Multimodal Machine Learning

Lecture 4.2: 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

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 1 8/30 & 9/1	**Course introduction** • Multimodal core challenges • Course syllabus	**Multimodal applications and datasets** • Research tasks and datasets • Team projects
Week 2 9/6 & 9/8 Read due: 9/9	**Basic concepts: neural networks** • Loss functions and neural networks • Gradient and optimization	**Unimodal representations** • Dimensions of heterogeneity • Visual representations
Week 3 9/13 & 9/15 Read due: 9/16 Proj. Due: 9/14	**Unimodal representations** • Language representations • Signals, graphs and other modalities	**Multimodal representations** • Cross-modal interactions • Multimodal fusion
Week 4 9/20 & 9/22 *Proj. due: 9/25*	**Multimodal representations** • Coordinated representations • Multimodal fission	**Multimodal alignment** • Explicit alignment • Multimodal grounding
Week 5 9/27 & 9/29 Read due: 9/30	***Project hours (Research ideas)***	**Aligned representations** • Self-attention transformer models • Masking and self-supervised learning
Week 6 10/4 & 10/6 *Proj. due: 10/9*	**Multimodal aligned representations** • Multimodal transformers • Video and graph representations	**Multimodal Reasoning** • Structured and hierarchical mo • Memory models

First assignment due on Sunday 9/25

Second assignment due on Sunday 10/9

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 7 10/11 & 10/13 Read due: 10/14	**Multimodal Reasoning** • Reinforcement learning • Discrete structure learning	**Multimodal Reasoning** • Logical and causal inference • External knowledge
Week 8 10/18 & 10/20	***Fall Break – No lectures***	
Week 9 10/25 & 10/27 *Proj. due: 10/30*	**Generation** • Translation, summarization, creation • Generative models: VAEs	**Generation** • GANs and diffusion models • Model evaluation and ethics
Week 10 11/1 & 11/3	***Project presentations (midterm)***	***Project presentations (midterm)***
Week 11 11/8 & 11/10 Read due: 11/12	**Transference** • Modality transfer • Multimodal co-learning	**Quantification** • Heterogeneity and interactions • Biases and fairness
Week 12 11/15 & 11/17 Read due: 11/21	***Project hours (Research ideas)***	**New research directions** • Recent approaches in multimodal ML

Midterm assignment due on Sunday 10/30

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 13 11/22 & 11/24	***Thanksgiving Week – No Class –***	
Week 14 11/30 & 12/2	***Language, Vision, and Actions*** • Robots, navigation and embodied AI • Guest lecturer: Yonatan Bisk	***Multimodal Language Grounding*** • Grounded semantics and pragmatics • Guest lecturer: Daniel Fried
Week 15 12/6 & 12/8 *Proj. due: 12/11*	***Project presentations (final)***	***Project presentations (final)***

Final assignment due on Sunday 12/11

Team Meetings with Instructor

Sign-up deadline: Sunday 9/25 at 11pm

- No lecture on Tuesday 9/27
- 15-mins meeting with instructor
  - Optional, but highly suggested
  - Not all teammates are required to attend
  - Prepare 2 slides to summarize your research ideas
- Meetings on Tuesday 9/27 and Wednesday 9/28
- Signup form:

  https://calendly.com/morency/student-meetings

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 4.2: Multimodal alignment

Louis-Philippe Morency

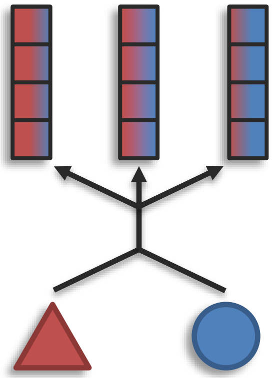
** Co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk.*

Lecture objectives

- Fine-grained fission
  - Cluster-based approach
- 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

Fine-Grained Fission

Sub-Challenge 1c: Representation Fission



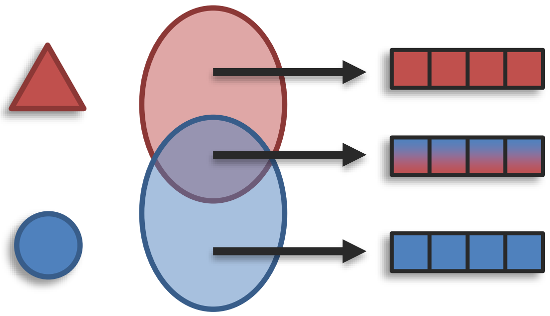
Definition: learning a new set of representations that reflects multimodal internal structure such as data factorization or clustering

How to automatically discover these internal clusters, factors?

Modality-level fission:

Modality A

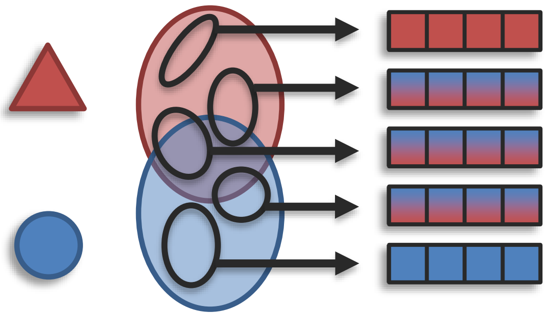
Modality B



Fine-grained fission:

