) = R - S$

factorizes 3-way mutual information into:

R: redundancy

S: Synergy



Task-relevant **unique** information

$I(\triangle;Y|\bigcirc)$

Task-relevant **shared** information

$I(\triangle;\bigcirc;Y)$

…but how to compute 3-way mutual information?

Task-relevant **unique** information

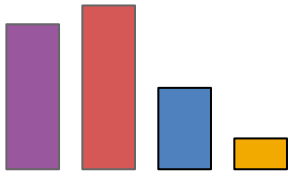
$I(\bigcirc;Y|\triangle)$

More details:
Liang et al., Quantifying & Modeling Feature Interactions: An Information Decomposition Framework. Neurips 2023

Quantifying Multimodal Interactions

These interactions can be efficiently estimated – gives a path towards understanding interactions

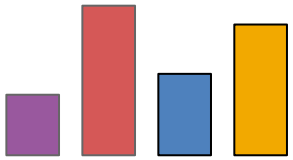
Sentiment



$R\ U_\ell U_{av}\ S$

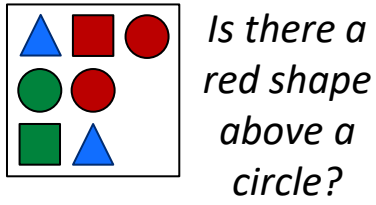
Language/Agreement

Sarcasm

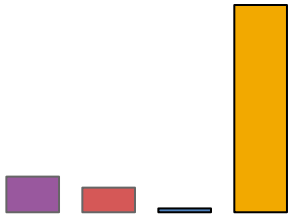


$R\ U_\ell U_{av}\ S$

Multimodal Transformer



VQA



$R\ U_\ell\ U_i\ S$

Multiplicative/Transformer

Also matches human judgment of interactions, and other sanity checks on synthetic datasets
They can also be used to choose most appropriate models.

[Liang et al., Quantifying & Modeling Feature Interactions: An Information Decomposition Framework. ICMI 2023]

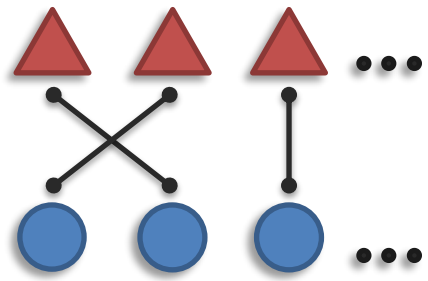
Challenge 2: Alignment

Challenge 2: Alignment

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

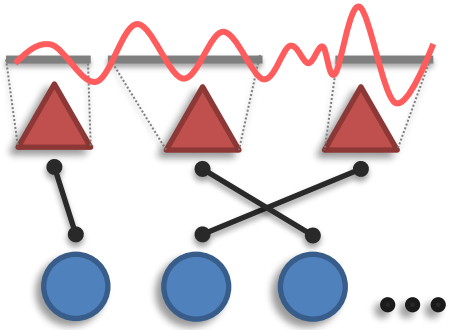
Sub-challenges:

Discrete Alignment



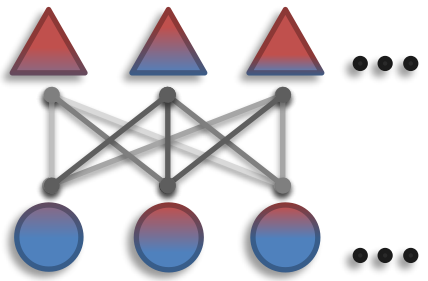
Discrete elements and connections

Continuous Alignment



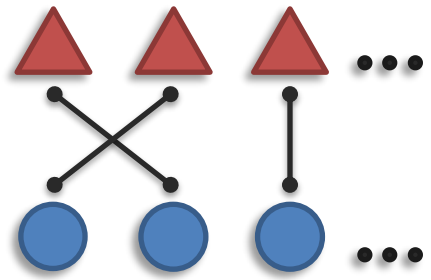
Segmentation and continuous warping

Contextualized Representation

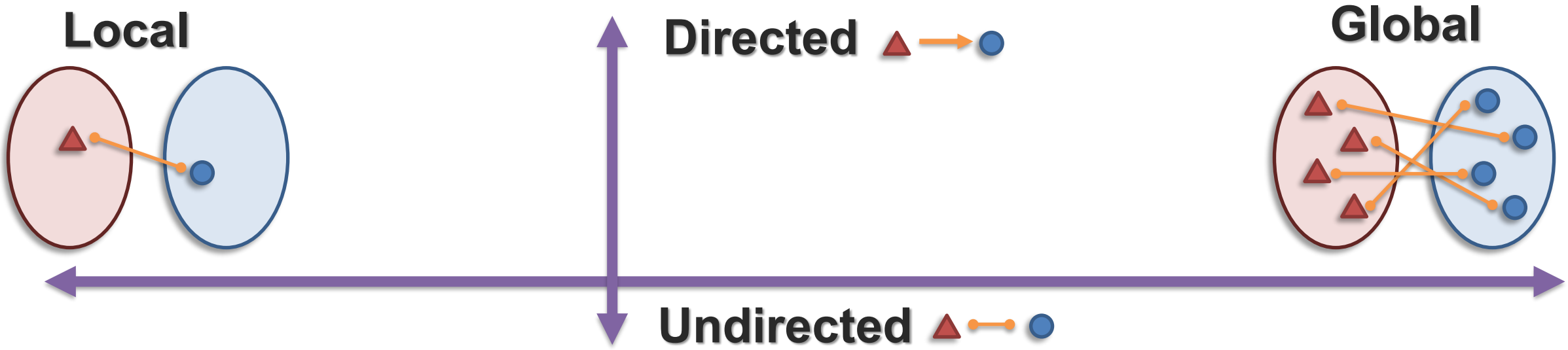


Alignment + representation

Sub-Challenge 2a: Discrete Alignment



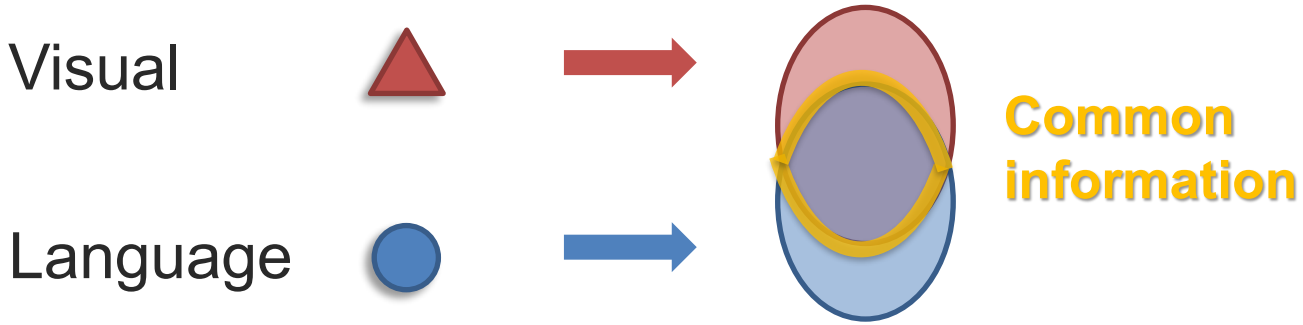
Definition: Identify and model discrete connections between elements of multiple modalities



