Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 4.1: Multimodal alignment

Louis-Philippe Morency

** Co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk.*

Administrative Stuff

Primary TAs

- Each team will have one primary TA
- Contact your primary TA anytime
  - Groups were 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 9/24 at 8pm

Four main sections:

- Introduction
- Related work
- Experimental setup
- Research ideas

The two main sections are related work and research ideas

teammates = # research ideas

Page limit depends on team size:

- 3 students : 4 pages + references
- 4 students : 4.5 pages + references
- 5 students : 5 pages + references
- 6 students : 5.5 pages + references

Follows ICML paper format

Team Meetings with Instructor

- No lecture on Tuesday 10/3
- 15-mins meeting with instructor
  - Optional, but highly suggested
  - Not all teammates are required to attend
- Meetings next week: Wednesday 9/27 until Friday 9/29
- Signup form will be shared via Piazza

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 4.1: Multimodal alignment

Louis-Philippe Morency

** Co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk.*

Lecture objectives

- A quick review
  - Connections, coordinated representations and mutual information
  - Modality interactions and factorized representations
- Discrete alignment
  - Local alignment
    - Coordinated representations; hard and soft attention
  - Global alignment
    - Assignment problem and optimal transport
- Continuous alignment
  - Continuous warping
    - Dynamic time warping
  - Discretization and segmentation

A Quick Review

Multimodal is the scientific study of heterogeneous, connected and interacting data

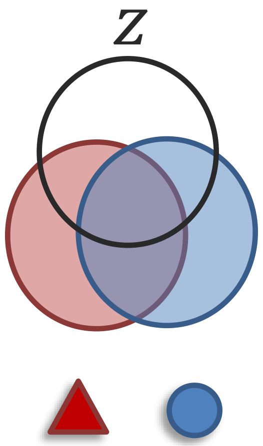
Interacting

+

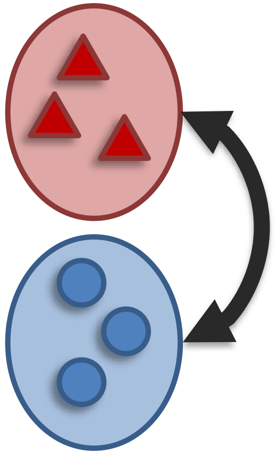
Connected

+

Heterogeneous

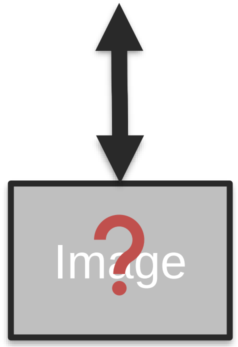


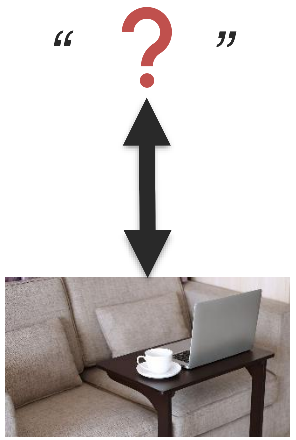
Connections between Modalities



Connections?
(associations)

“*kitchen*”

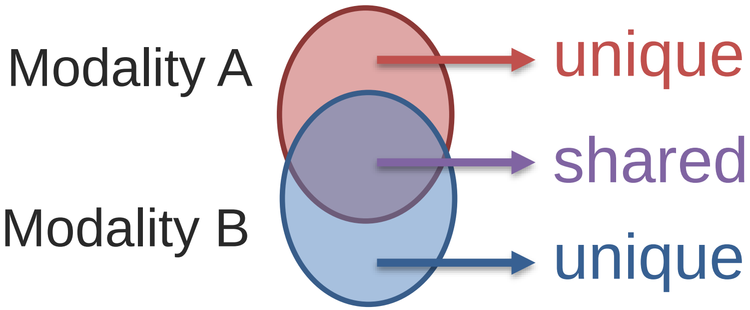


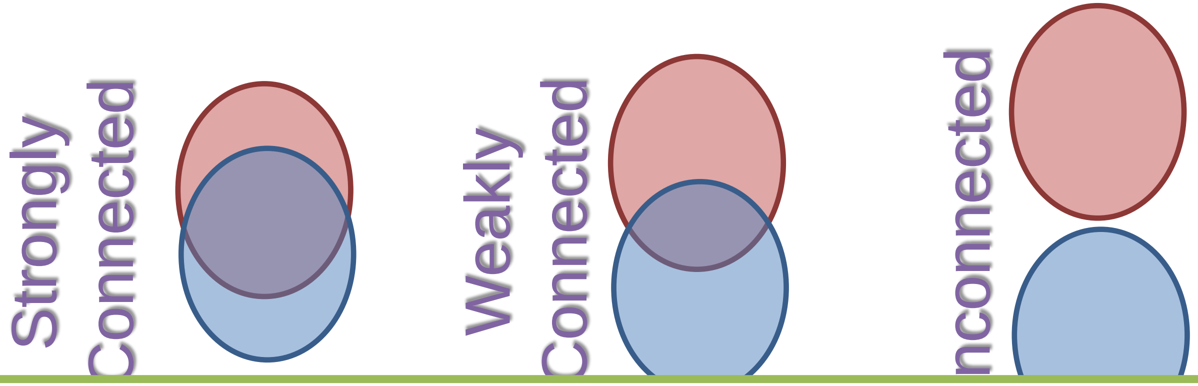


Connection types:

- Co-occurrence
- Correlation
- Causality
- Relationship
- ⋮

➡ knowledge of one modality provides information about the other modality



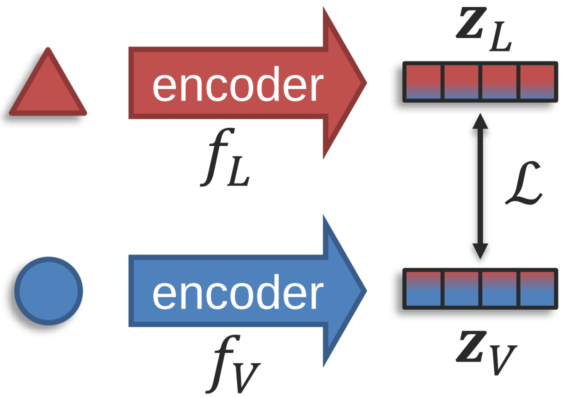


Connections are part of the data… and models will try to learn them.

Coordinated Representations – Example: CLIP

Language

Visual
(image)



Popular contrastive loss: InfoNCE

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

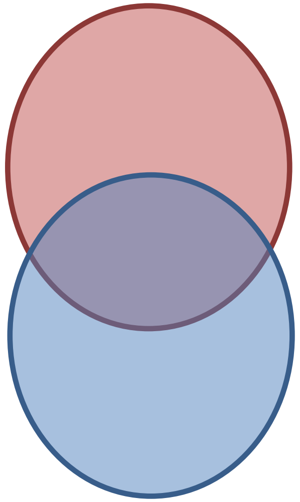
positive pairs

negative pairs
and positive pairs

Similarity function can
be cosine similarity

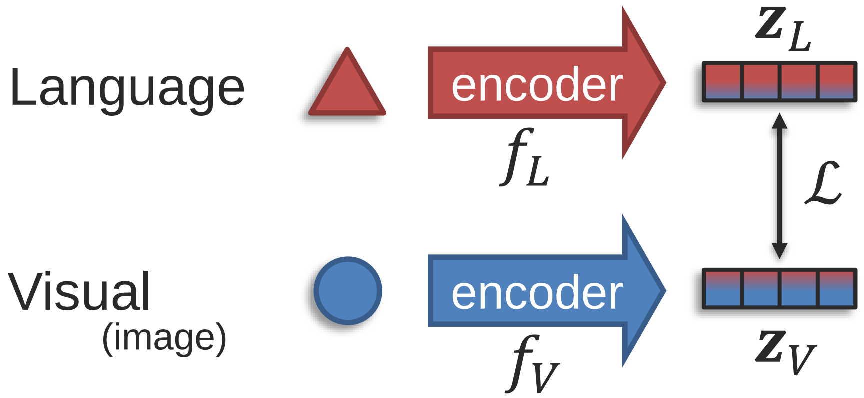
Connected modalities:

Images ▲
Language ●



Radford et al., Learning Transferable Visual Models From Natural Language Supervision, arxiv 2021

Coordinated Representations – Example: CLIP



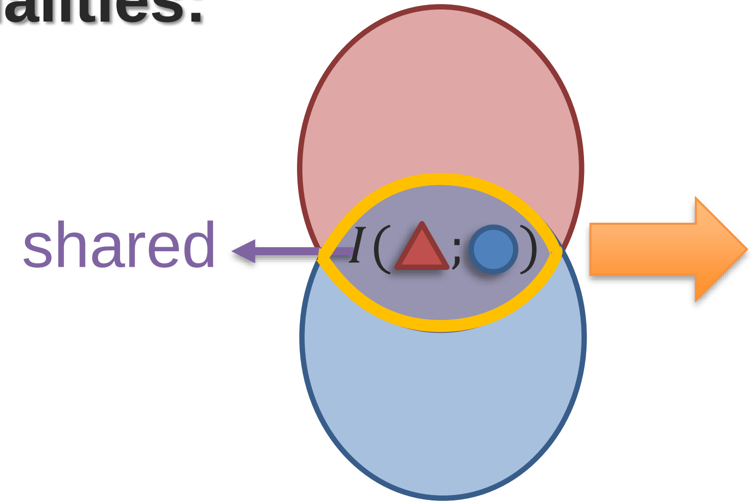
Popular contrastive loss: InfoNCE

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

CLIP focuses on shared connections

Connected modalities:

Images	▲
Language	●



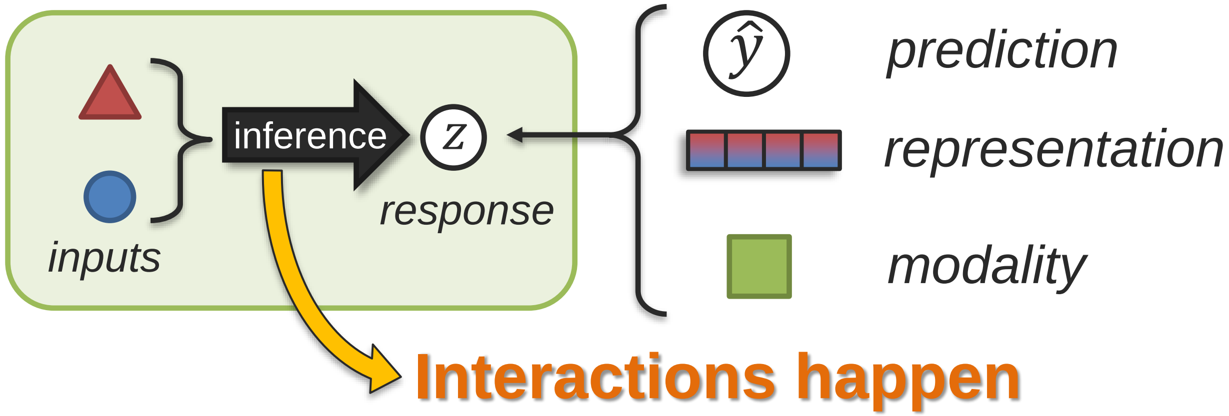
Mutual information $I(X;Y)$

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

Radford et al., Learning

More about connections and alignment today!