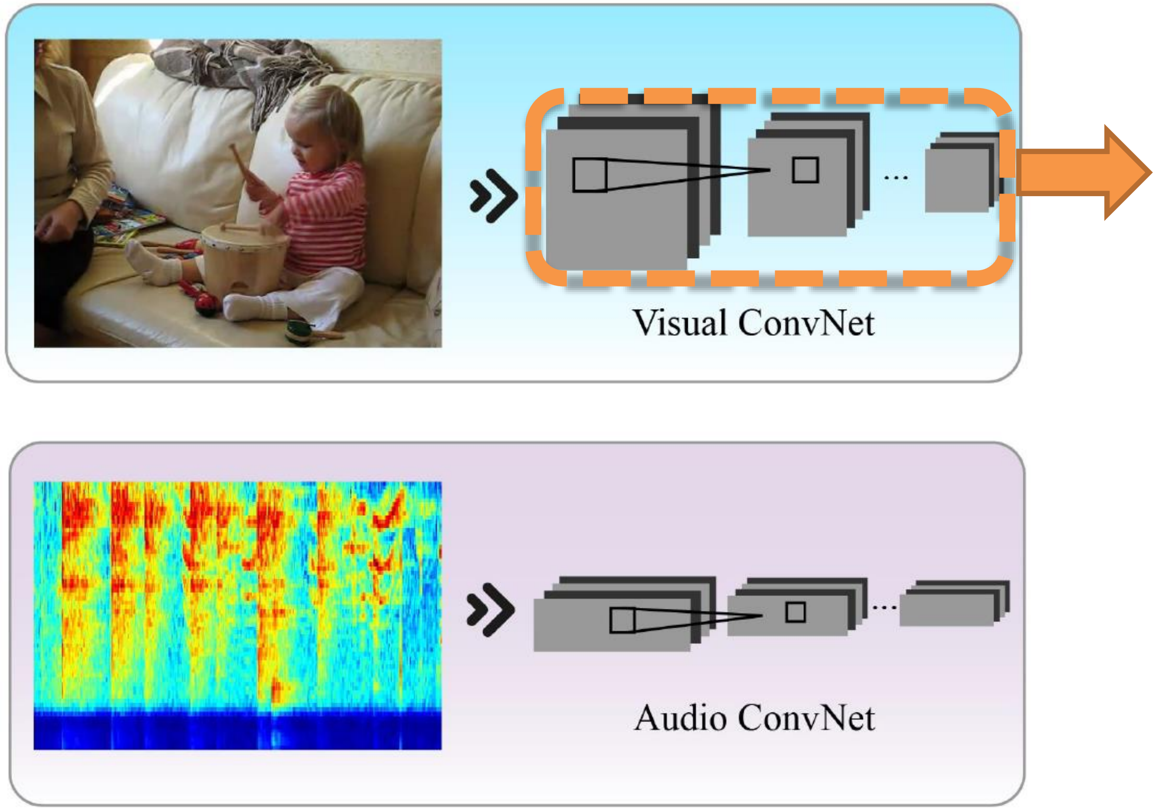
Modality A

Modality B

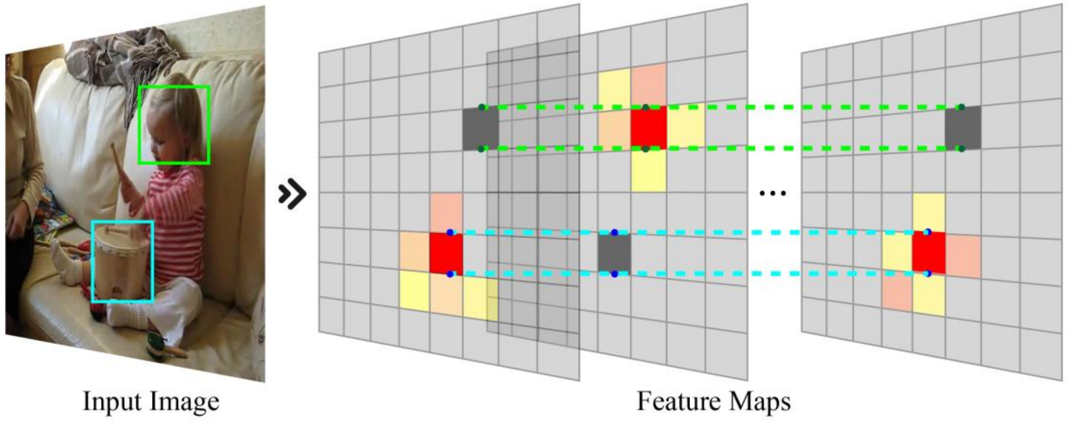


Fine-Grained Fission – A Clustering Approach

Unimodal Encoders

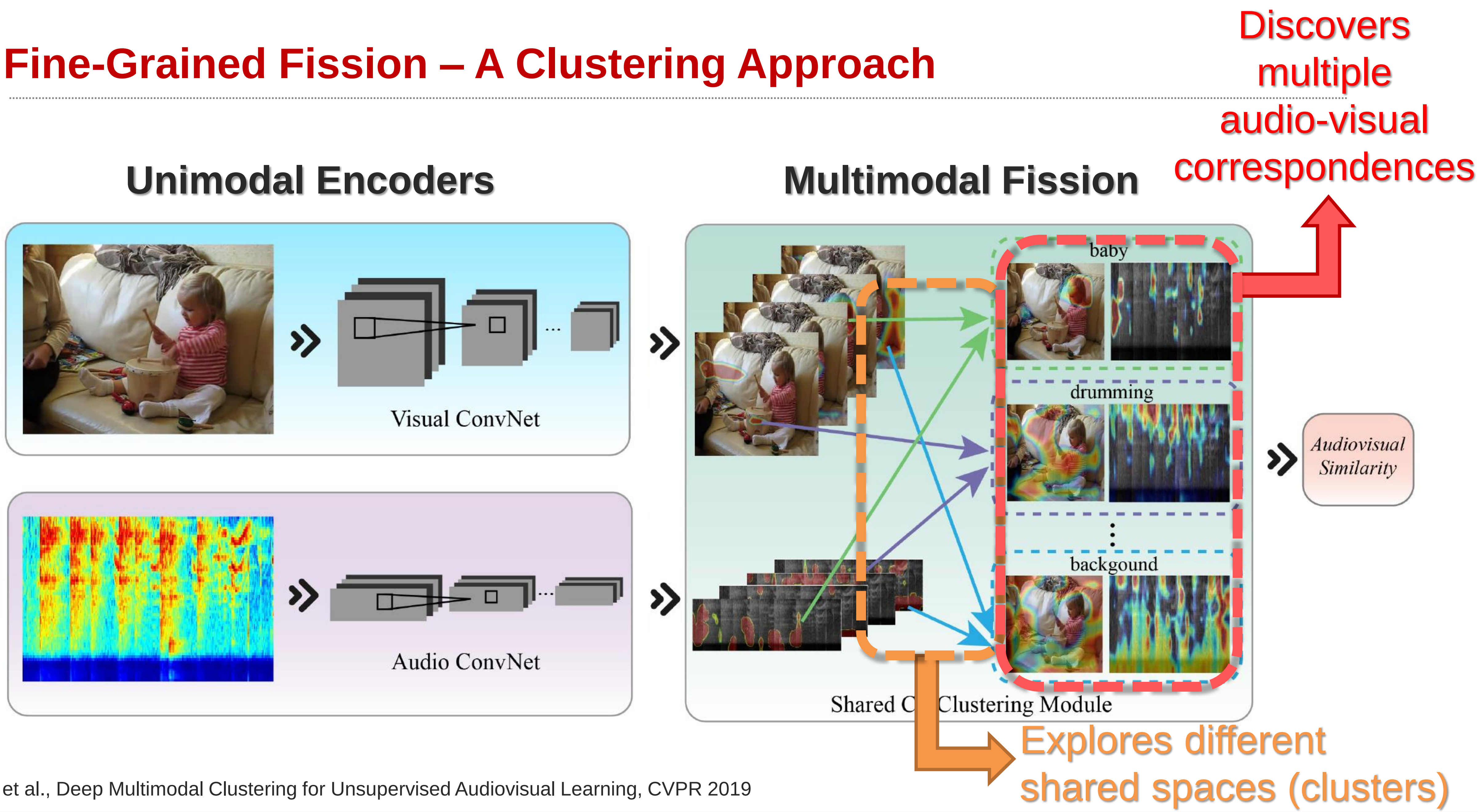


Localized activations for different objects



Hu et al., Deep Multimodal Clustering for Unsupervised Audiovisual Learning, CVPR 2019