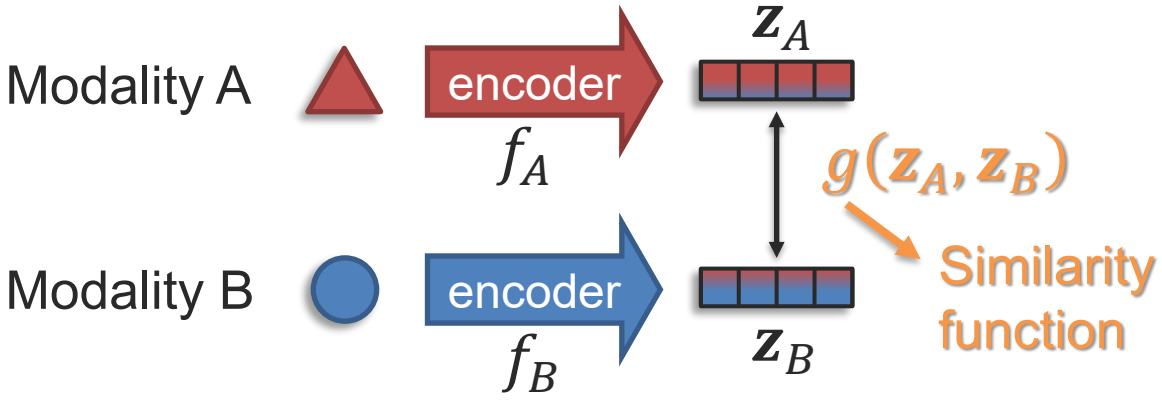
Language Grounding

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

Local Alignment – Coordinated Representations

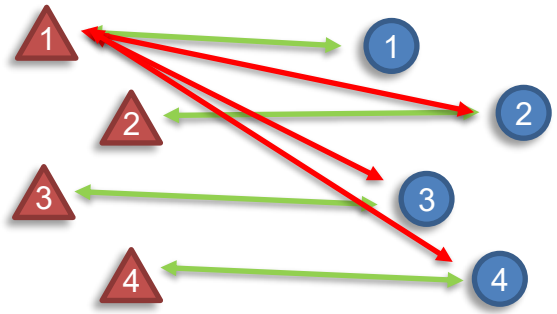


Learning coordinated representations:

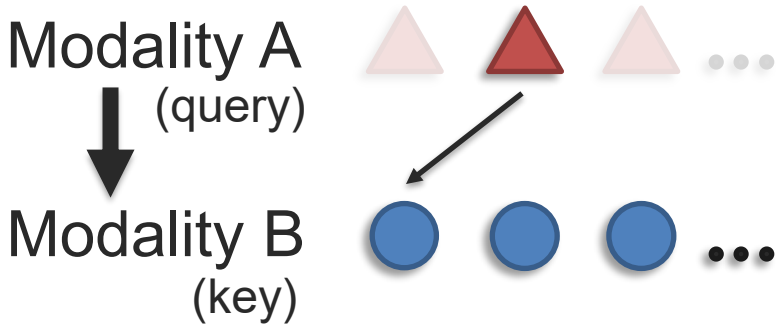


or contrastive learning

Supervision: Paired data



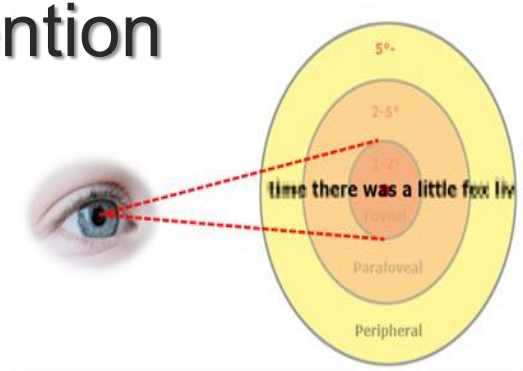
Directed Alignment



A woman is throwing a frisbee

Which object?