Modality Interactions

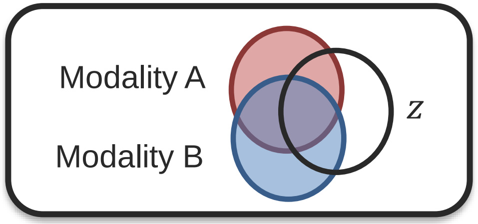


Interactions happen during inference! (from human or model)

Interactions require more than the input modalities!

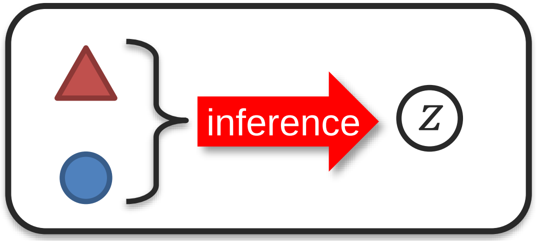
Interactions taxonomy:

Level 1: Response(s) and Input Modalities



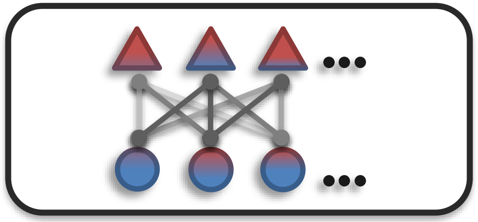
- ❑ Co-occurrence
- ❑ Redundancy
- ❑ Dominance
- ❑ Emergence
- …

Level 2: Interactions – Internal Mechanics



- ❑ Additive
- ❑ Multiplicative
- ❑ Polynomial
- ❑ Nonlinear
- …

Level 3: Contextualized Interactions



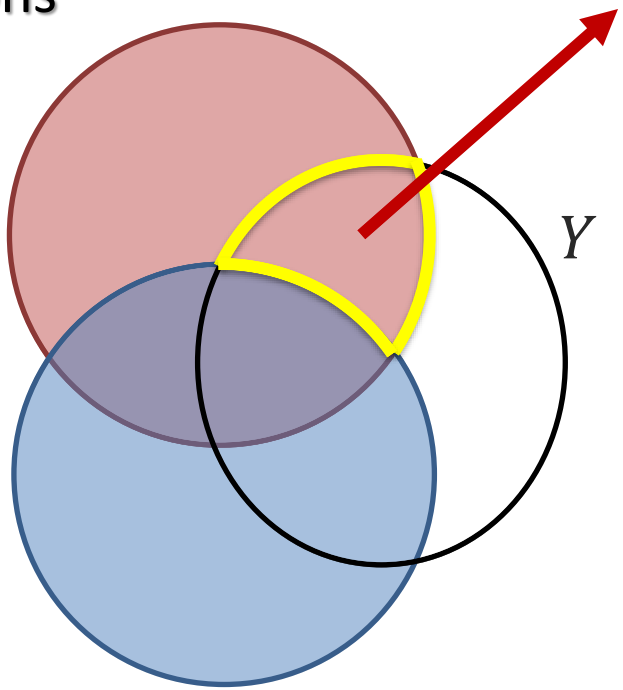
- ❑ Temporal
- ❑ Hierarchy
- ❑ Multimodal
- …

Modeling Interactions – Response and Input Modalities (Level 1)

Information theory as a framework for modeling Level 1 interactions between input modalities and responses

Modality A ▲

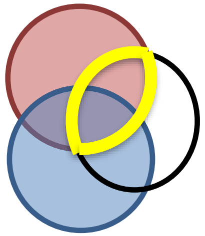
Modality B ●



Task-relevant **unique** information

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

Mutual information…



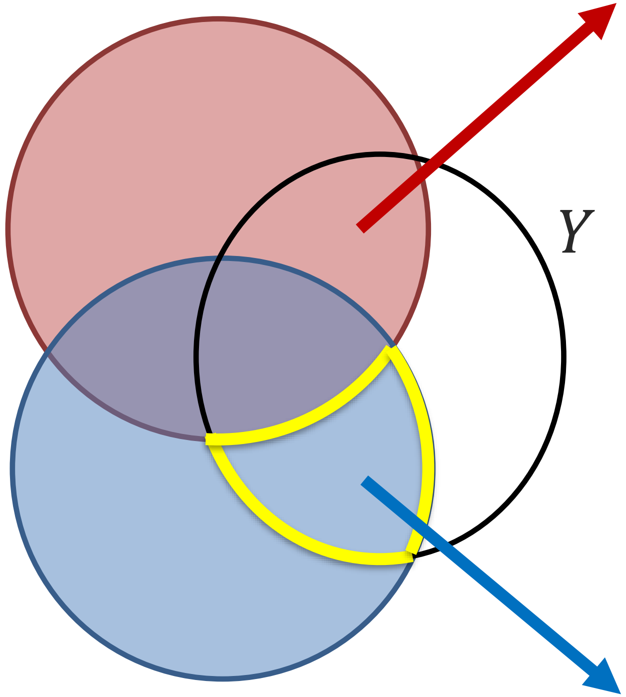
$$I(\blacktriangle; Y)$$

…but without ●

Modeling Interactions – Response and Input Modalities (Level 1)

Modality A $\blacktriangle$

Modality B $\bullet$



Task-relevant **unique** information

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

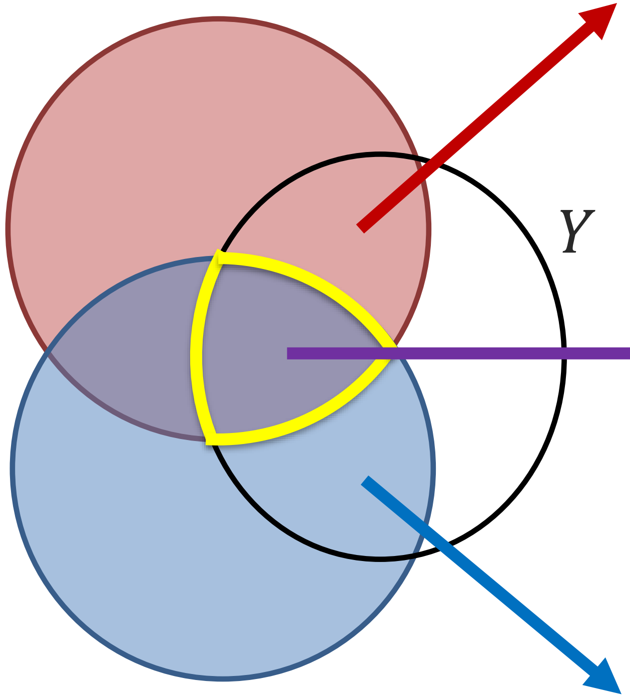
Task-relevant **unique** information

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

Modeling Interactions – Response and Input Modalities (Level 1)

Modality A $\blacktriangle$

Modality B $\bullet$



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 Interactions – Response and Input Modalities (Level 1)

Partial Information Decomposition (PID)

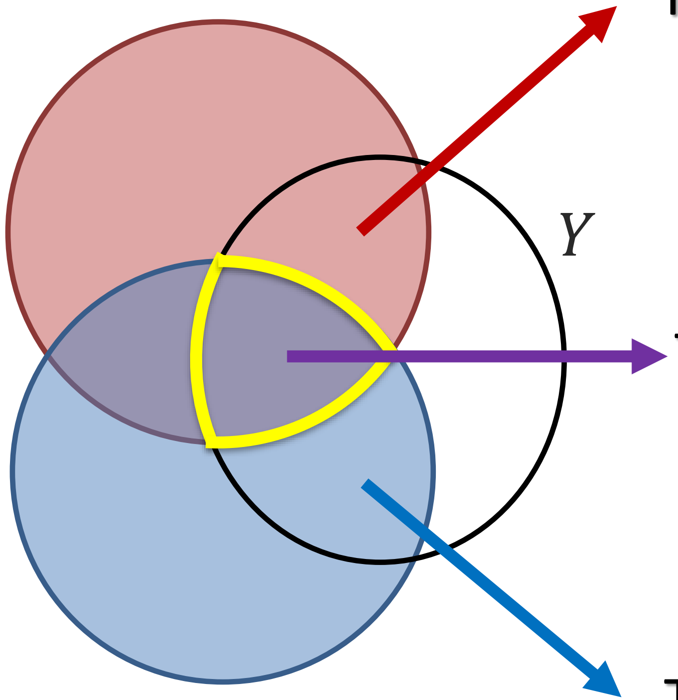
$$I(\triangle;\bigcirc;Y) = R - S$$

factorizes 3-way mutual information into:

R: redundancy

S: Synergy

More about PID in future lectures and reading assignments



Task-relevant **unique** information

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

Task-relevant **shared** information

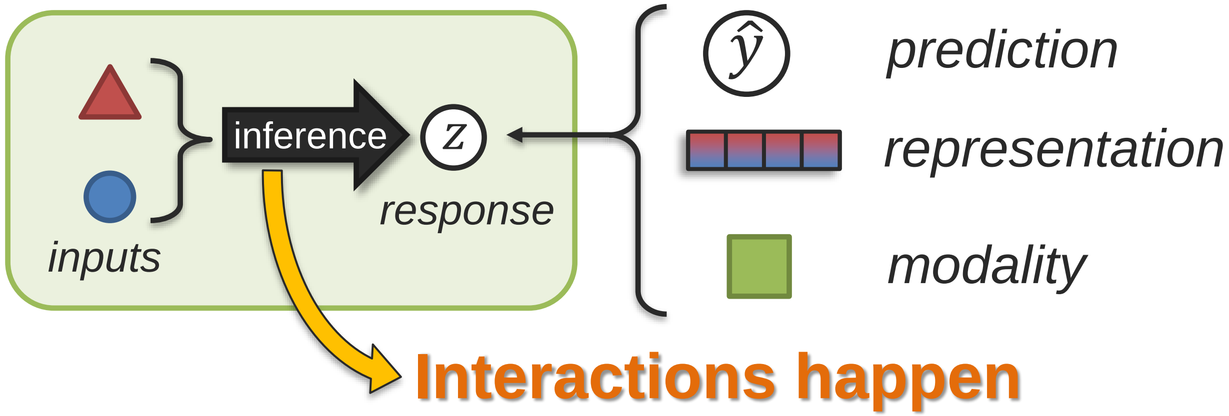
$$I(\triangle;\bigcirc;Y)$$

…but how to compute 3-way mutual information?

Task-relevant **unique** information

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

Modality Interactions

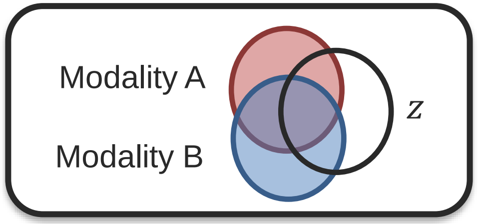


Interactions happen during inference! (from human or model)

Interactions require more than the input modalities!