Fine-Grained Fission – A Clustering Approach



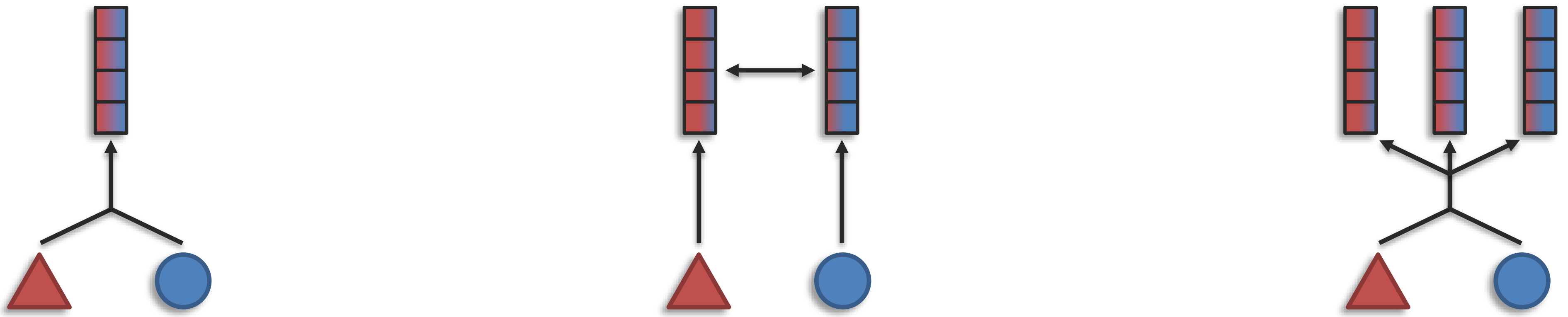
Hu et al., Deep Multimodal Clustering for Unsupervised Audiovisual Learning, CVPR 2019

Challenge 1: Representation

Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

Sub-challenges:

Fusion **Coordination** **Fission**



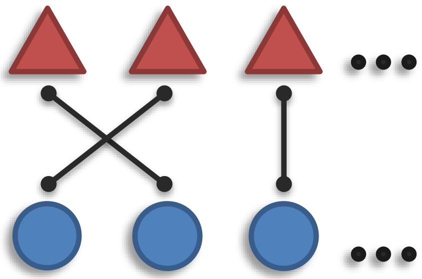
Challenge 2: Alignment

Challenge 2: Alignment

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

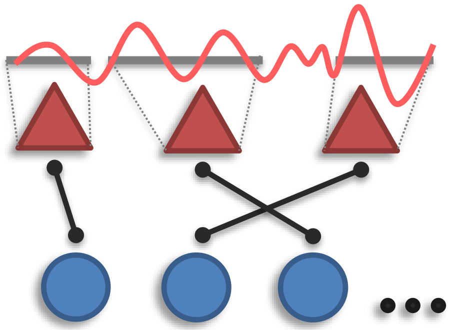
Sub-challenges:

Discrete Alignment



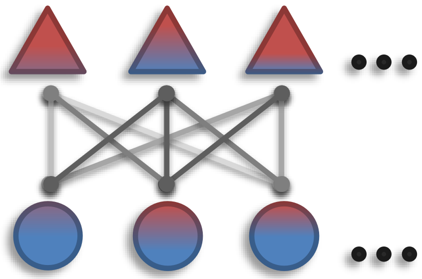
Discrete elements and connections

Continuous Alignment



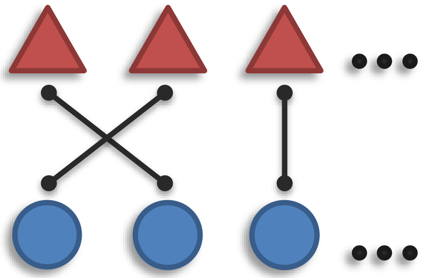
Segmentation and continuous warping

Contextualized Representation

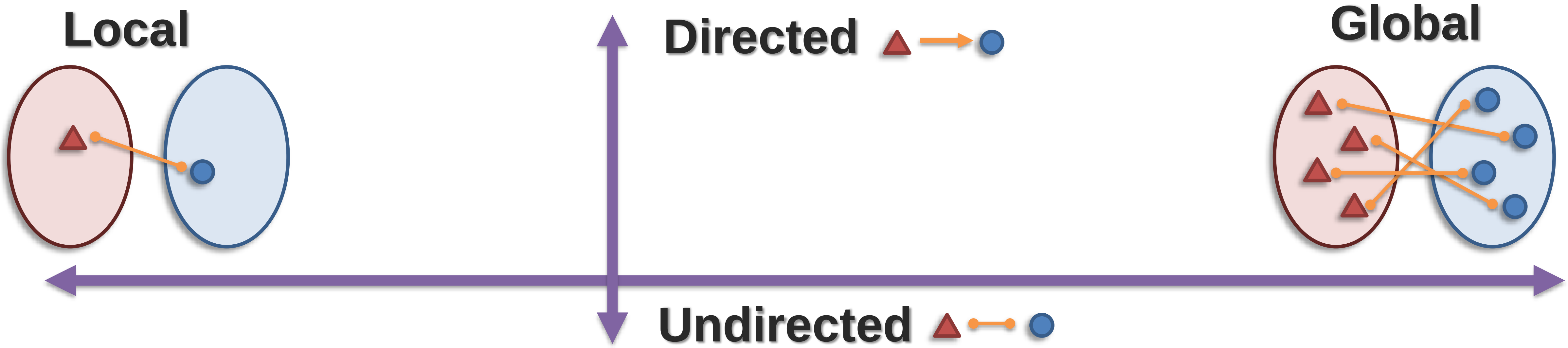


Alignment + representation

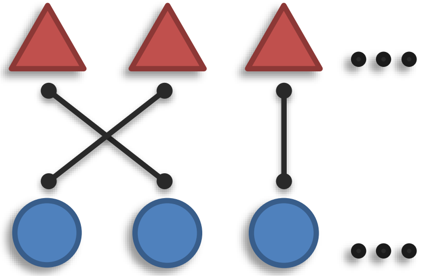
Sub-Challenge 2a: Discrete Alignment



Definition: Identify and model connections between elements of multiple modalities



Connections



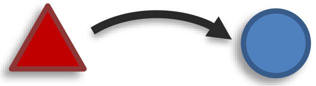
Why should 2 elements be connected?

Statistical

Association

e.g., correlation,
co-occurrence

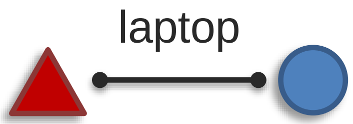
Dependency



e.g., causal,
temporal

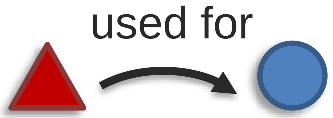
Semantic

Correspondence



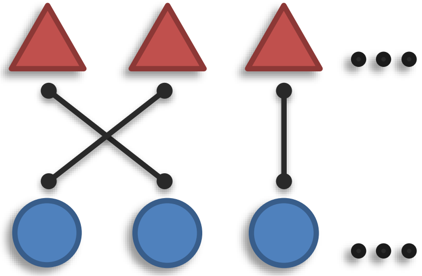
e.g., grounding

Relationship



e.g., function

Connections



Why should 2 elements be connected?