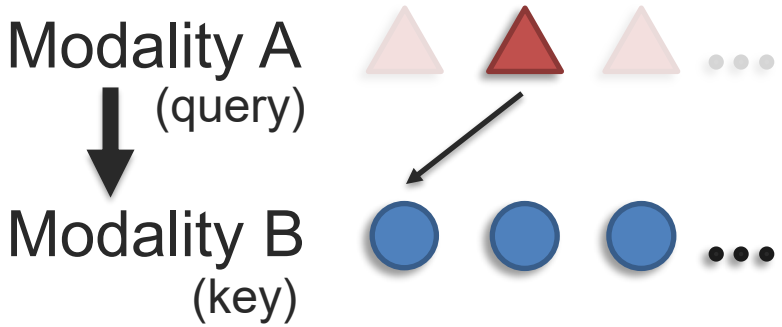
Attention

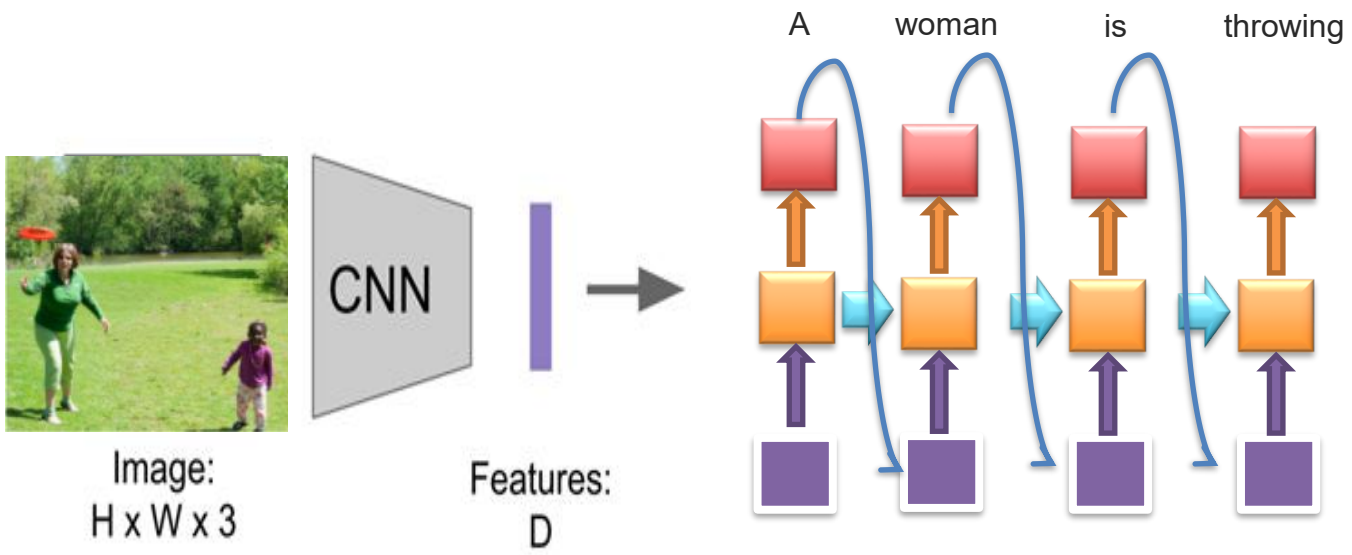


1 Soft attention

2 Hard attention

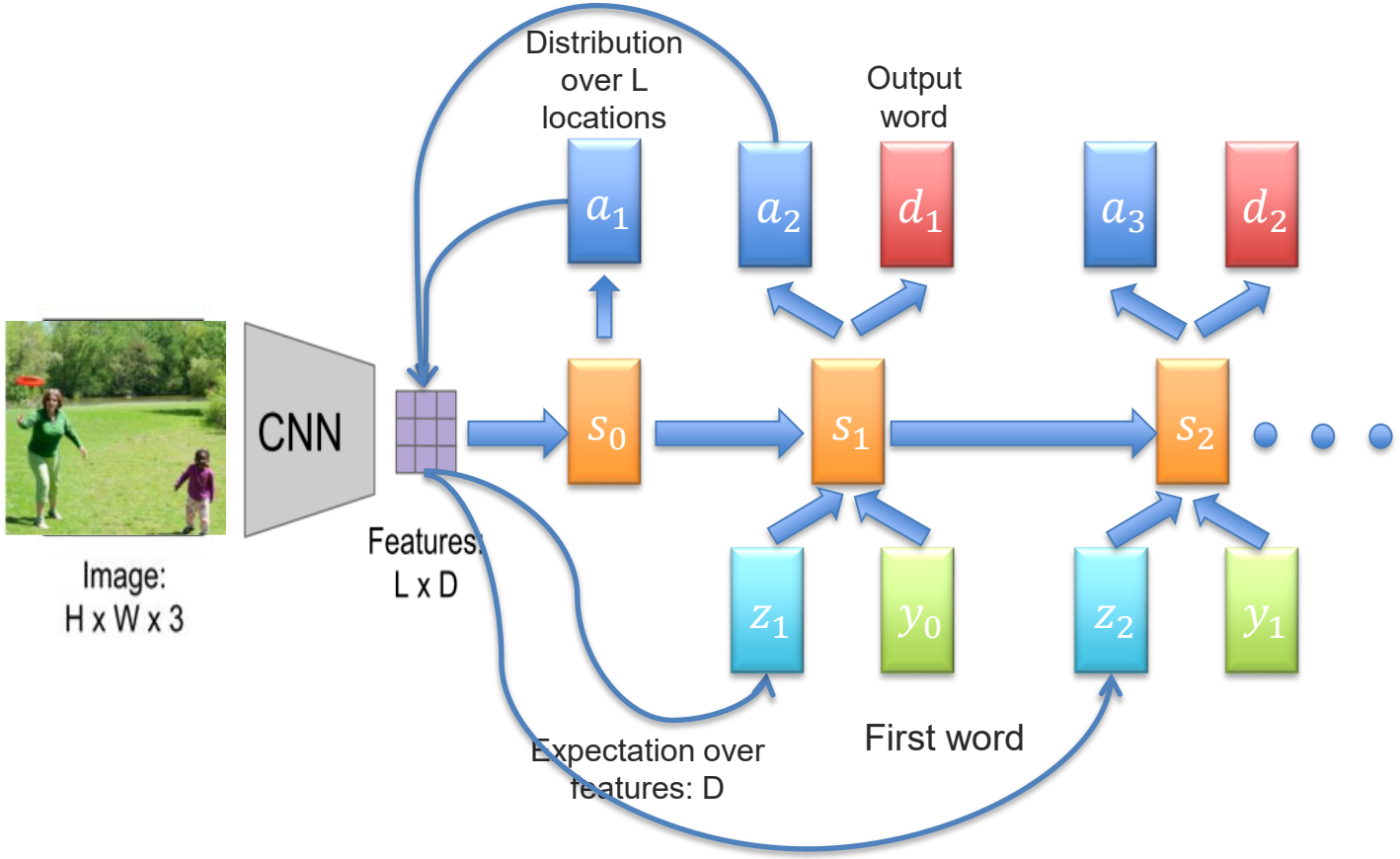
Directed Alignment – Image Captioning





Should we always use the final layer of the CNN for all generated words?

Directed Alignment – Image Captioning



Attention Gates

Before:

$$p(y_i|y_1, \ldots, y_{i-1}, \boldsymbol{x}) = g(y_{i-1}, \boldsymbol{s}_i, \boldsymbol{z}),$$

where $\boldsymbol{z} = \boldsymbol{h}_T$, last encoder state and $\boldsymbol{s}_i$ is the current state of the decoder

Now:

$$p(y_i|y_1, \ldots, y_{i-1}, \boldsymbol{x}) = g(y_{i-1}, \boldsymbol{s}_i, \boldsymbol{z}_i)$$

Have an attention "gate"

- A different context $\boldsymbol{z}_i$ used at each time step!
- $\boldsymbol{z}_i = \sum_{j=i}^{T_x} \alpha_{ij} \boldsymbol{h}_j$

α_{ij} is the (scalar) attention for word j at generation step i

Attention Gates

So how do we determine α_{ij}?

$$\alpha_{i,j} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$ => softmax, making sure they sum to 1

where:

$$e_{ij} = \boldsymbol{v}^T \sigma\left(W s_{i-1} + U h_j\right)$$

a feedforward network that can tell us how important the current encoding is

$\boldsymbol{v}, W, U$– learnable weights

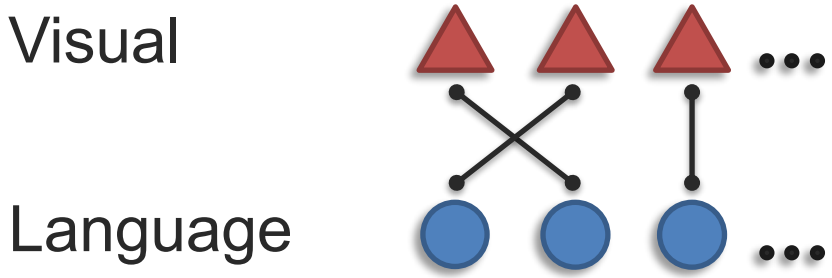
$$z_i = \sum_{j=i}^{T_x} \alpha_{ij} h_j$$

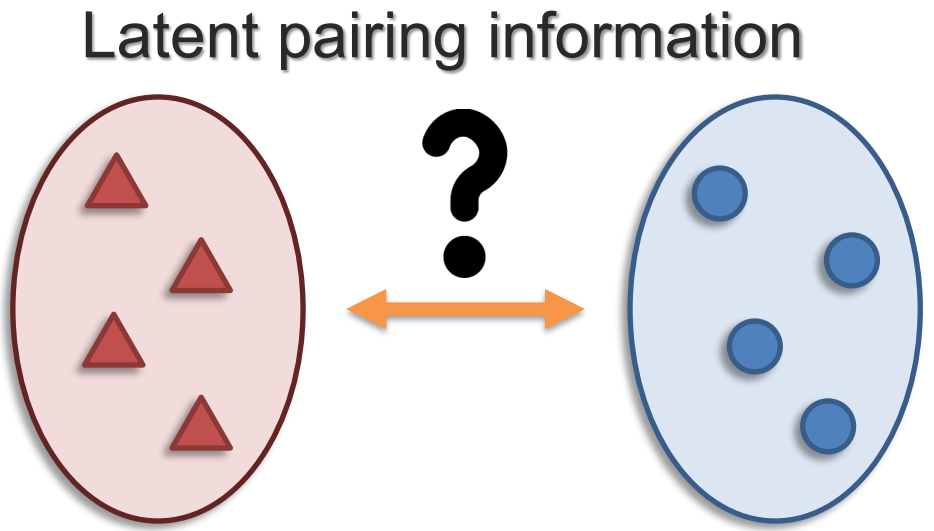
expectation of the context (a fancy way to say it's a weighted average)

Example – Image Captioning

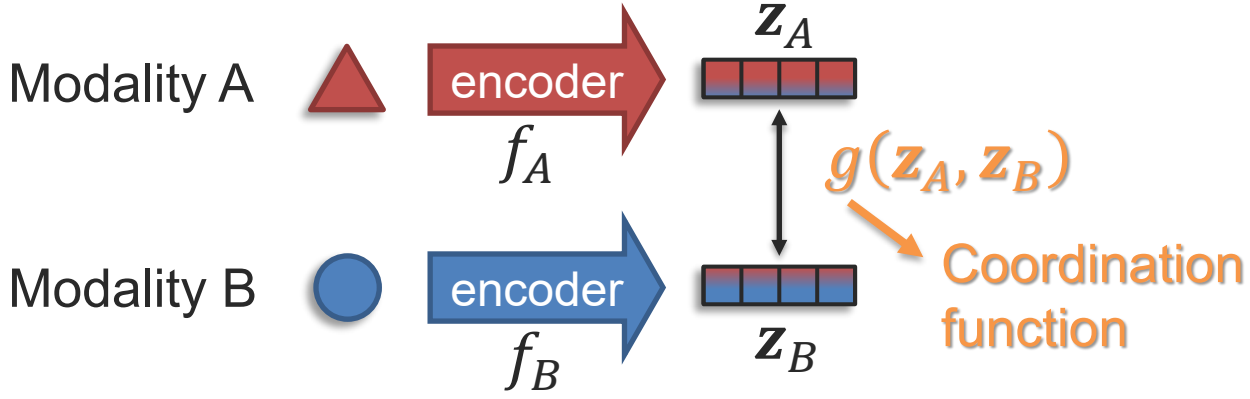
[Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, Xu et al., 2015]