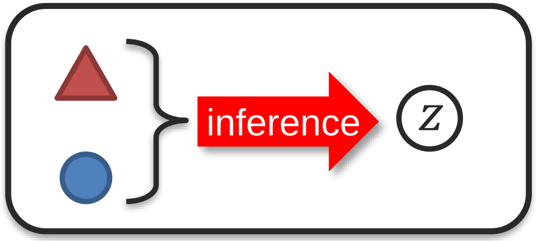
Interactions taxonomy:

Level 1: Response(s) and Input Modalities



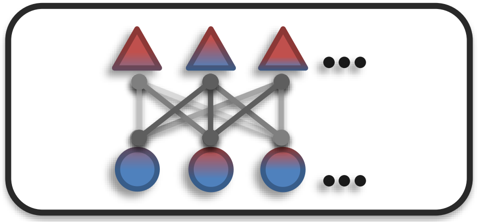
- ❑ Co-occurrence
- ❑ Redundancy
- ❑ Dominance
- ❑ Emergence
- …

Level 2: Interactions – Internal Mechanics



- ❑ Additive
- ❑ Multiplicative
- ❑ Polynomial
- ❑ Nonlinear
- …

Level 3: Contextualized Interactions



- ❑ Temporal
- ❑ Hierarchy
- ❑ Multimodal
- …

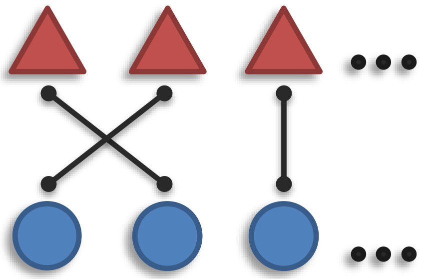
Challenge 2: Alignment

Challenge 2: Alignment

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

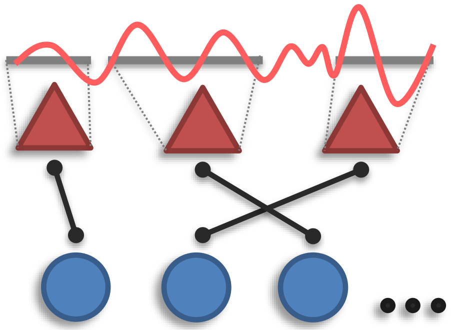
Sub-challenges:

Discrete Alignment



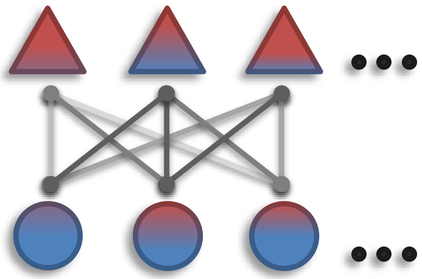
Discrete elements and connections

Continuous Alignment



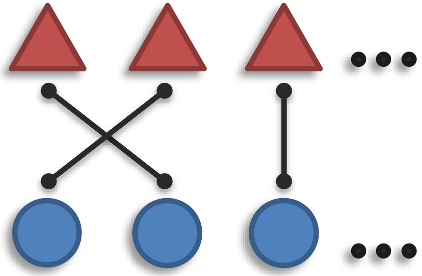
Segmentation and continuous warping

Contextualized Representation

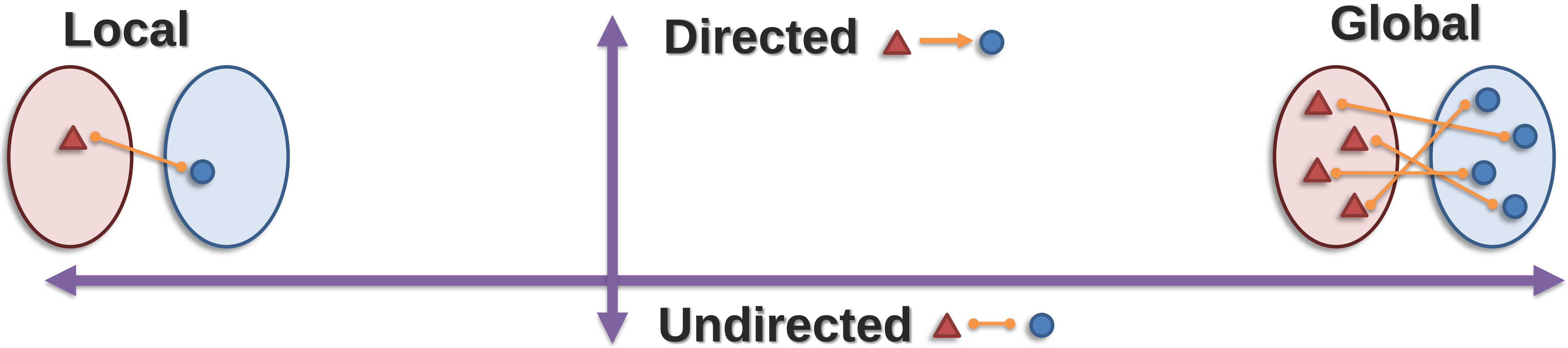


Alignment + representation

Sub-Challenge 2a: Discrete Alignment



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

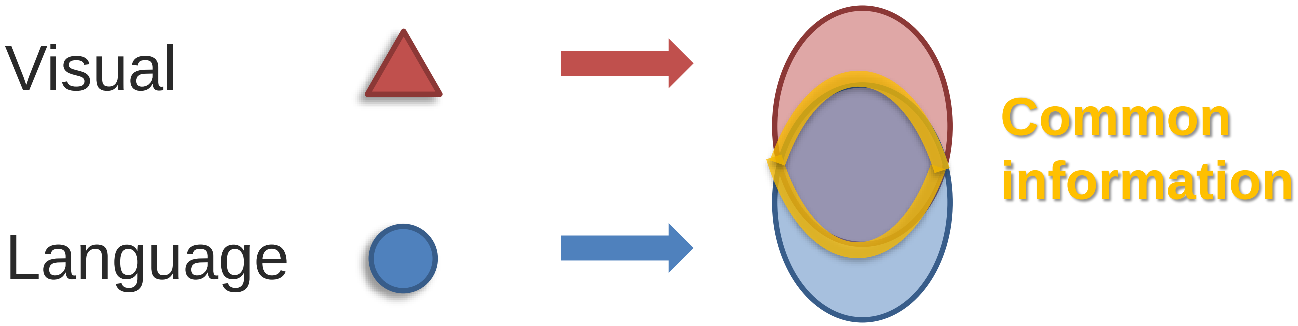


Language Grounding

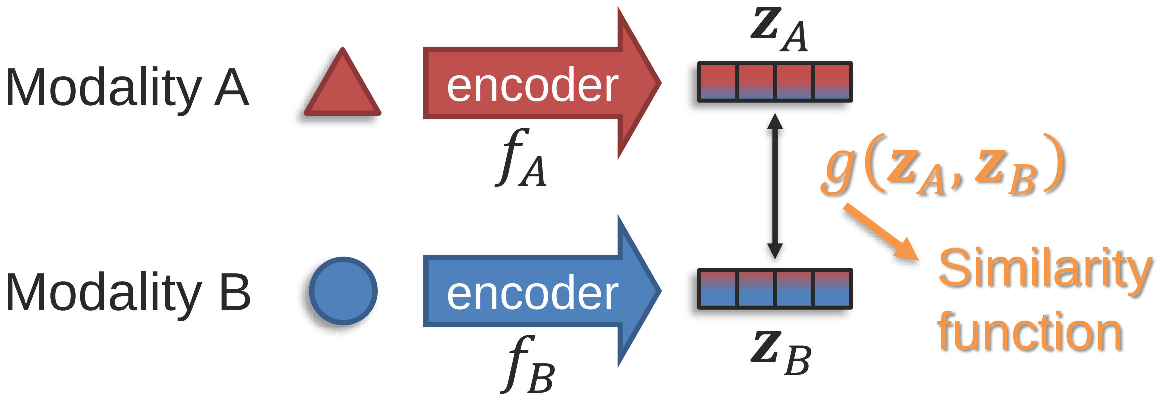
Definition: Tying language (words, phrases,…) to non-linguistic elements, such as the visual world (objects, people, …)

A woman reading newspaper

Local Alignment – Coordinated Representations

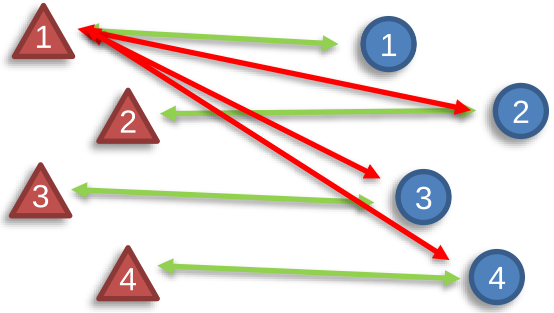


Learning coordinated representations:



or contrastive learning

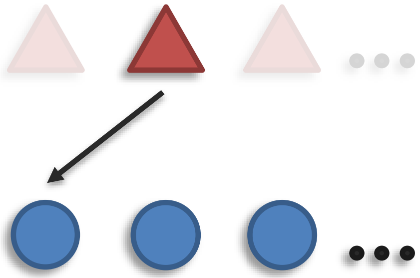
Supervision: Paired data



Directed Alignment

Modality A
(query)

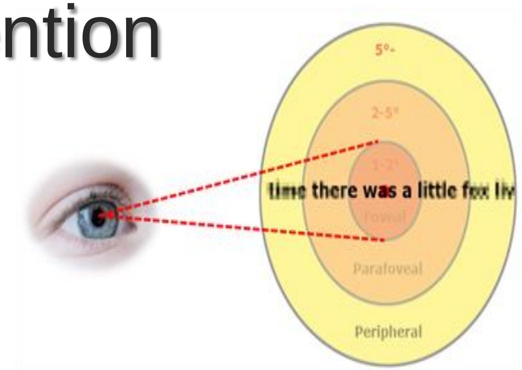
Modality B
(key)



A woman is throwing a frisbee

Which object?