Relationships and Dependencies will be discussed in more details in **Reasoning** challenge

Statistical

Association

=

e.g., correlation, co-occurrence

Dependency

e.g., causal, temporal

Semantic

Correspondence

laptop

e.g., grounding

Relationship

used for

e.g., function

Language Grounding

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

A woman reading newspaper

Statistical

Association

=

e.g., correlation, co-occurrence

Dependency

e.g., causal, temporal

Semantic

Correspondence

laptop

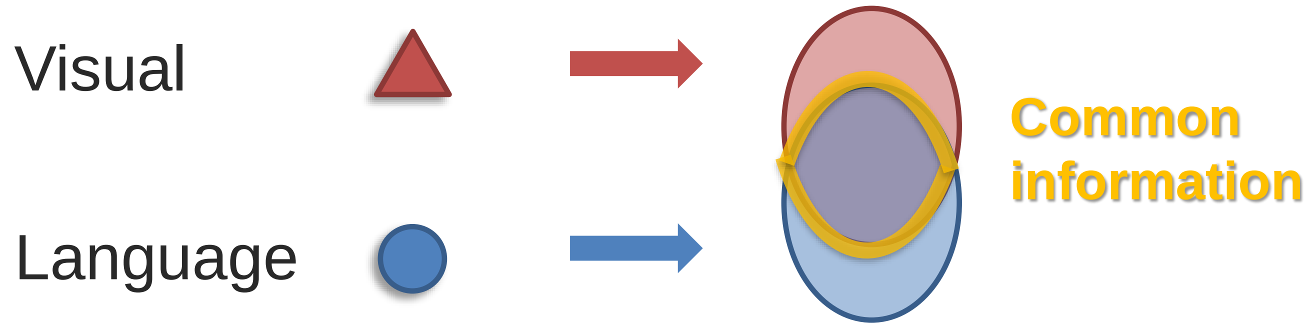
e.g., grounding

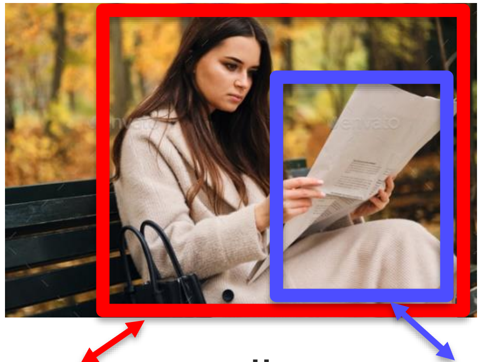
Relationship

used for

e.g., function

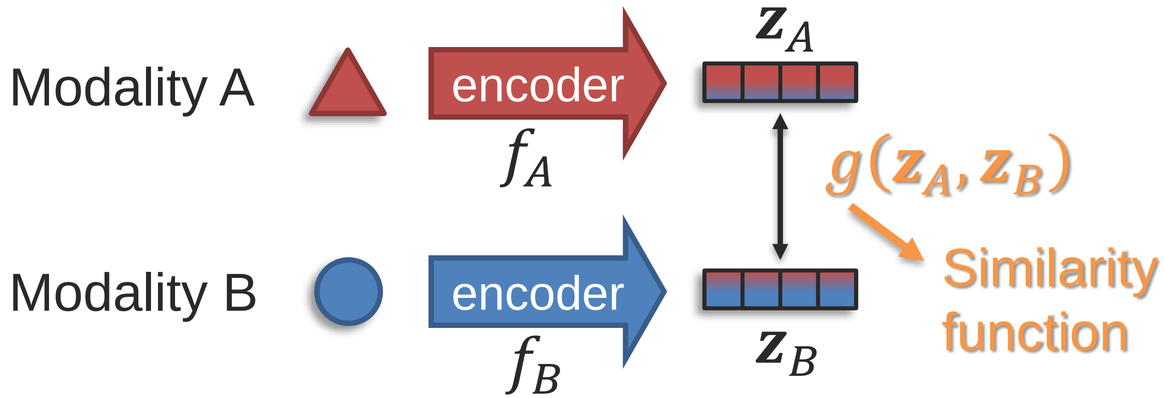
Local Alignment – Coordinated Representations





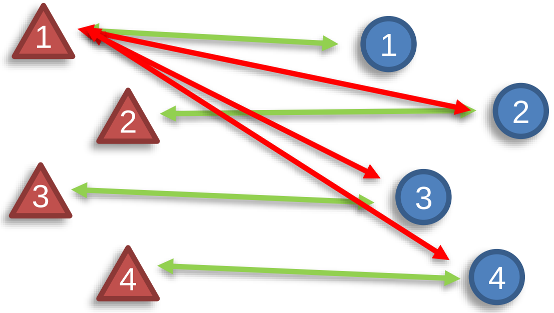
A woman reading newspaper

Learning coordinated representations:



or contrastive learning

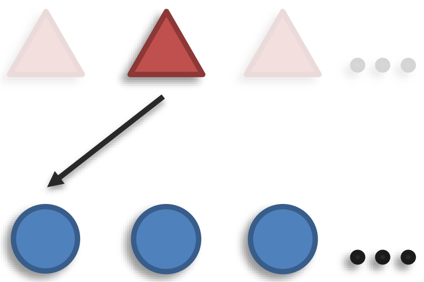
Supervision: Paired data



Directed Alignment

Modality A
(query)

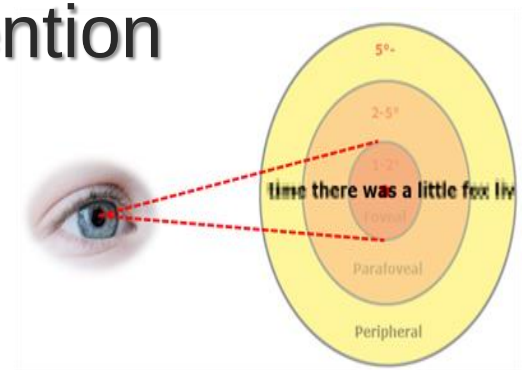
Modality B
(key)



A woman is throwing a frisbee

Which object?