Global Alignment

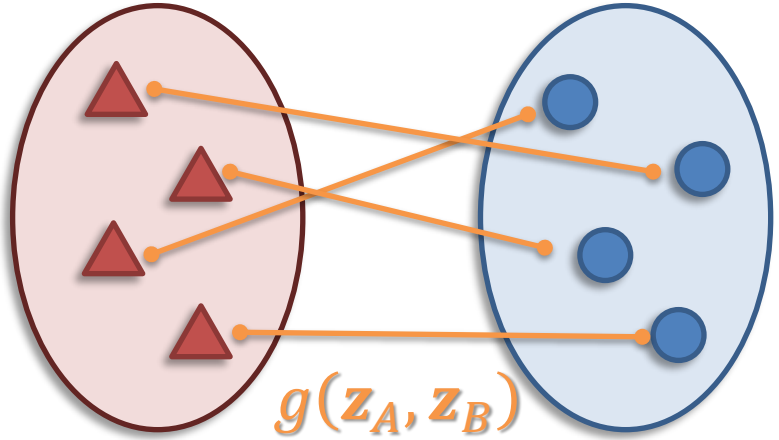




Jointly optimize representation + global alignment:

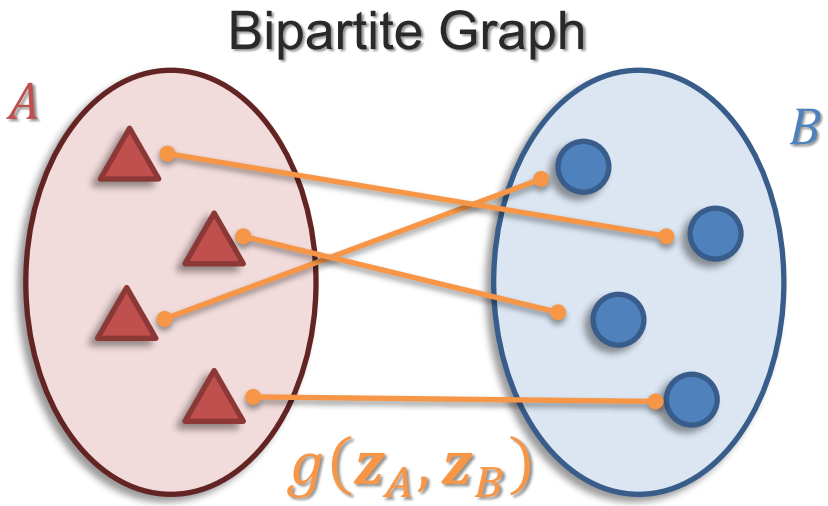


(representation)



(global alignment)

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:

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

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

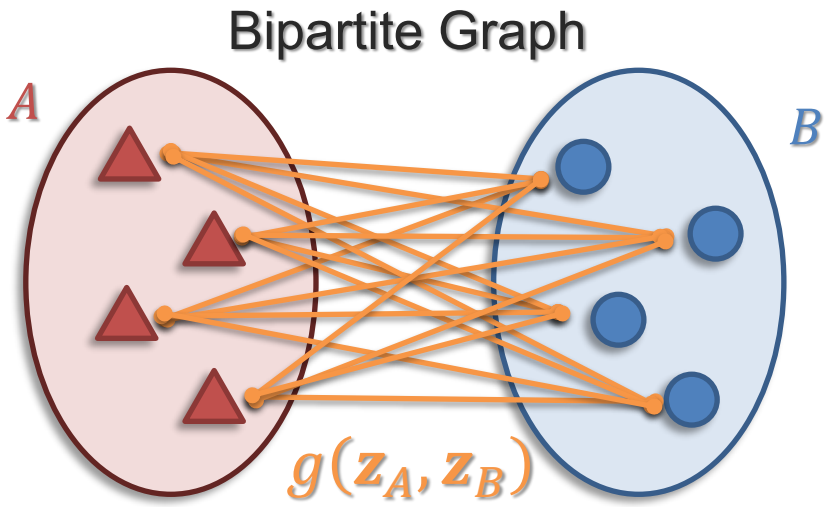
Maximize:

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

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

➡ Can be solved with simplex algorithm

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 $\boldsymbol{z}_B^j$

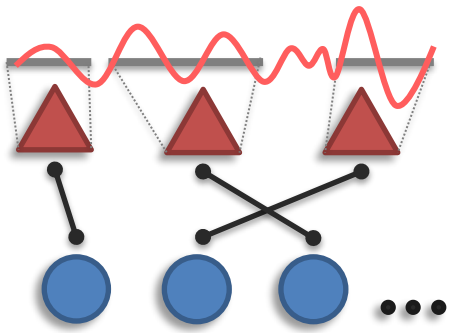
Similarity weights: $w_{(i,j)} = g(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)$

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

➡ Wassertein distance gives optimal transport

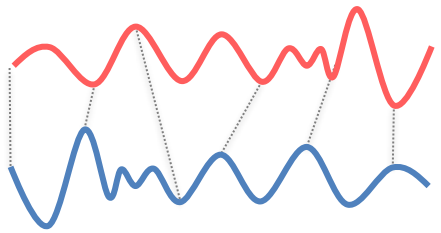
Continuous Alignment

Challenge 2b: Continuous Alignment

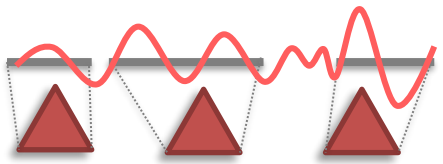


Definition: Model alignment between modalities with continuous signals and no explicit elements

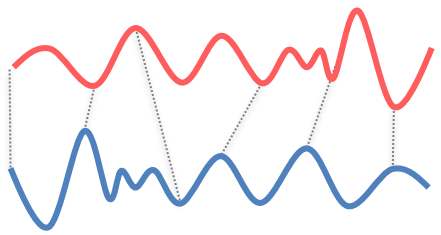
Continuous warping



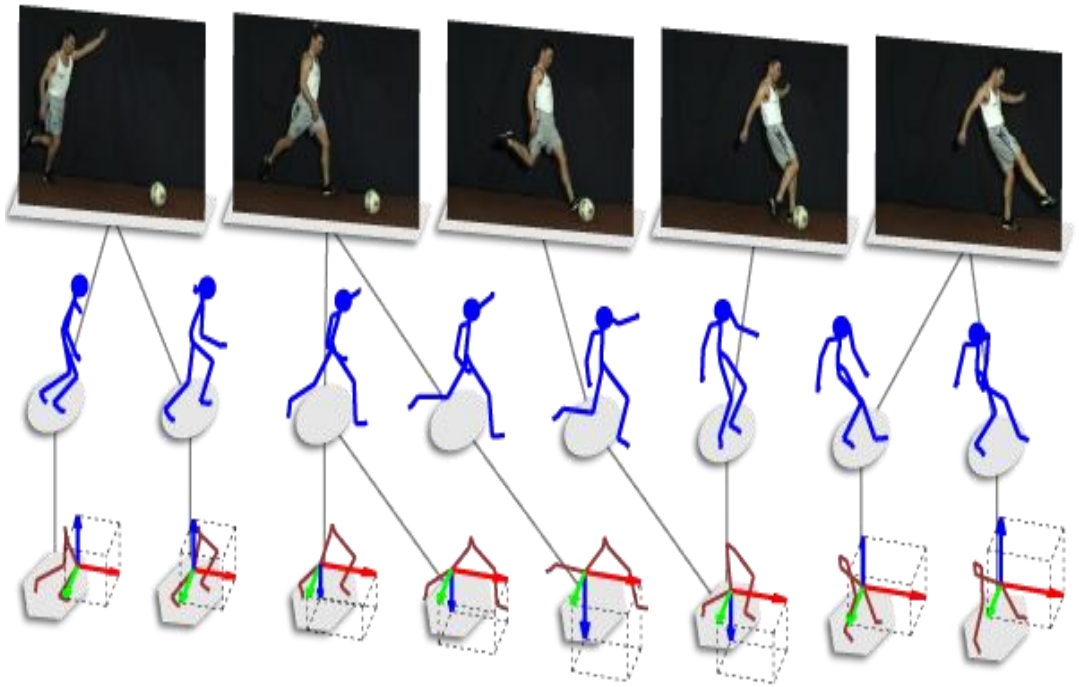
Discretization
(segmentation)