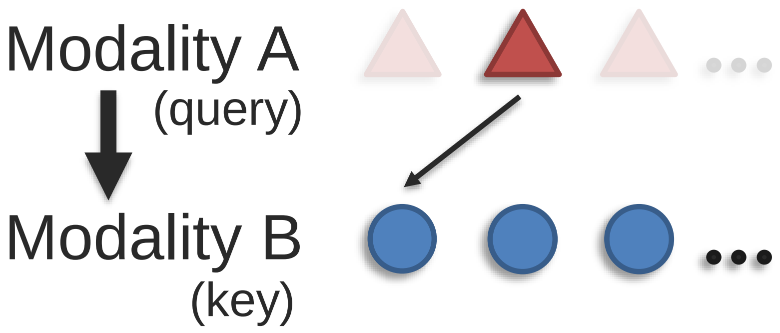
Attention

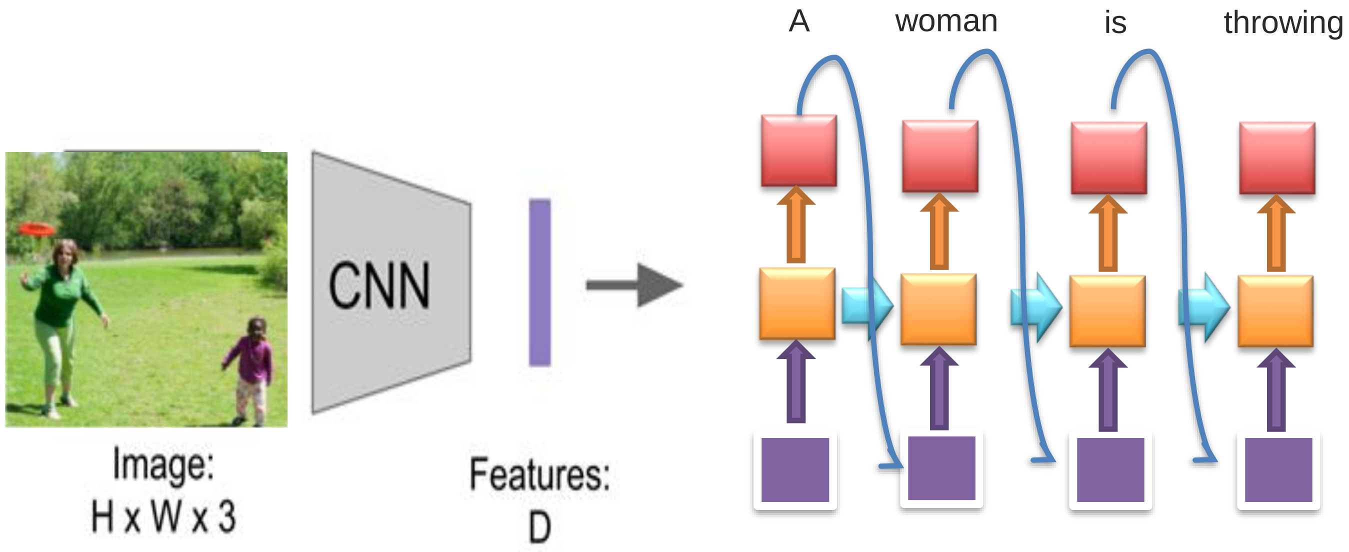


1 Soft attention

2 Hard attention

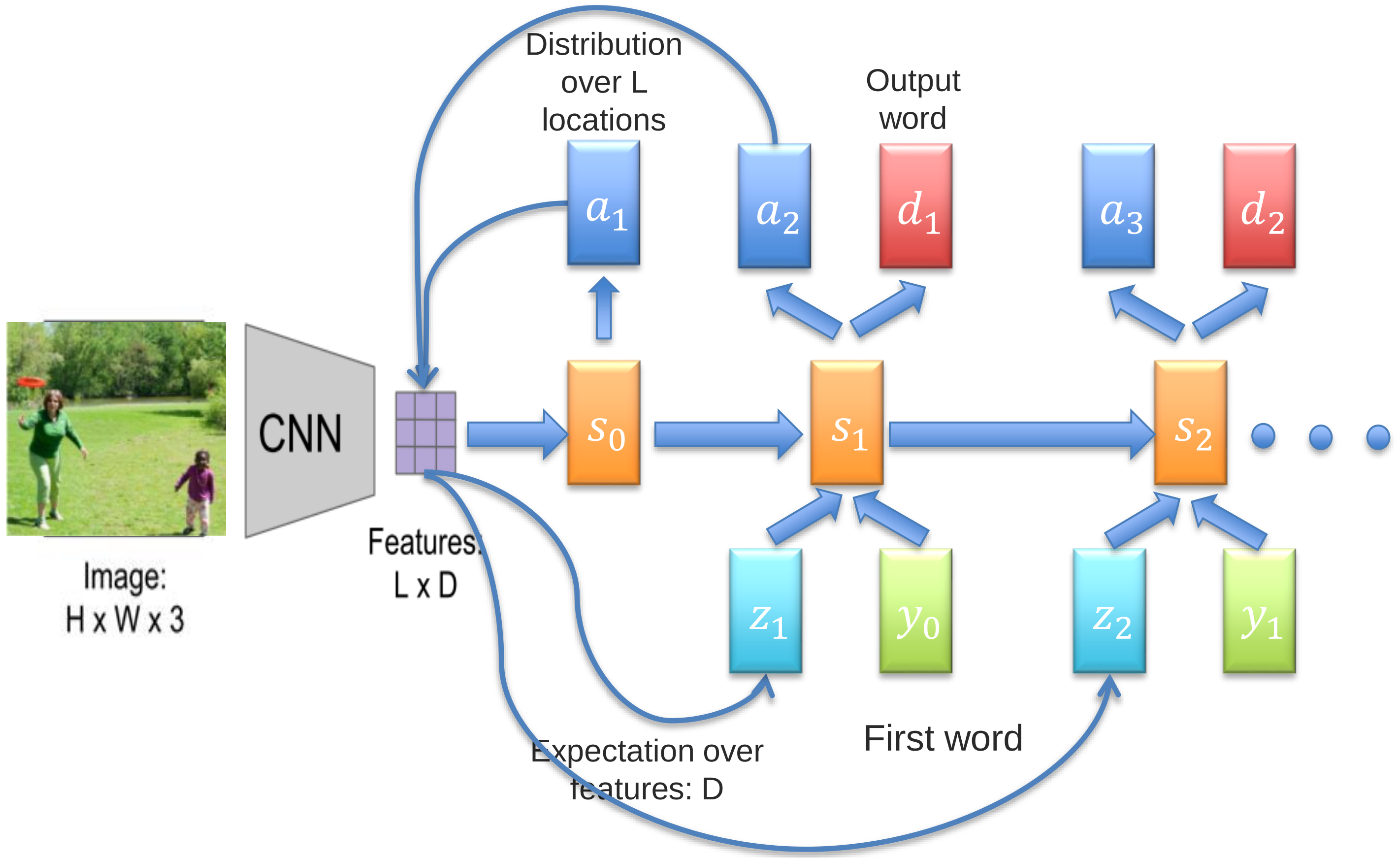
Directed Alignment – Image Captioning





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

Directed Alignment – Image Captioning



Attention Gates

Before:

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

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

Now:

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

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

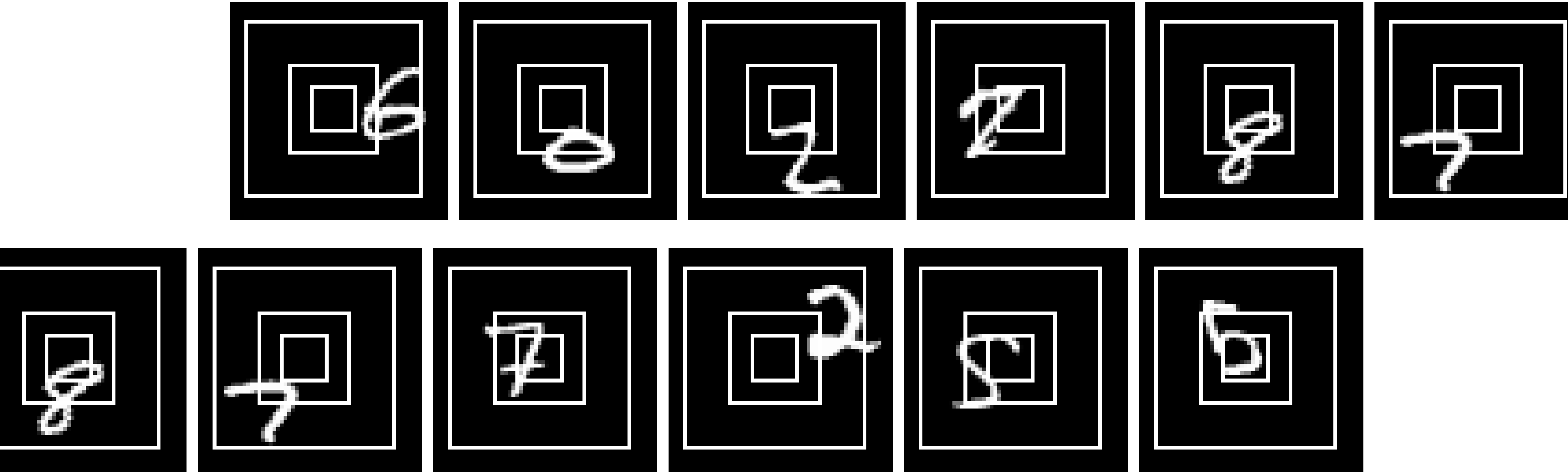
$$z_i = \sum_{j=i}^{T_x} \alpha_{ij} h_j$$