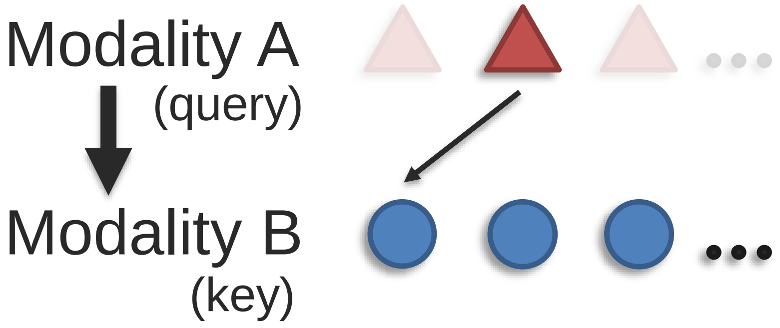
Attention

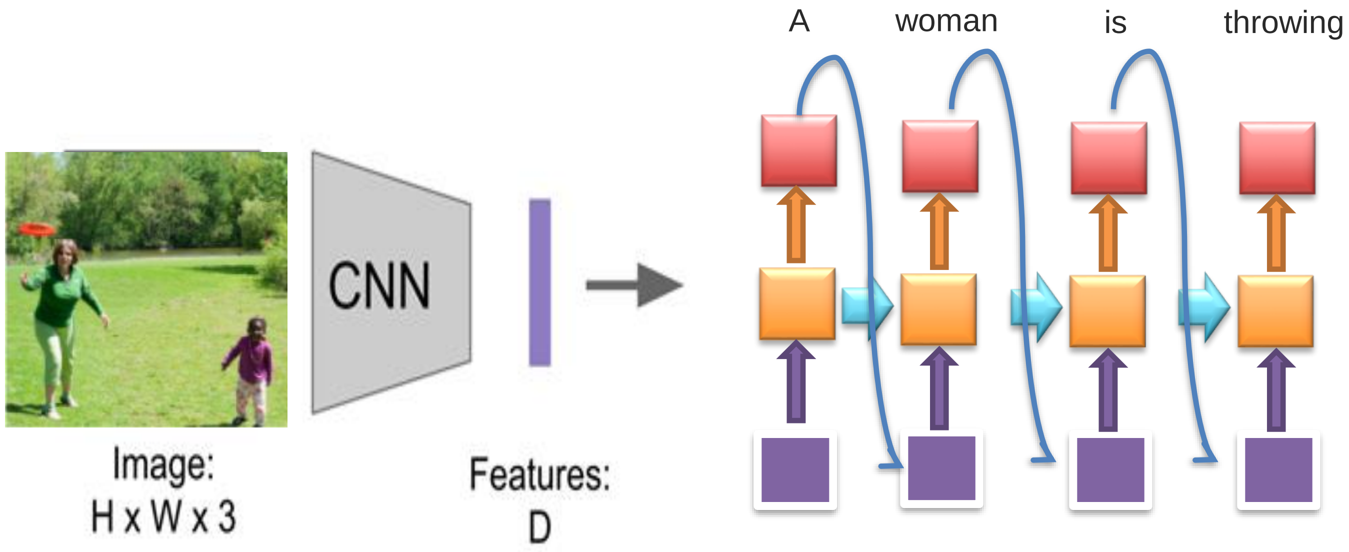


1 Soft attention

2 Hard attention

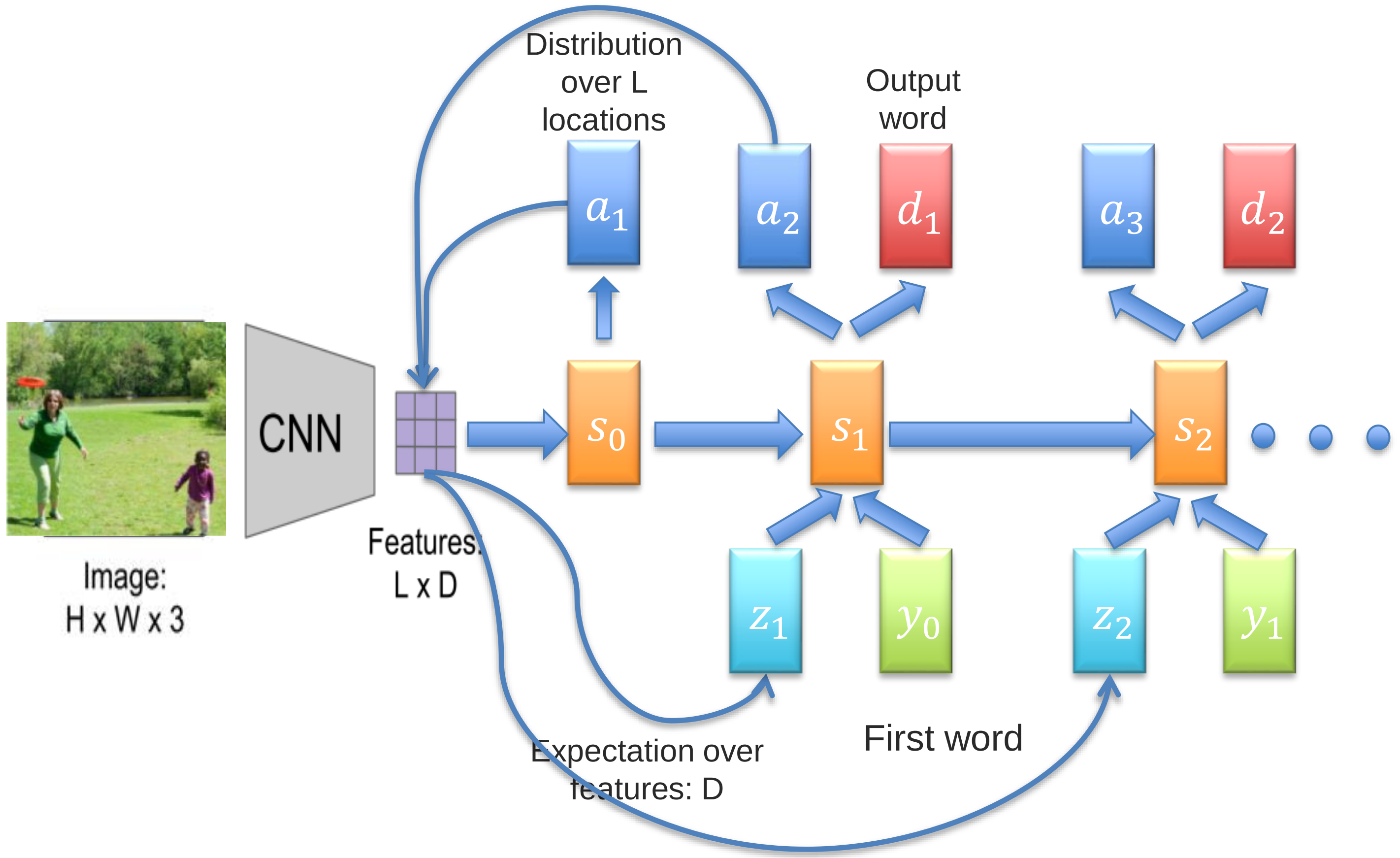
Directed Alignment – Image Captioning





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:

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

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

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

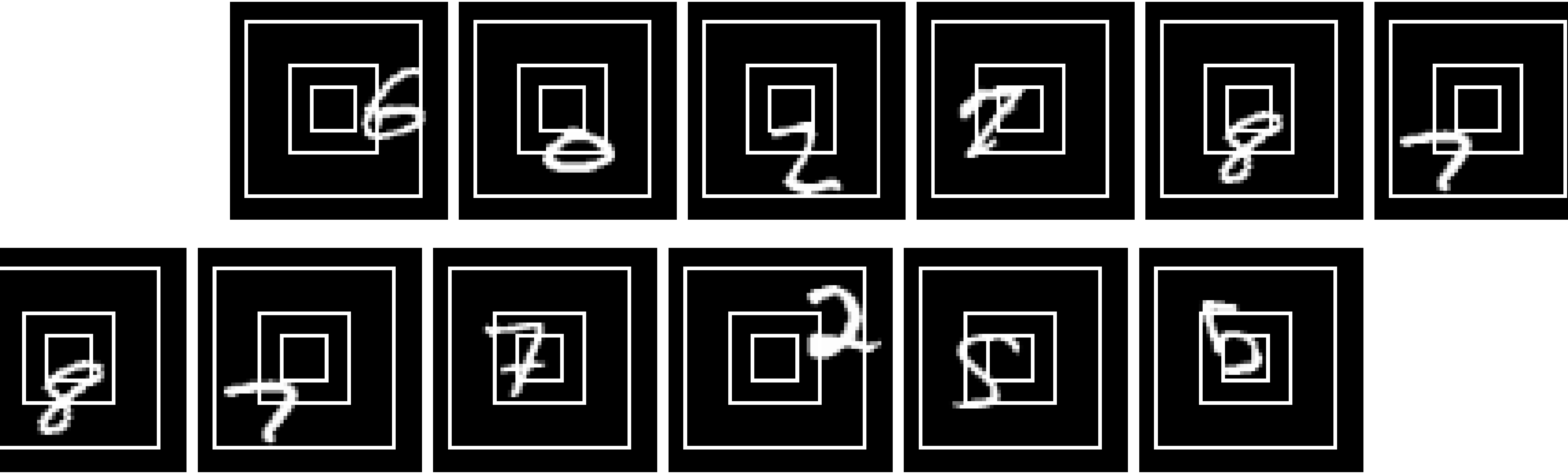
$$z_i = \sum_{j=i}^{T_x} \alpha_{ij} h_j$$

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

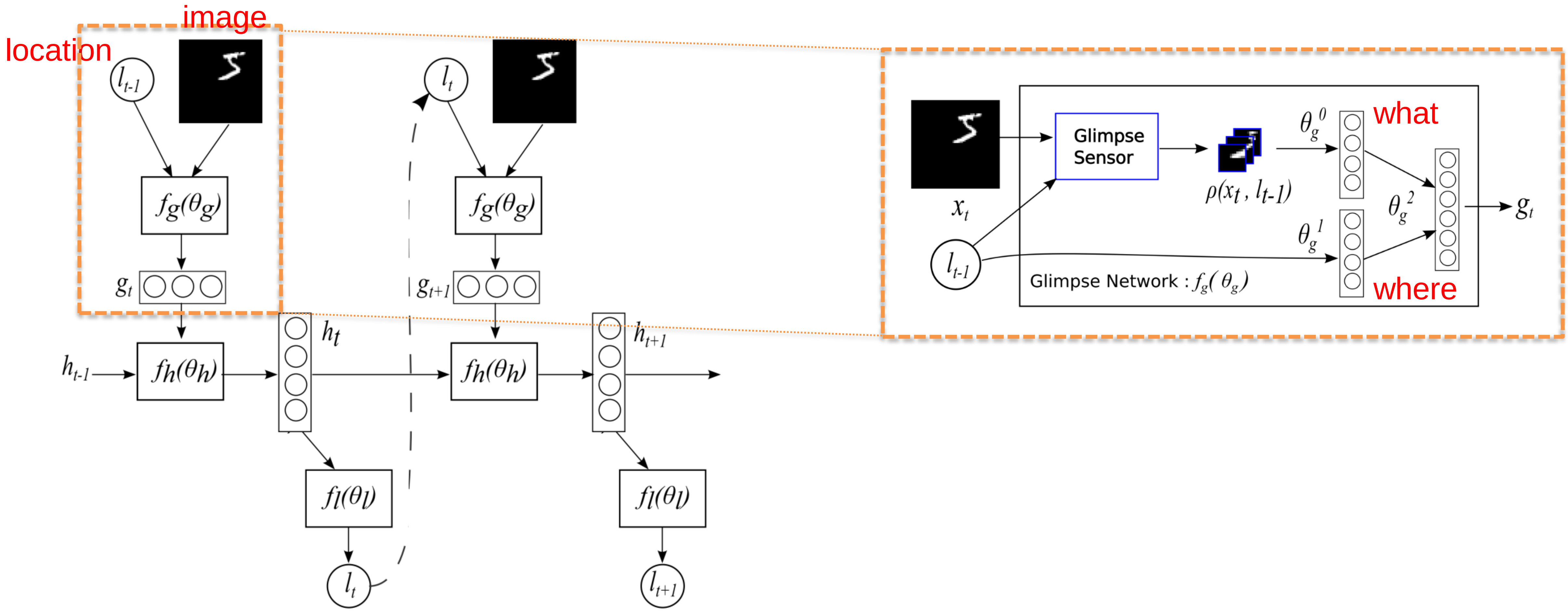
Example – Image Captioning

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

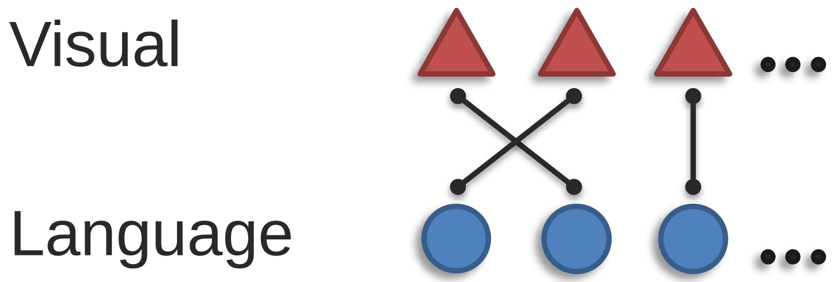
Hard attention - Example



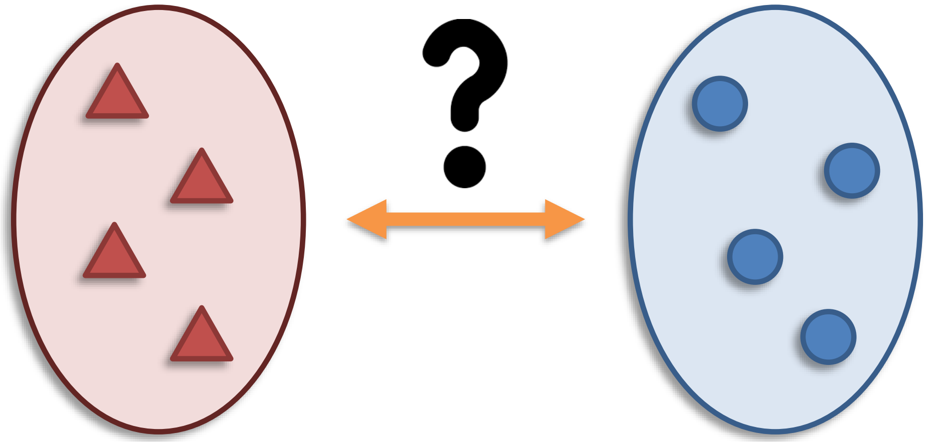
Hard Attention – Recurrent Model of Visual Attention