Continuous Warping – Example



Aligning video sequences



Dynamic Time Warping (DTW)

We have two unaligned temporal unimodal signals

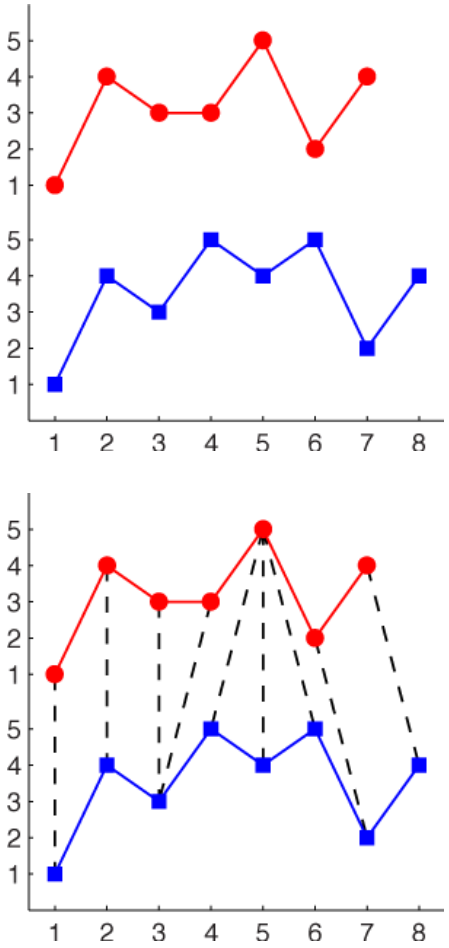
- $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_x}] \in \mathbb{R}^{d \times n_x}$
- $\mathbf{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{n_y}] \in \mathbb{R}^{d \times n_y}$

Find set of indices to minimize the alignment difference:

$$L(\boldsymbol{p}^x, \boldsymbol{p}^y) = \sum_{t=1}^{l} \left\| \boldsymbol{x}_{\boldsymbol{p}_t^x} - \boldsymbol{y}_{\boldsymbol{p}_t^y} \right\|_2^2$$

where $\boldsymbol{p}^x$ and $\boldsymbol{p}^y$ are index vectors of same length

Dynamic Time Warping is designed to find these index vectors!

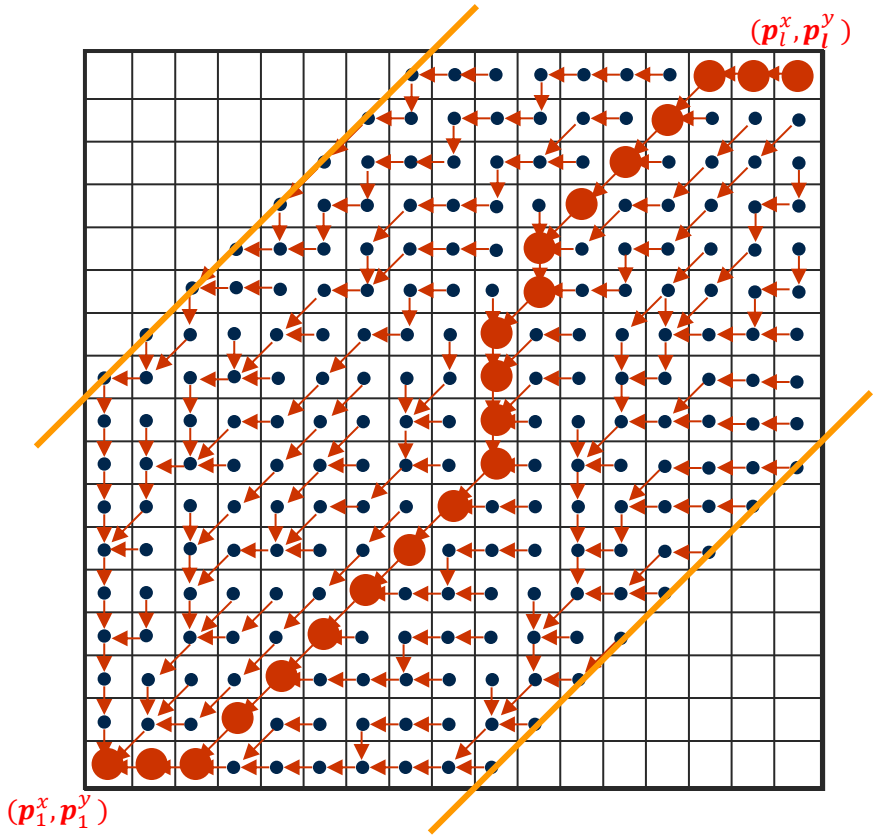


Dynamic Time Warping (DTW)

Lowest cost path in a cost matrix

- Restrictions?
  - Monotonicity – no going back in time
  - Continuity - no gaps
  - Boundary conditions - start and end at the same points
  - Warping window - don't get too far from diagonal
  - Slope constraint – do not insert or skip too much

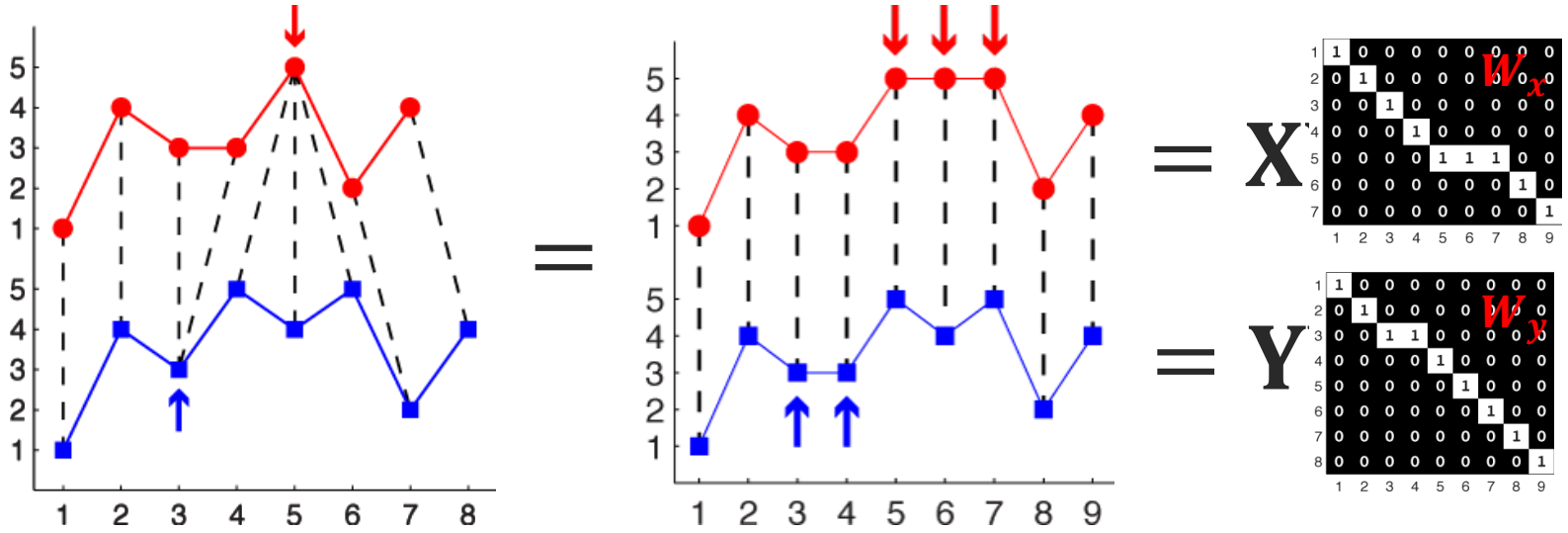
Solved using dynamic programming while respecting the restrictions



DTW alternative formulation

$$L(\boldsymbol{p}^x, \boldsymbol{p}^y) = \sum_{t=1}^{l} \left\| \boldsymbol{x}_{p_t^x} - \boldsymbol{y}_{p_t^y} \right\|_2^2$$

Replication doesn't change the objective!