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

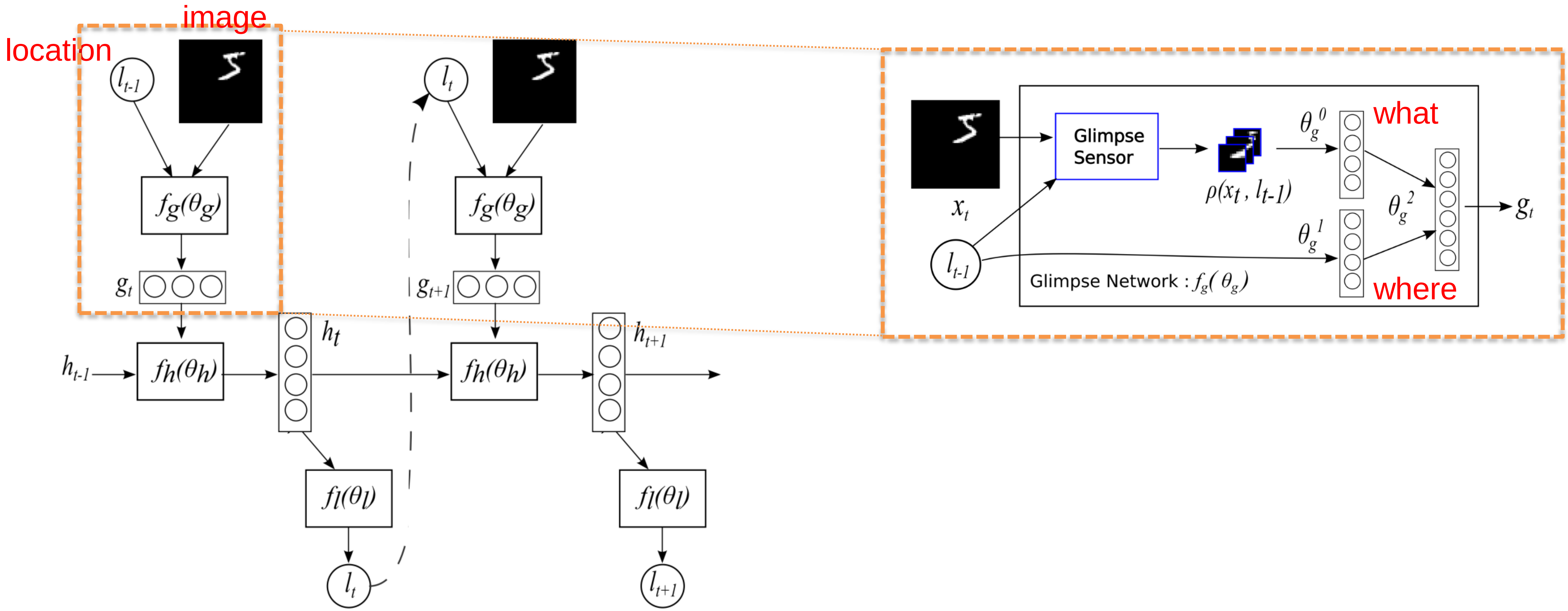
Example – Image Captioning

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

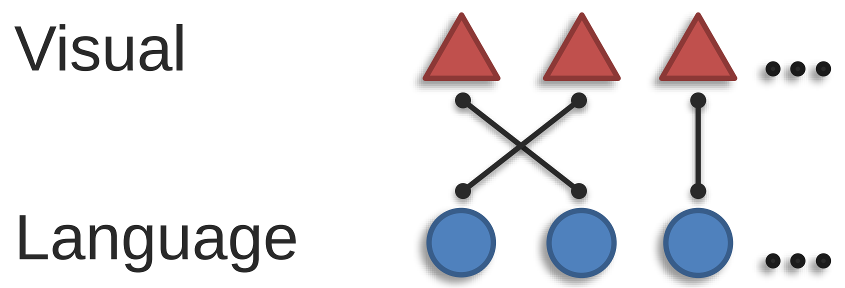
Hard attention - Example



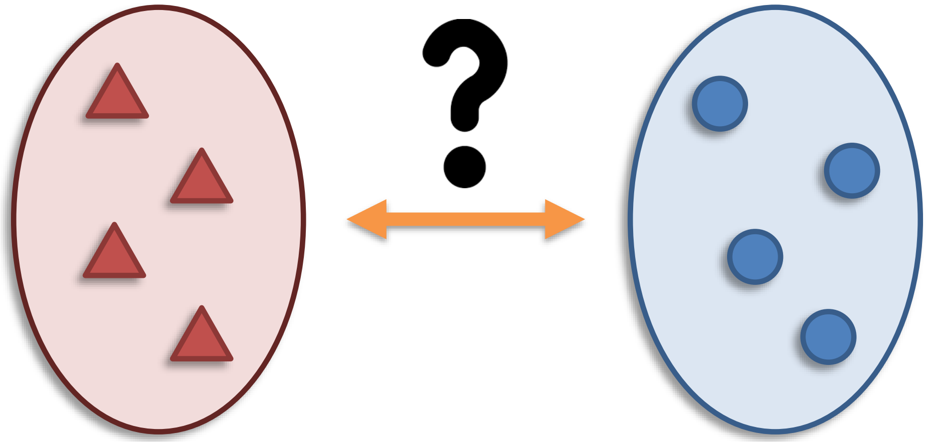
Hard Attention – Recurrent Model of Visual Attention



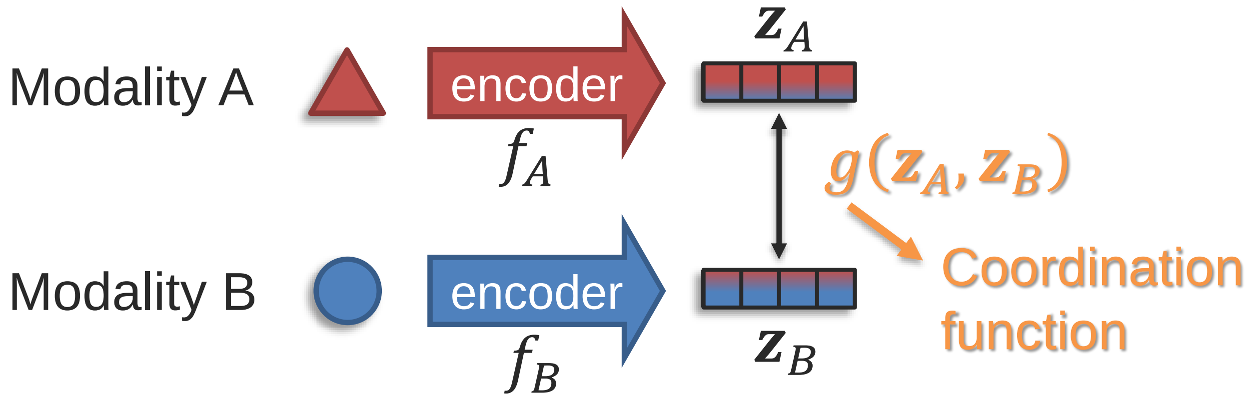
Global Alignment



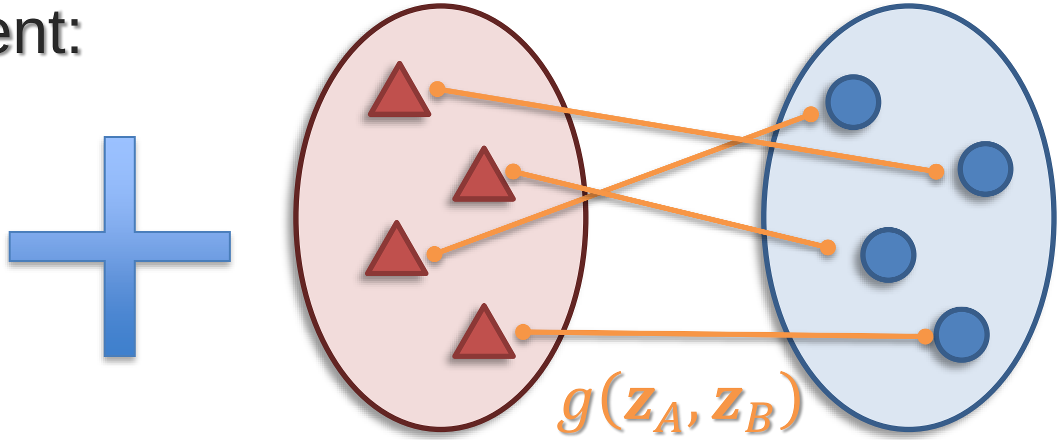
Latent pairing information



Jointly optimize representation + global alignment:



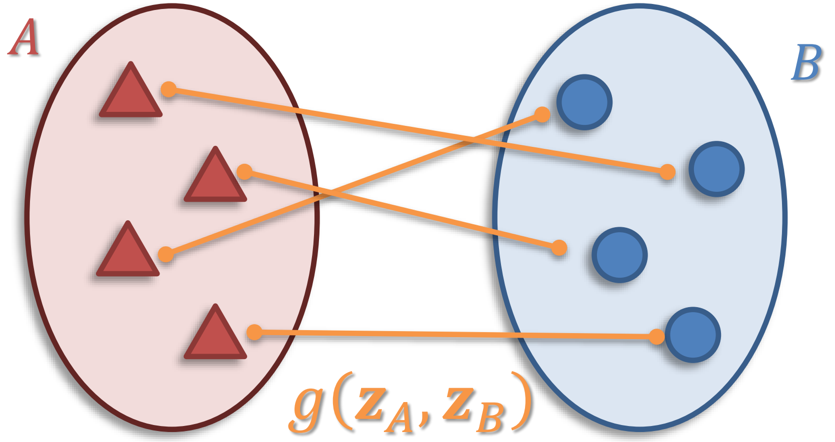
(representation)



(global alignment)

Assignment Problem

Bipartite Graph



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: $\cancel{f: A \rightarrow B}$
(vector of indices)

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

Similarity weights: $\cancel{w_{(i,f(i))} = g(\mathbf{z}_A^i, \mathbf{z}_B^{f(i)})}$

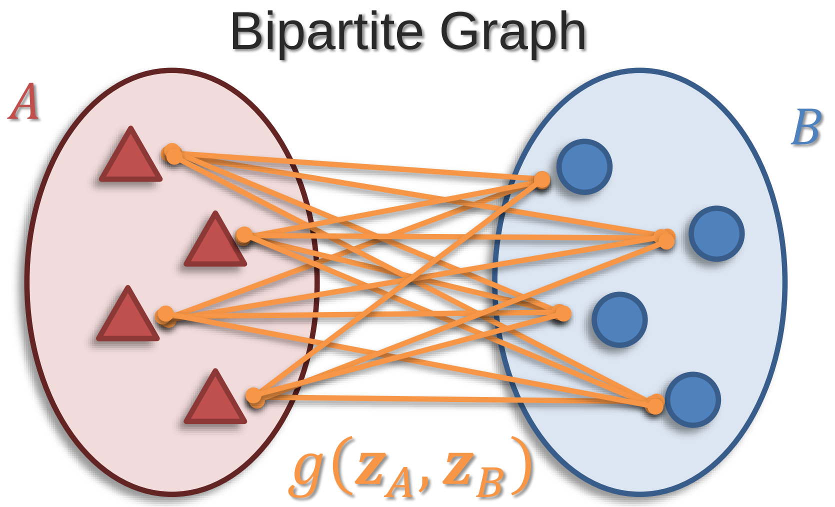
$w_{(i,j)} = g(\mathbf{z}_A^i, \mathbf{z}_B^j)$

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

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

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 $\boldsymbol{z}_B^j$

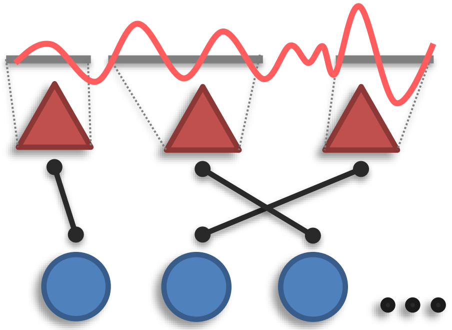
Similarity weights: $w_{(i,j)} = g(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)$

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

Wassertein distance gives optimal transport

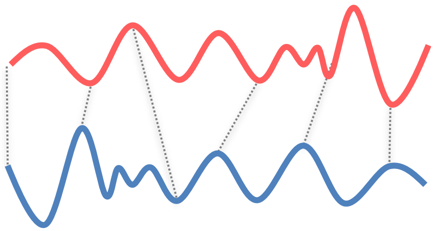
Continuous Alignment

Challenge 2b: Continuous Alignment

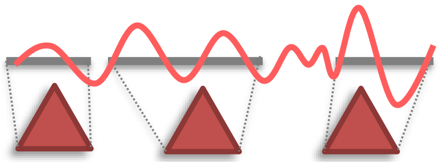


Definition: Model alignment between modalities with continuous signals and no explicit elements