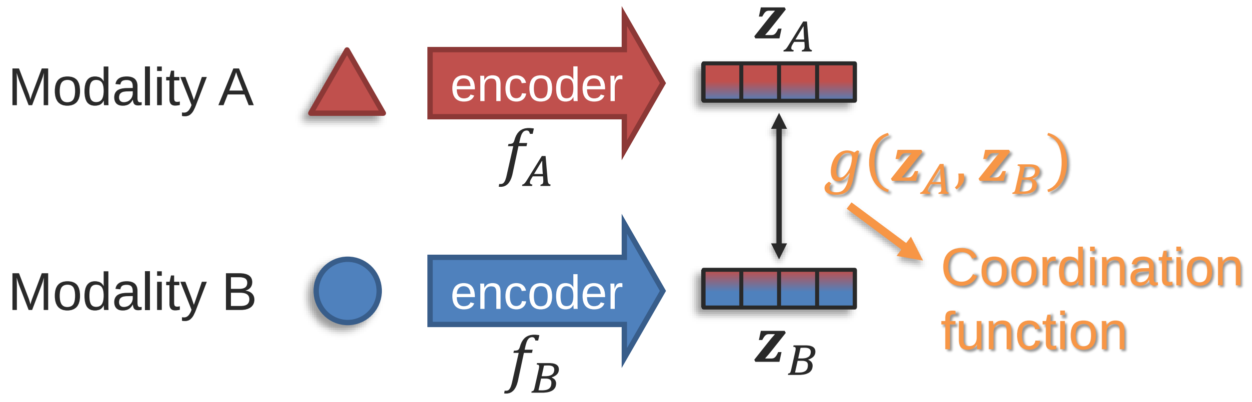
Global Alignment



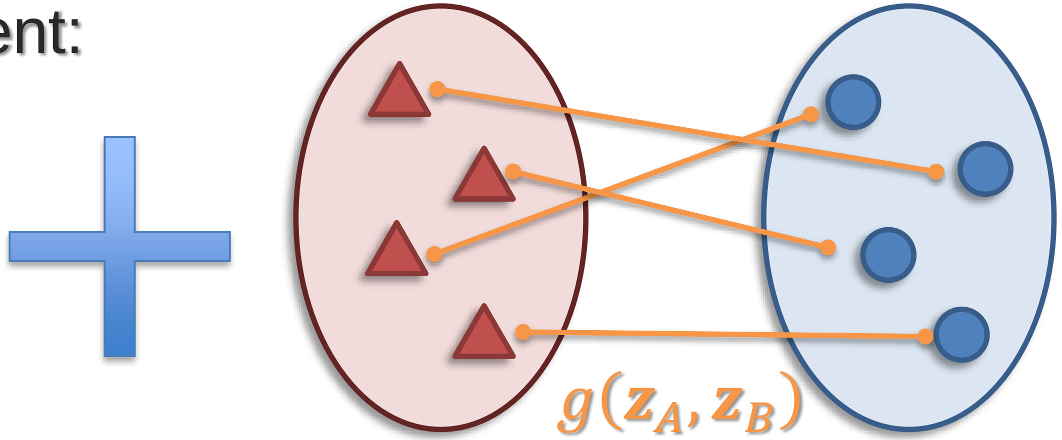
Latent pairing information



Jointly optimize representation + global alignment:



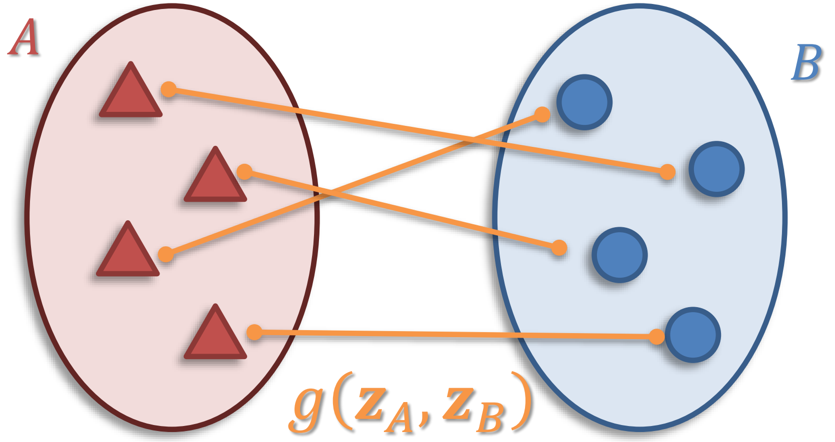
(representation)



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

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

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

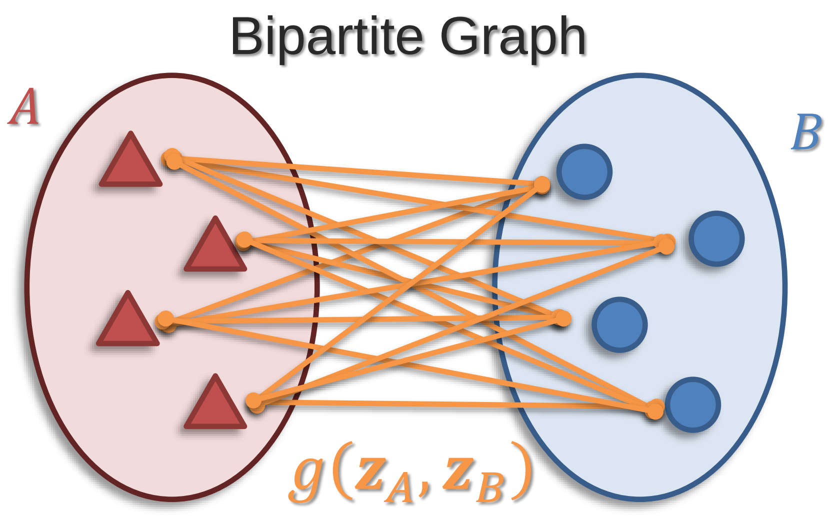
$w_{(i,j)} = g(\mathbf{z}_A^i, \mathbf{z}_B^j)$

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 $\mathbf{z}_A^i$ and $\mathbf{z}_B^j$

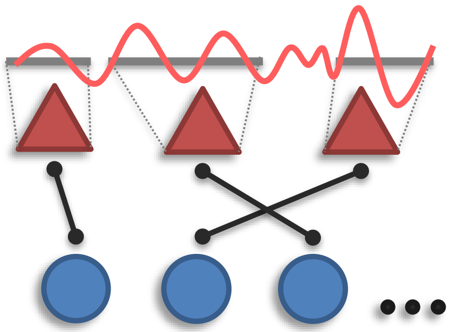
Similarity weights: $w_{(i,j)} = g(\mathbf{z}_A^i, \mathbf{z}_B^j)$

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

➡ Wassertein distance give optimal transport

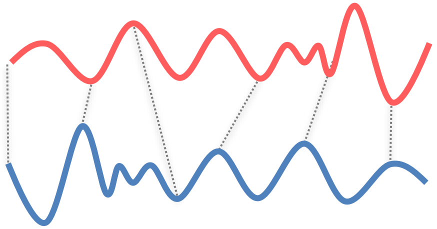
Continuous Alignment

Challenge 2b: Continuous Alignment

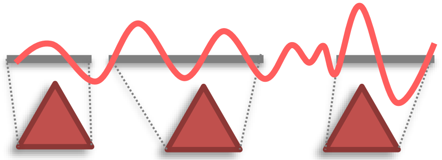


Definition: Model alignment between modalities with continuous signals and no explicit elements