Alternative objective:

$$L(\boldsymbol{W}_x, \boldsymbol{W}_y) = \left\| \boldsymbol{X}\boldsymbol{W}_x - \boldsymbol{Y}\boldsymbol{W}_y \right\|_F^2$$

Frobenius norm $\|\boldsymbol{A}\|_F^2 = \sum_i \sum_j |a_{i,j}|^2$

$\boldsymbol{X}, \boldsymbol{Y}$ – original signals (same #rows, possibly different #columns)

$\boldsymbol{W}_x, \boldsymbol{W}_y$ - alignment matrices

A differentiable version of DTW also exists…
https://arxiv.org/pdf/1703.01541.pdf

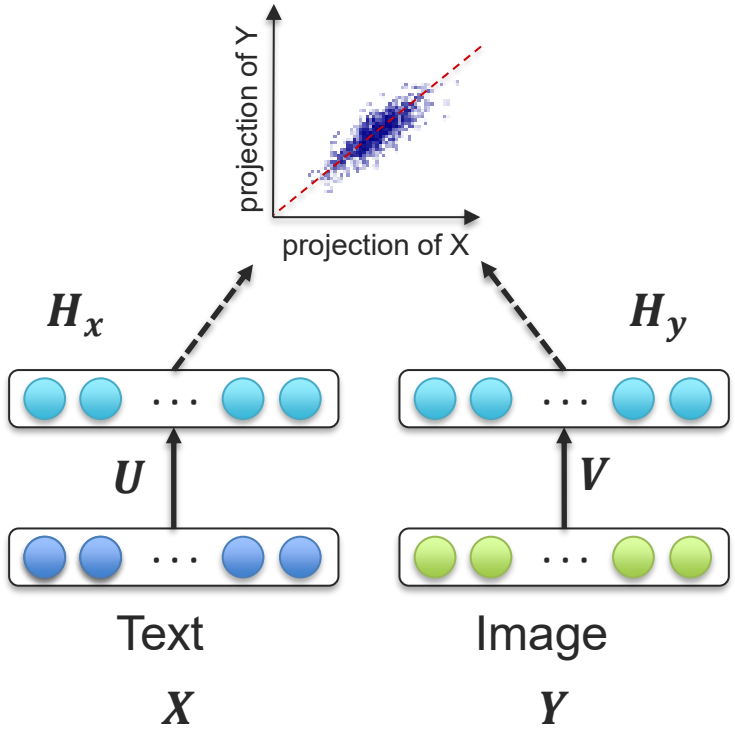
Canonical Correlation Analysis – Reminder

CCA loss can also be re-written as:

$$L(\boldsymbol{U}, \boldsymbol{V}) = \|\mathbf{U}^T\mathbf{X} - \mathbf{V}^T\mathbf{Y}\|_F^2$$

subject to:

$$\boldsymbol{U^T\Sigma_{YY}U} = \boldsymbol{V^T\Sigma_{YY}V} = \boldsymbol{I},\ \boldsymbol{u_{(j)}^T\Sigma_{XY}v_{(i)}} = \mathbf{0}$$



Canonical Time Warping

Dynamic Time Warping + Canonical Correlation Analysis = Canonical Time Warping

$$L(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}_x, \boldsymbol{W}_y) = \left\| \mathbf{U}^T \mathbf{X} \mathbf{W}_\mathbf{x} - \mathbf{V}^T \mathbf{Y} \mathbf{W}_\mathbf{y} \right\|_F^2$$

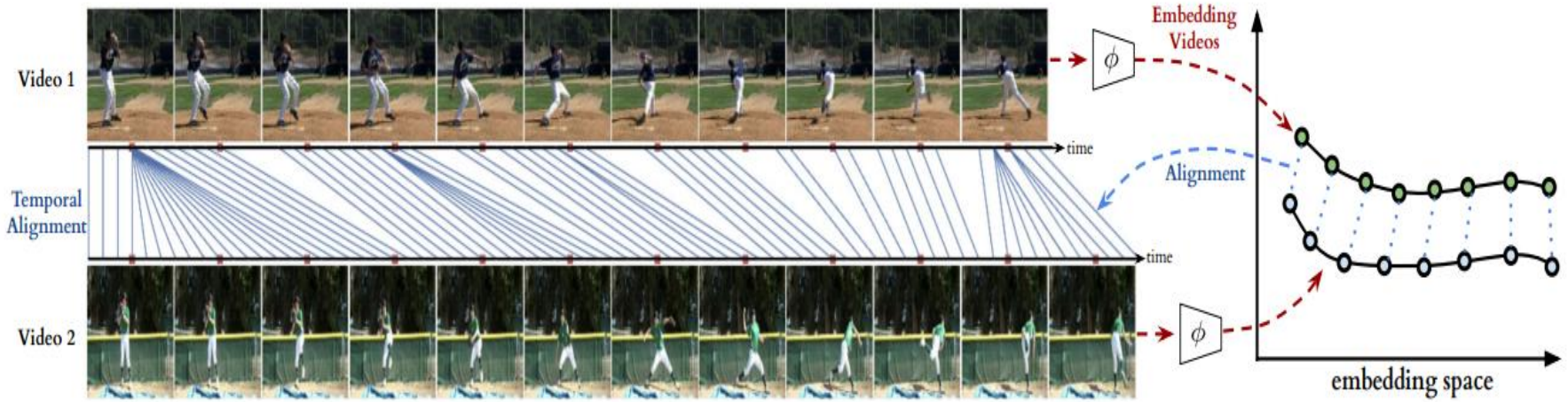
Allows to align multi-modal or multi-view (same modality but from a different point of view)

- $\boldsymbol{W}_x, \boldsymbol{W}_y$ – temporal alignment
- $\boldsymbol{U}, \boldsymbol{V}$ – cross-modal (spatial) alignment

[Canonical Time Warping for Alignment of Human Behavior, Zhou and De la Tore, 2009]

Temporal Alignment and Neural Representation Learning

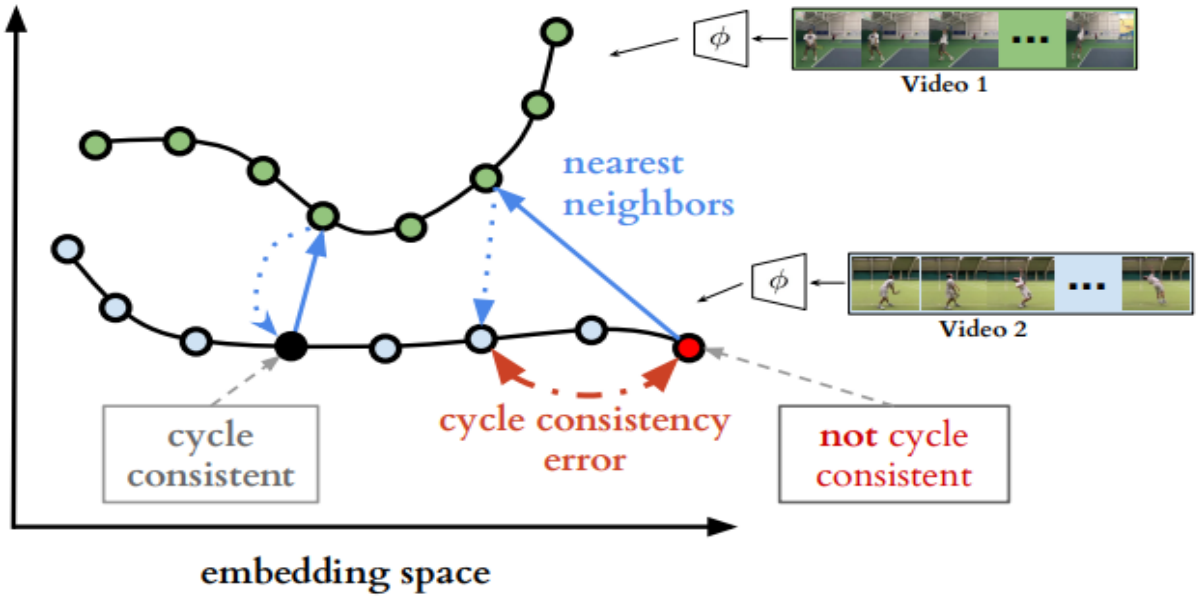
Premise: we have paired video sequences that can be be temporally aligned



How can we define a loss function to enforce the alignment between sequences while at the same time learning good representations?