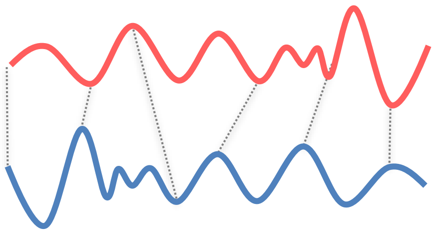
Continuous warping



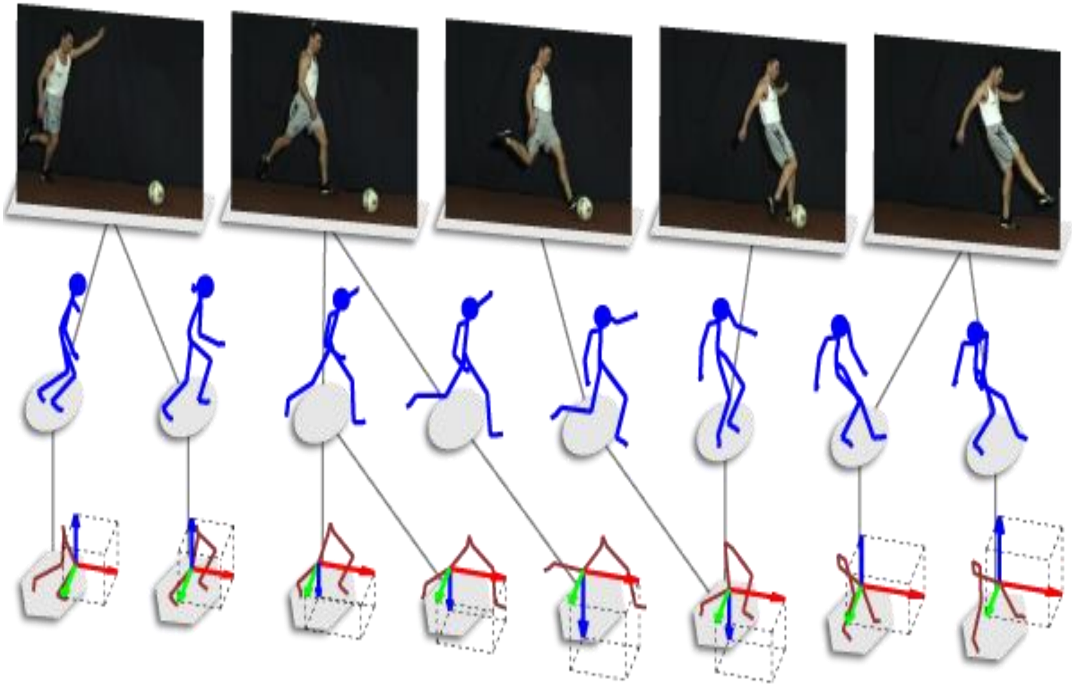
Discretization (segmentation)



Continuous Warping – Example



Aligning video sequences



Dynamic Time Warping (DTW)

We have two unaligned temporal unimodal signals

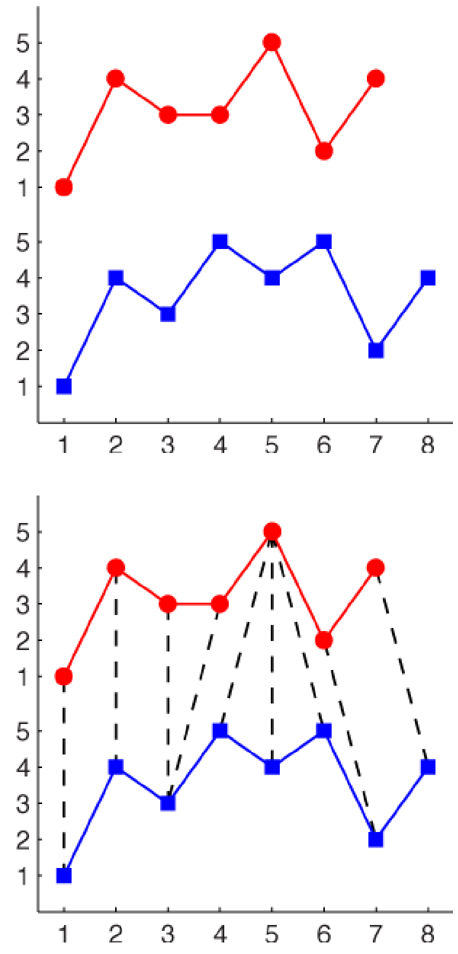
- $\mathbf{X} = \left[\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_{n_x}\right] \in \mathbb{R}^{d \times n_x}$
- $\mathbf{Y} = \left[\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_{n_y}\right] \in \mathbb{R}^{d \times n_y}$

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!

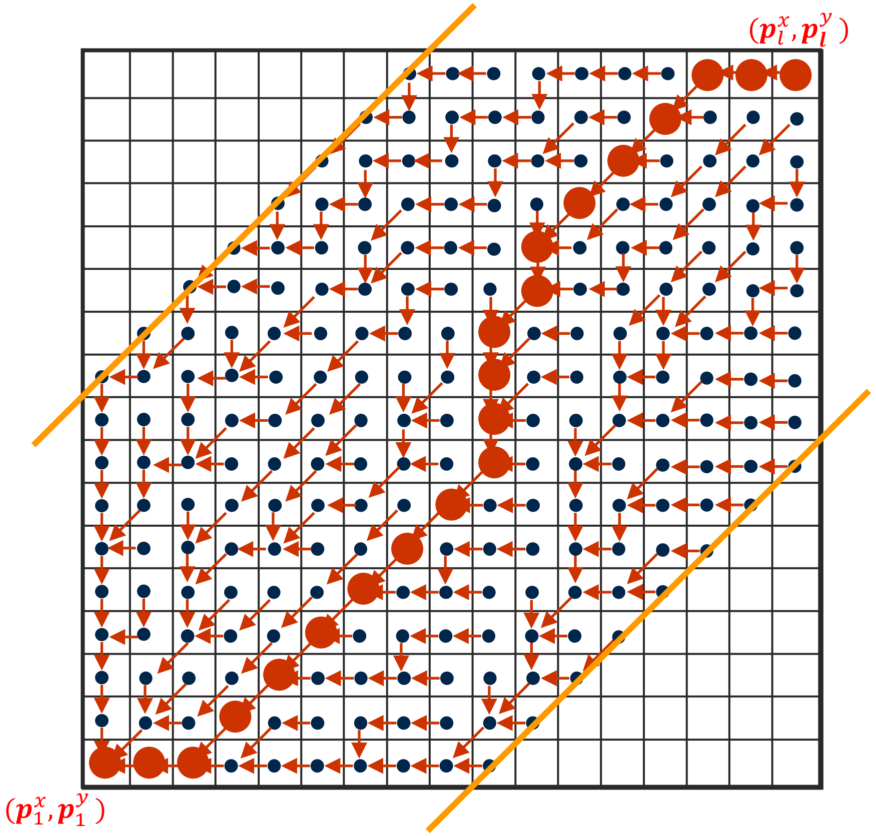


Dynamic Time Warping (DTW)

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

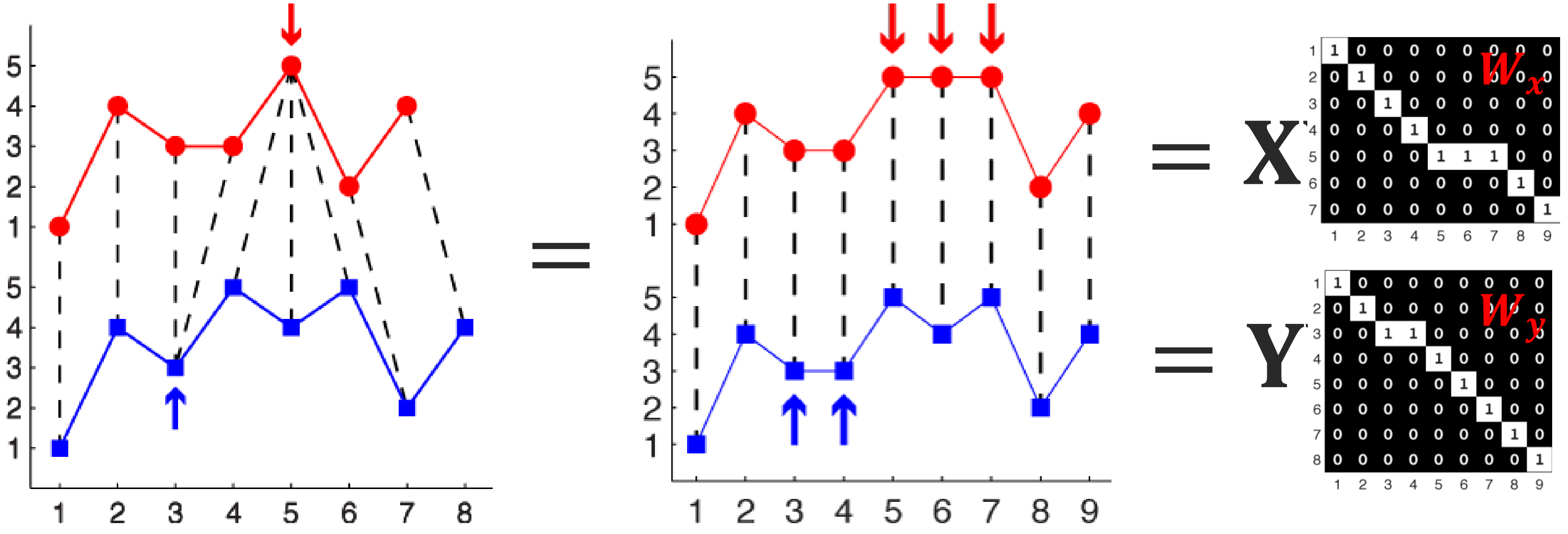
Solved using dynamic programming while respecting the restrictions



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

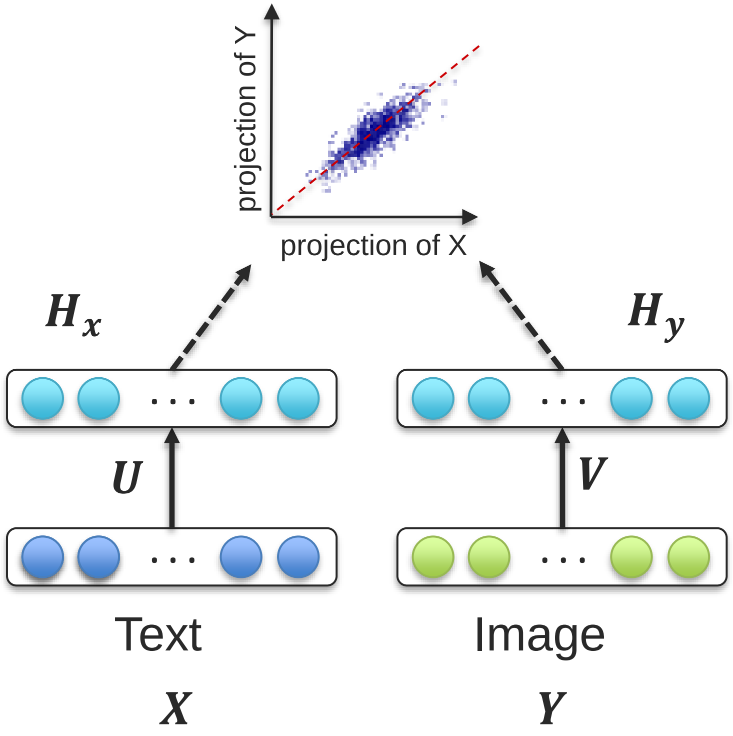
Canonical Correlation Analysis – Reminder

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:

$$\boldsymbol{U}^T\boldsymbol{\Sigma}_{YY}\boldsymbol{U} = \boldsymbol{V}^T\boldsymbol{\Sigma}_{YY}\boldsymbol{V} = \boldsymbol{I},\ \boldsymbol{u}_{(j)}^T\boldsymbol{\Sigma}_{XY}\boldsymbol{v}_{(i)} = \boldsymbol{0}$$



Canonical Time Warping

Dynamic Time Warping + Canonical Correlation Analysis = Canonical Time Warping

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

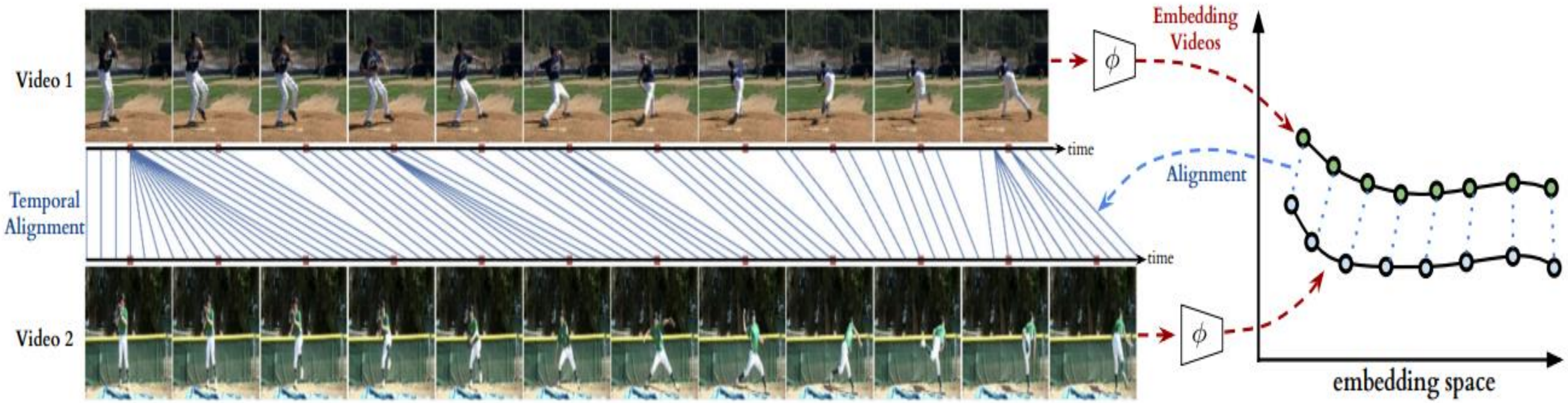
Allows to align multi-modal or multi-view (same modality but from a different point of view)

- $\boldsymbol{W}_x, \boldsymbol{W}_y$ – 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

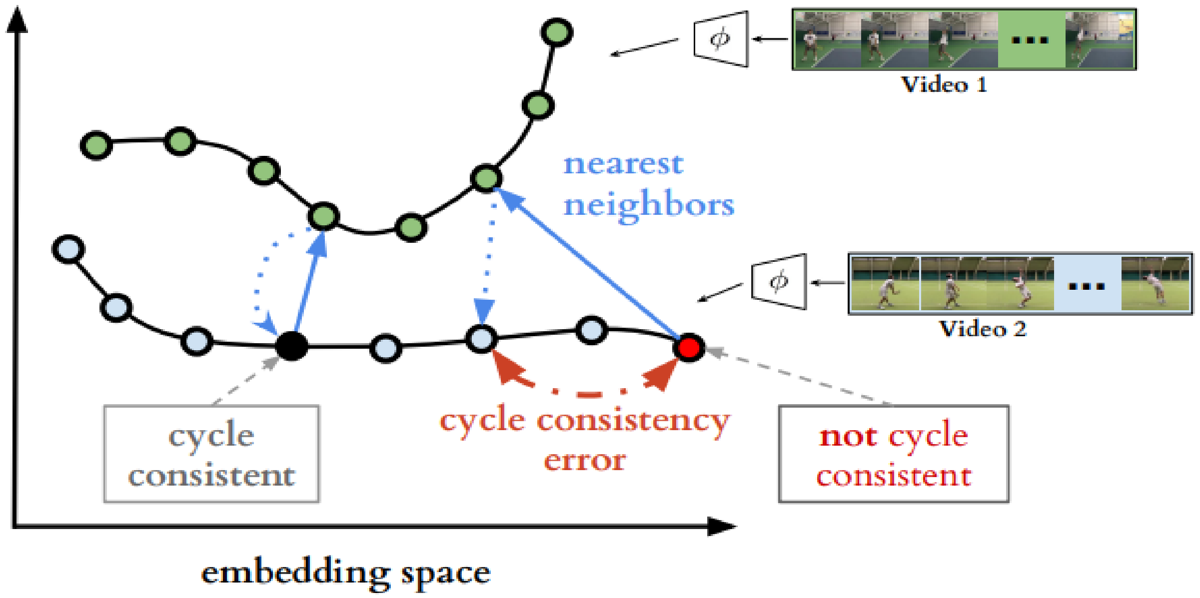
Premise: we have paired video sequences that can be be temporally aligned



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

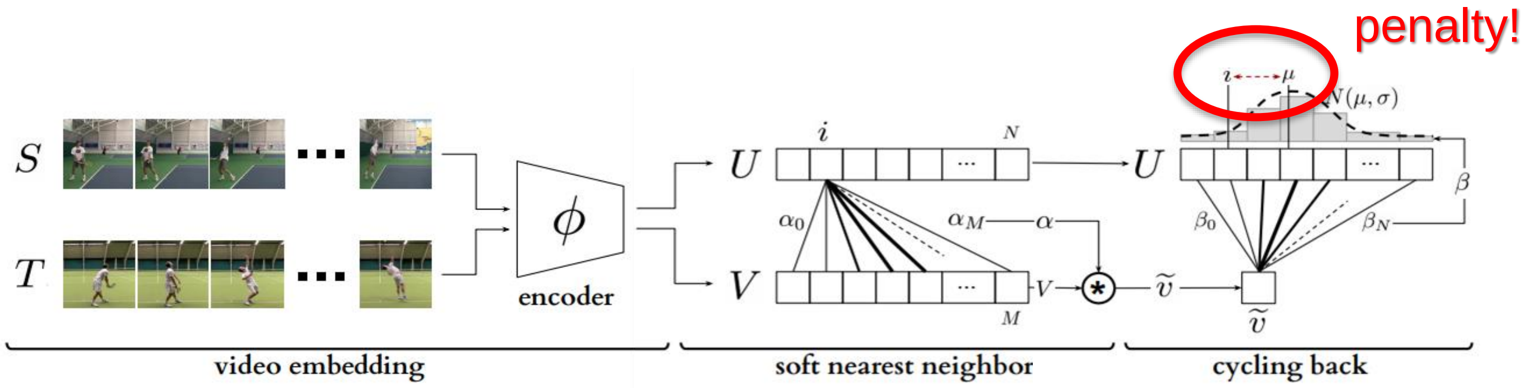
Temporal Cycle-Consistency Learning

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: $$\alpha_j = \frac{e^{-||u_i - v_j||^2}}{\sum_k^M e^{-||u_i - v_k||^2}}$$

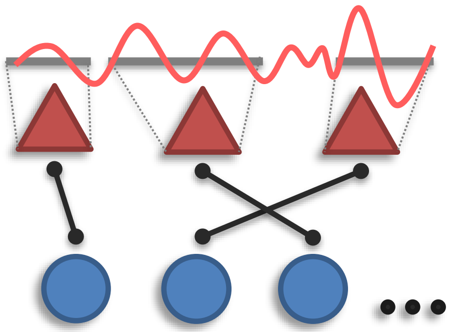
Soft nearest neighbor: $$\widetilde{v} = \sum_j^M \alpha_j v_j,$$

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

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

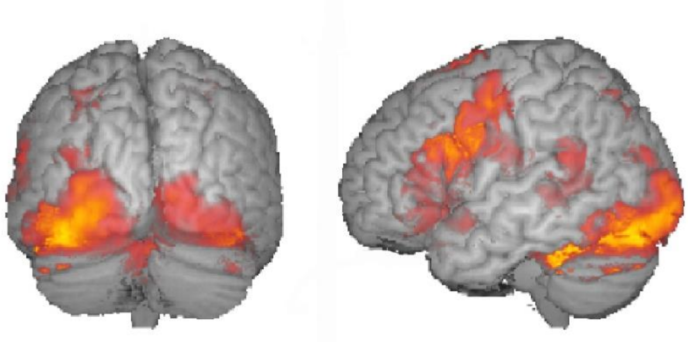
$$L_{cbr} = \frac{|i - \mu|^2}{\sigma^2} + \lambda \log(\sigma)$$

Discretization (aka Segmentation)

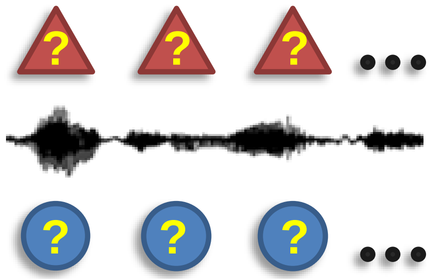


Common assumptions: ① Segmented elements

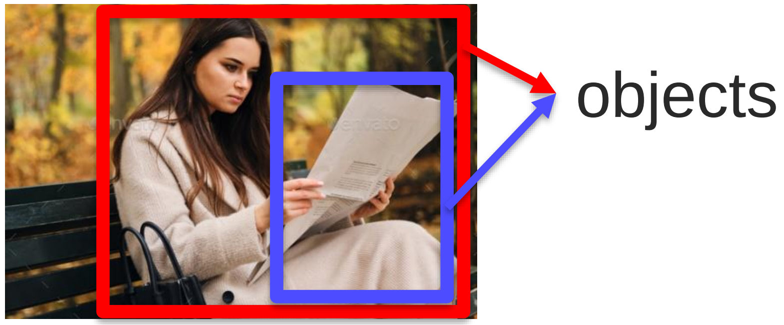
Examples:



Medical imaging



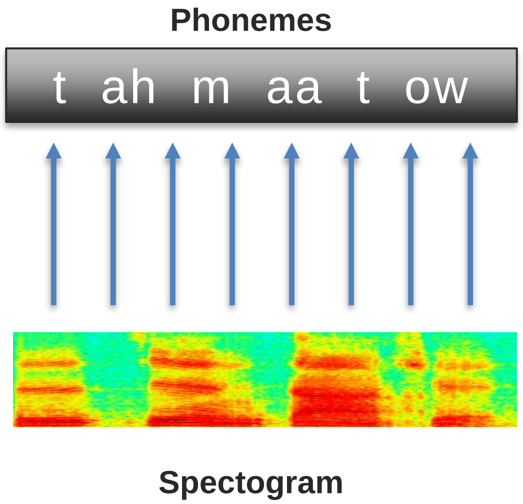
Signals



Images

Discretization – Example

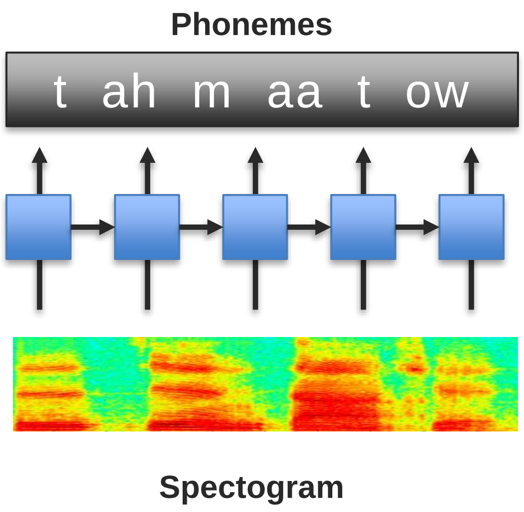
Sequence Labeling and Alignment



How can we predict the sequence of phoneme labels?