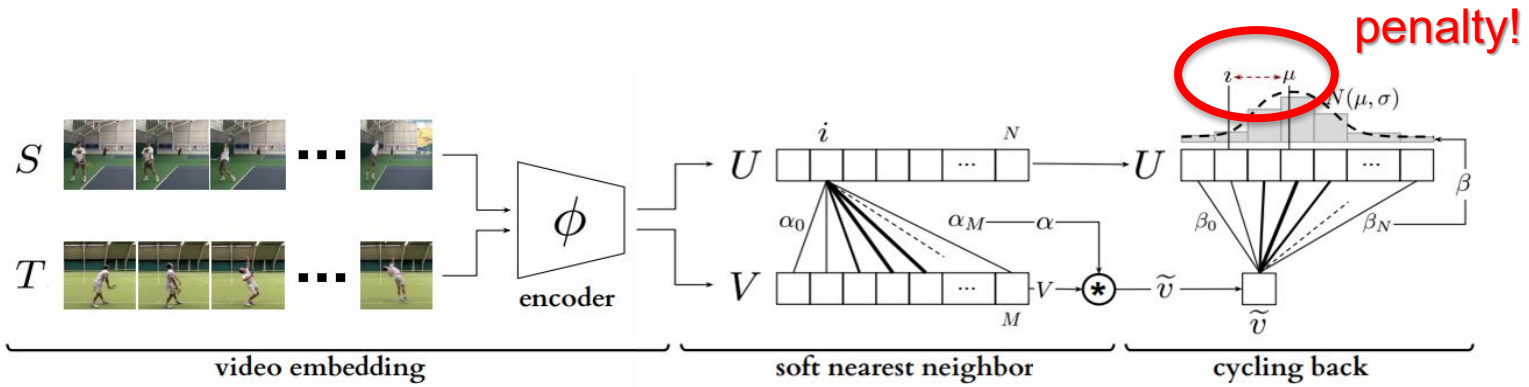
Temporal Cycle-Consistency Learning

Solution: Representation learning by enforcing **Cycle consistency**



Main idea: My closest neighbor also views me as their closest neighbor

Temporal Cycle-Consistency Learning



Compute “soft” / “weighted” nearest neighbour:

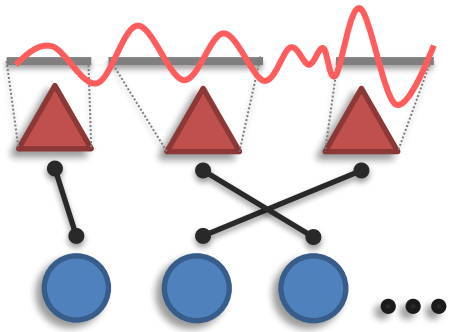
distances: $\alpha_j = \frac{e^{-||u_i - v_j||^2}}{\sum_k^M e^{-||u_i - v_k||^2}}$ Soft nearest neighbor: $\widetilde{v} = \sum_j^M \alpha_j v_j,$

Find the nearest neighbor the other way and then penalize the distance:

$$\beta_k = \frac{e^{-||\widetilde{v} - u_k||^2}}{\sum_j^N e^{-||\widetilde{v} - u_j||^2}}$$

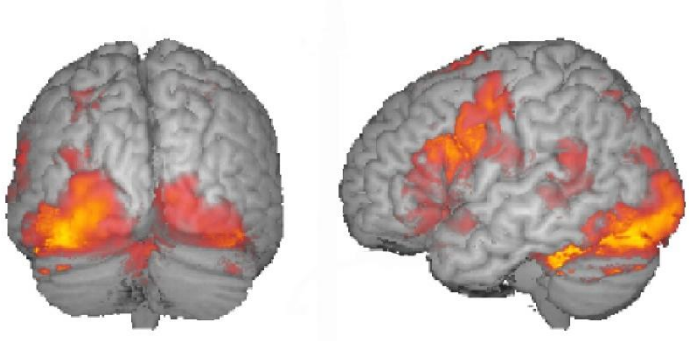
$$L_{cbr} = \frac{|i - \mu|^2}{\sigma^2} + \lambda \log(\sigma)$$

Discretization (aka Segmentation)

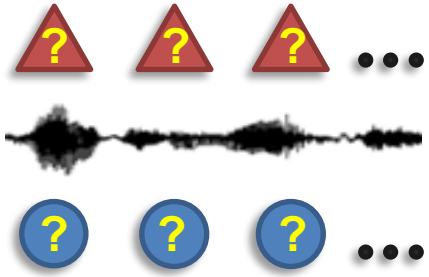


Common assumptions: ① Segmented elements

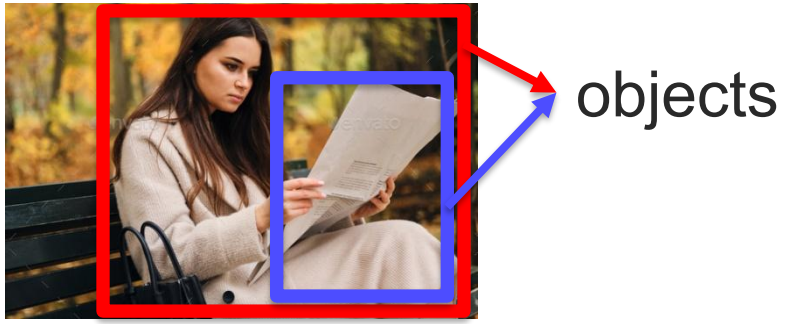
Examples:



Medical imaging



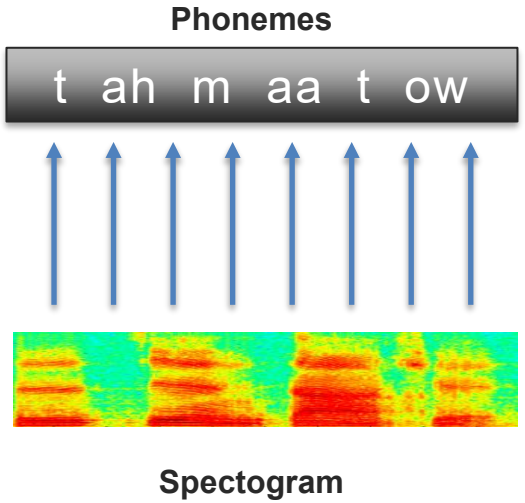
Signals



Images

Discretization – Example

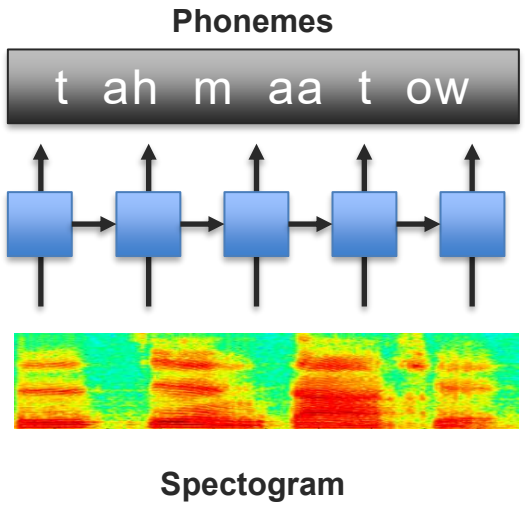
Sequence Labeling and Alignment



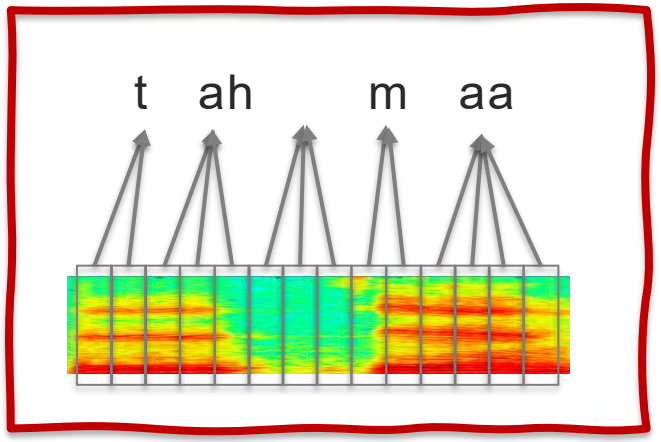
How can we predict the sequence of phoneme labels?

Discretization – Example

Sequence Labeling and Alignment



Challenge: many-to-1 alignment



How can we predict the sequence of phoneme labels?

Discretization – A Classification Approach

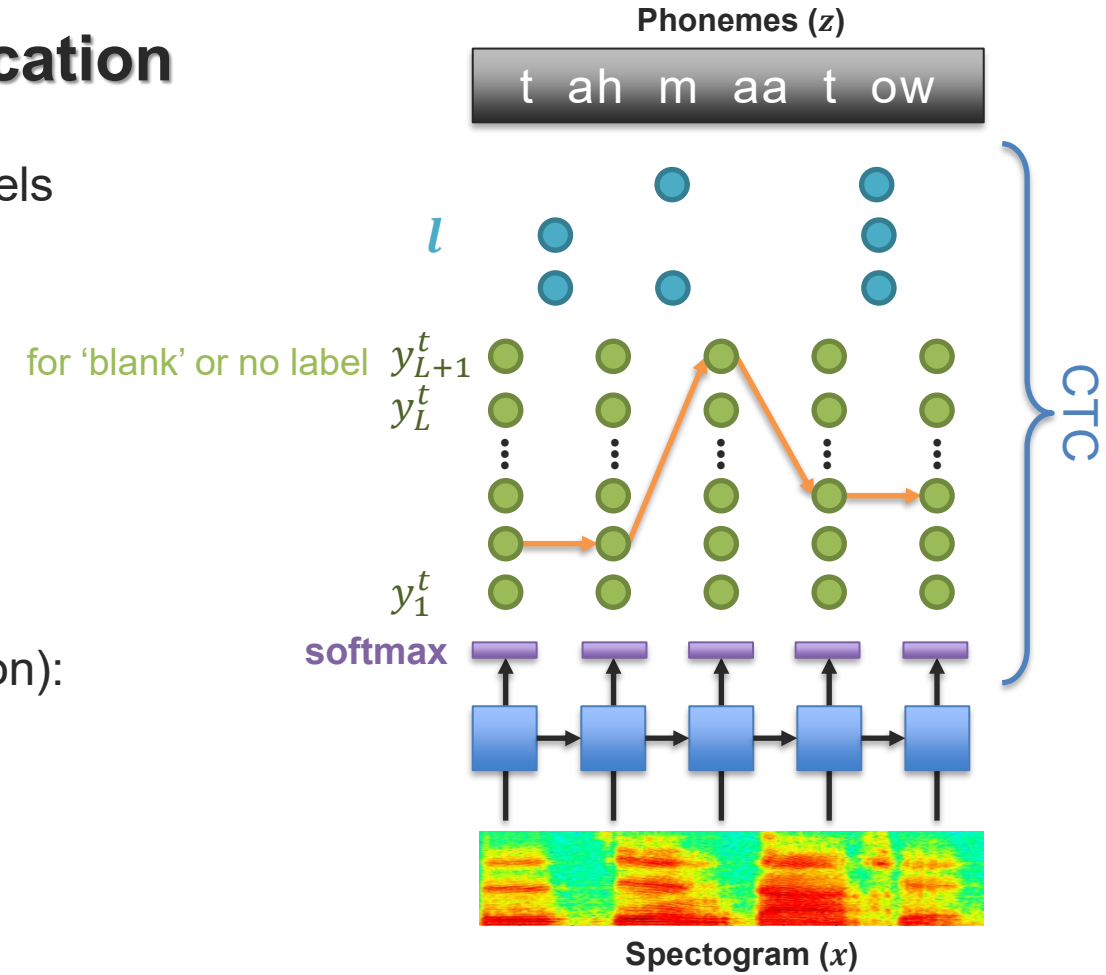
Connectionist Temporal Classification

4. Most probable sequence labels

3. Predicted labels l

2. Path π over the activations:

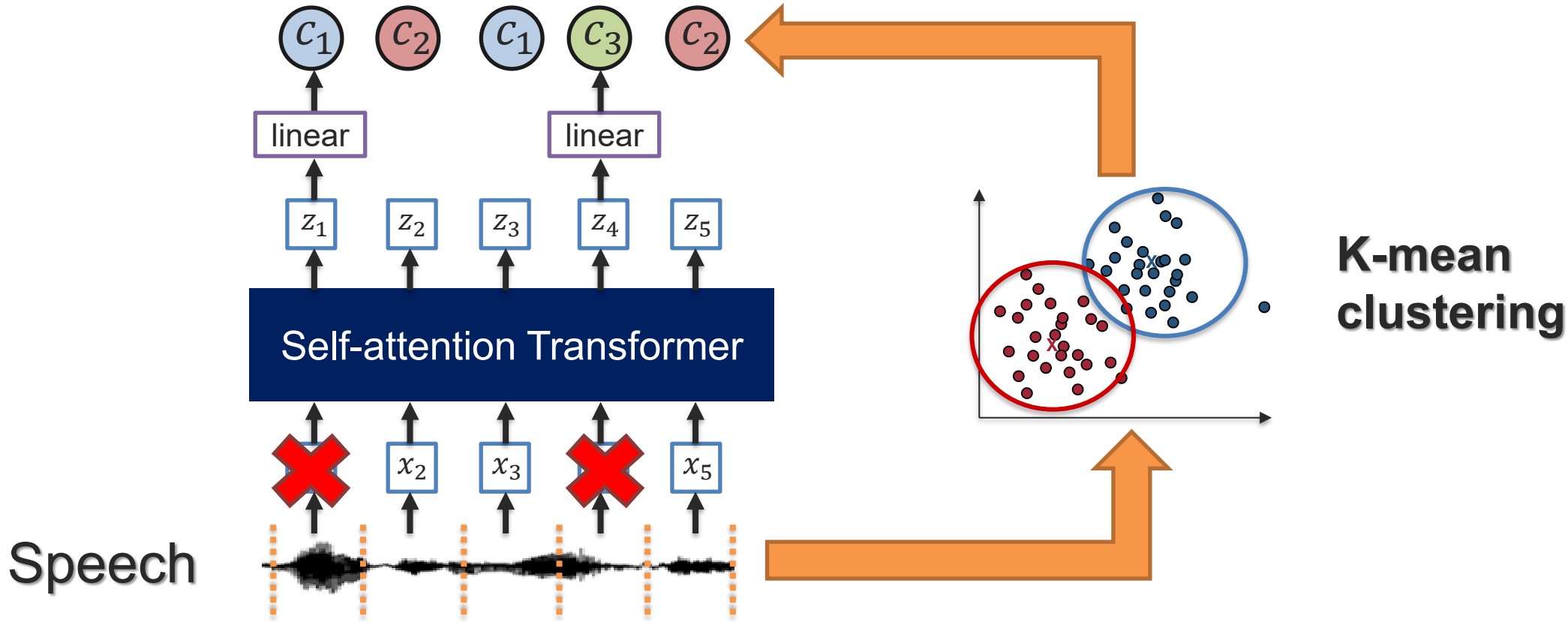
1. Output activations (distribution):



Grave et al., Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks, ICML 2006

Discretization and Representation – Cluster-based Approaches

HUBERT: Hidden-Unit BERT



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