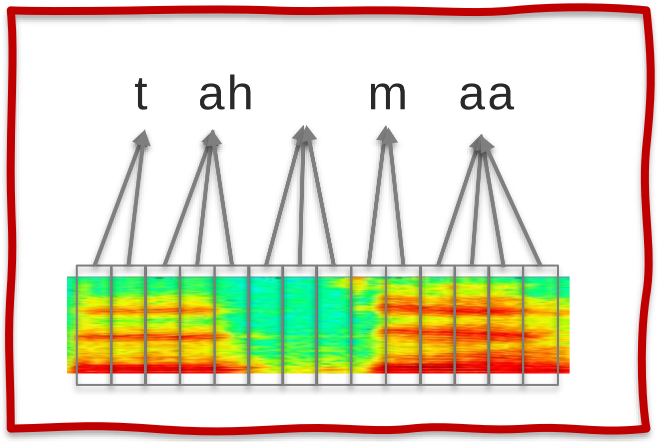
Discretization – Example

Sequence Labeling and Alignment



How can we predict the sequence of phoneme labels?

Challenge: many-to-1 alignment



Discretization – A Classification Approach

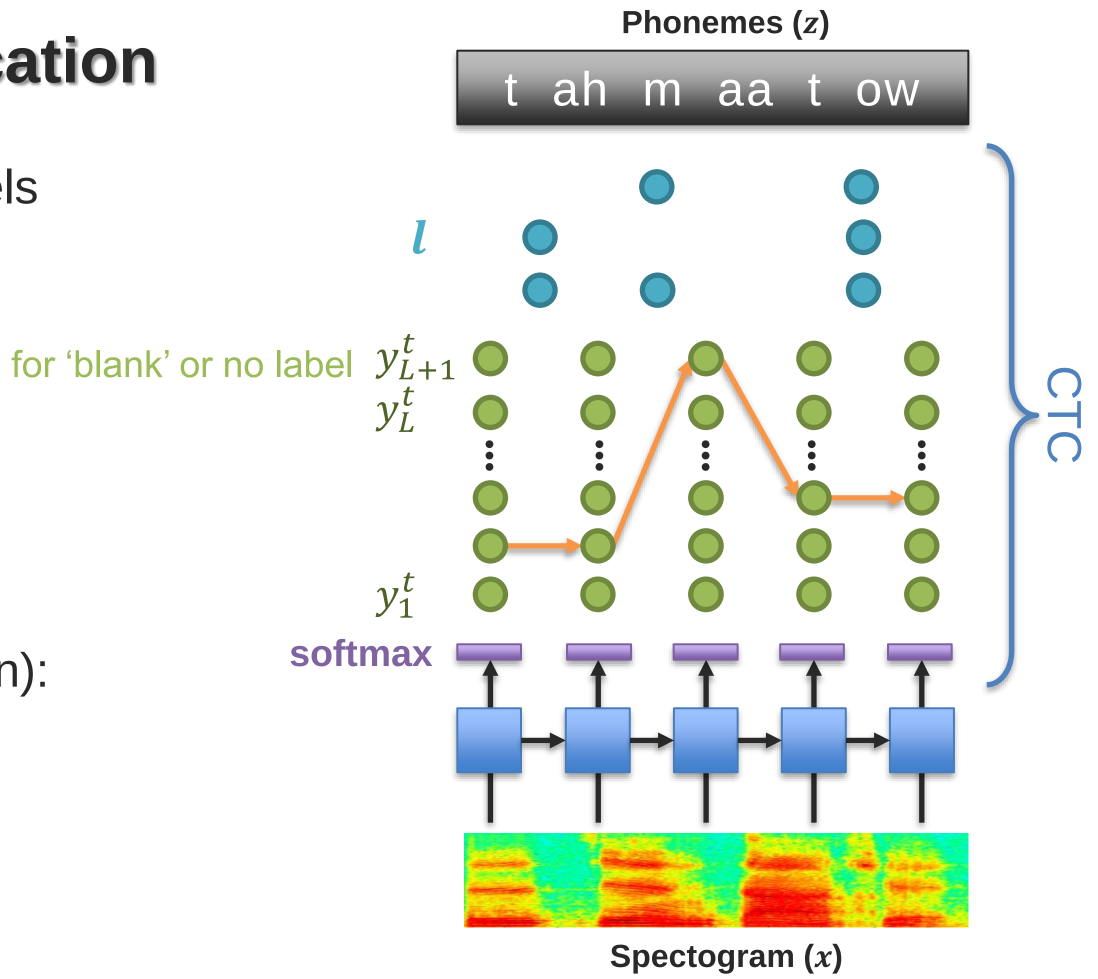
Connectionist Temporal Classification

4 Most probable sequence labels

3 Predicted labels $\boldsymbol{l}$

2 Path $\boldsymbol{\pi}$ over the activations:

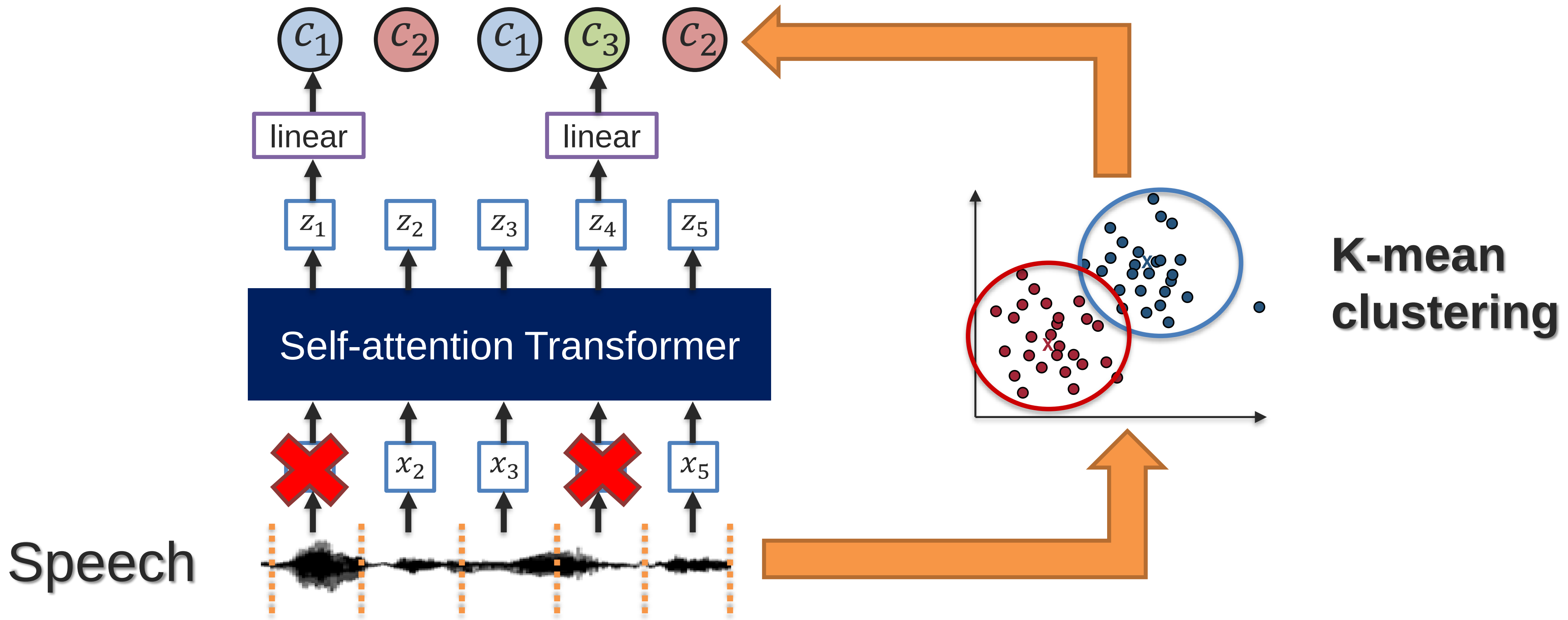
1 Output activations (distribution):



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

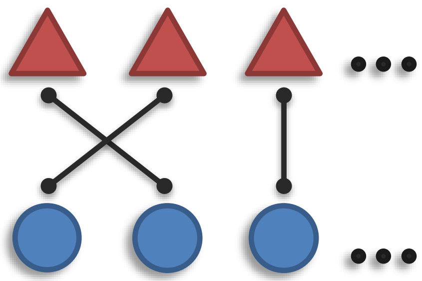
Challenge 2: Alignment

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

Sub-challenges:

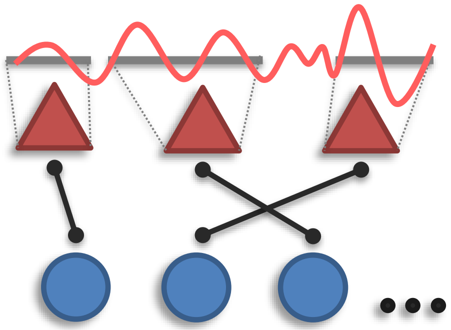
Thursday!

Discrete Alignment



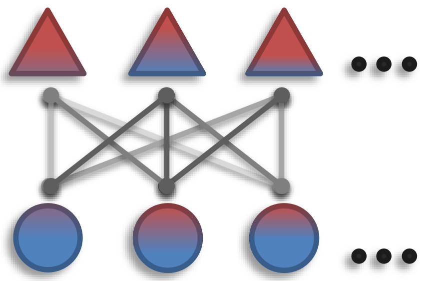
Discrete elements and connections

Continuous Alignment



Segmentation and continuous warping

Contextualized Representation



Alignment + representation