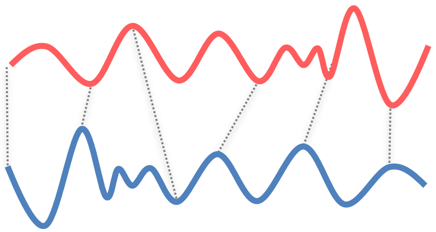
Continuous warping



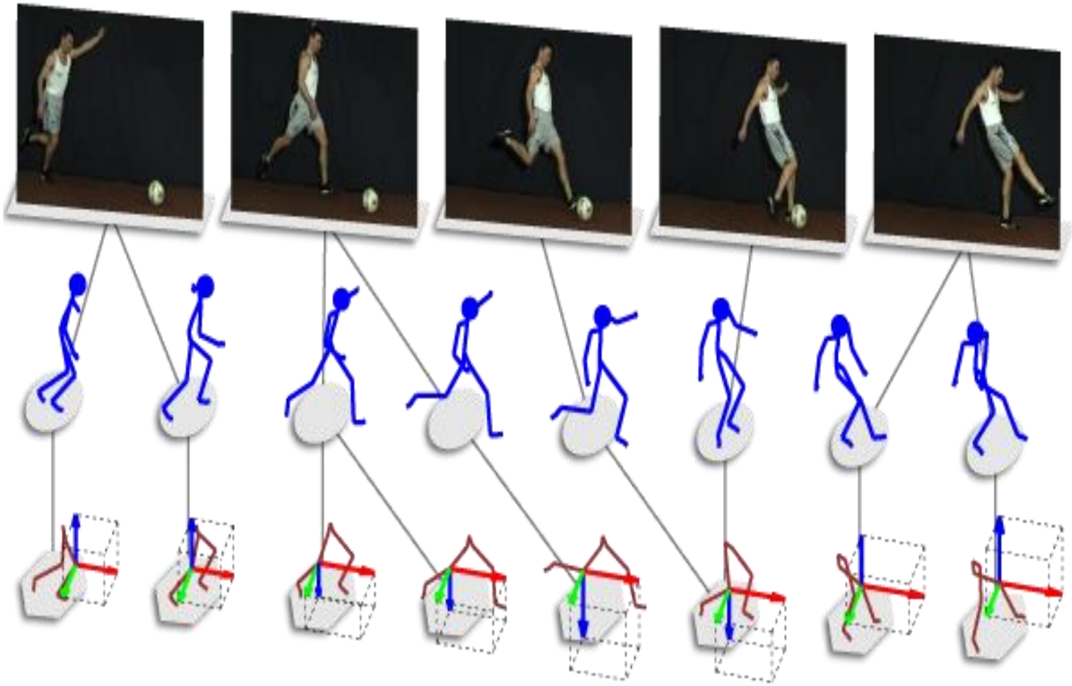
Discretization (segmentation)



Continuous Warping – Example



Aligning video sequences



Dynamic Time Warping (DTW)

We have two unaligned temporal unimodal signals

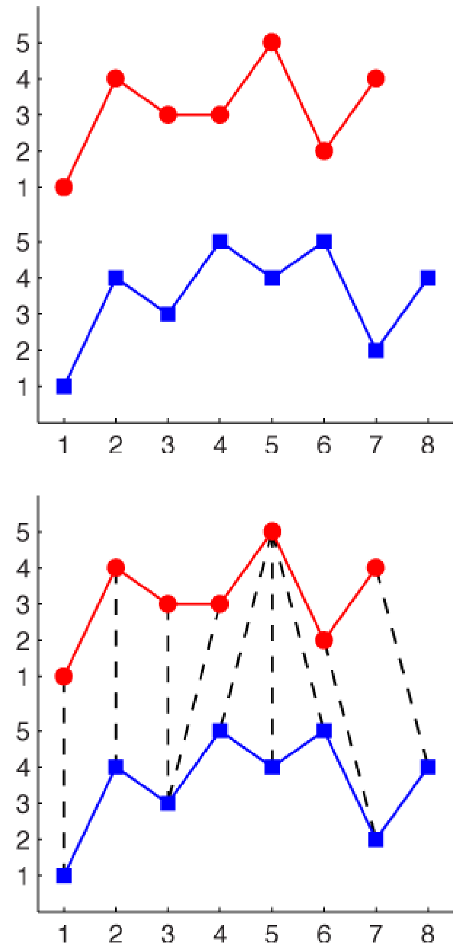
- $\mathbf{X} = \left[\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_{n_x}\right] \in \mathbb{R}^{d \times n_x}$
- $\mathbf{Y} = \left[\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_{n_y}\right] \in \mathbb{R}^{d \times n_y}$

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!

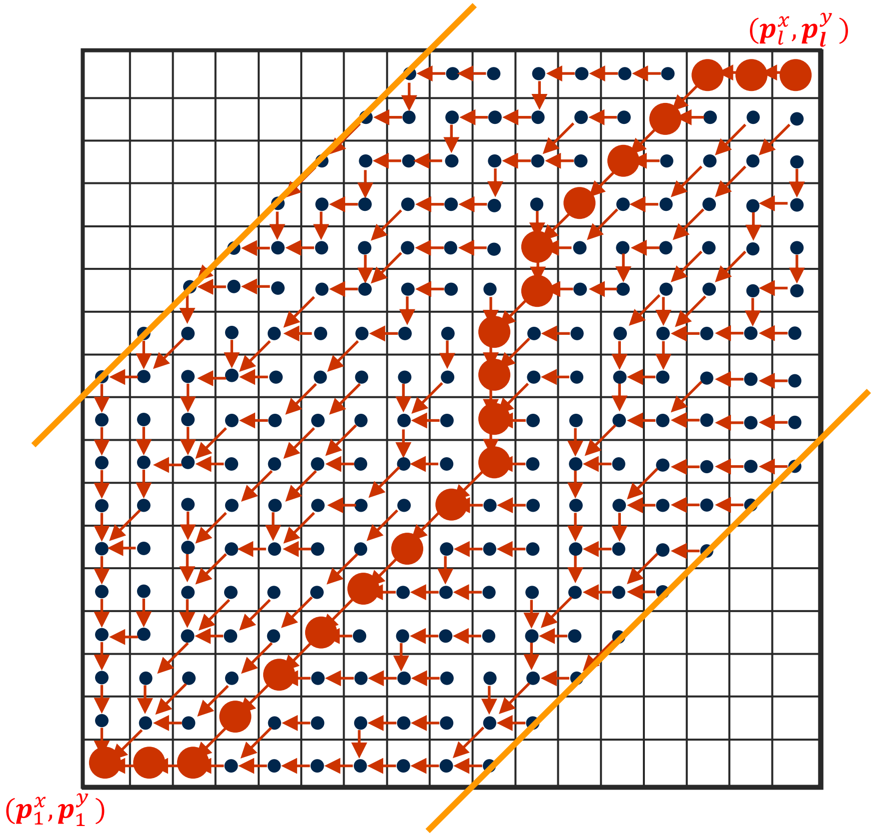


Dynamic Time Warping (DTW)

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

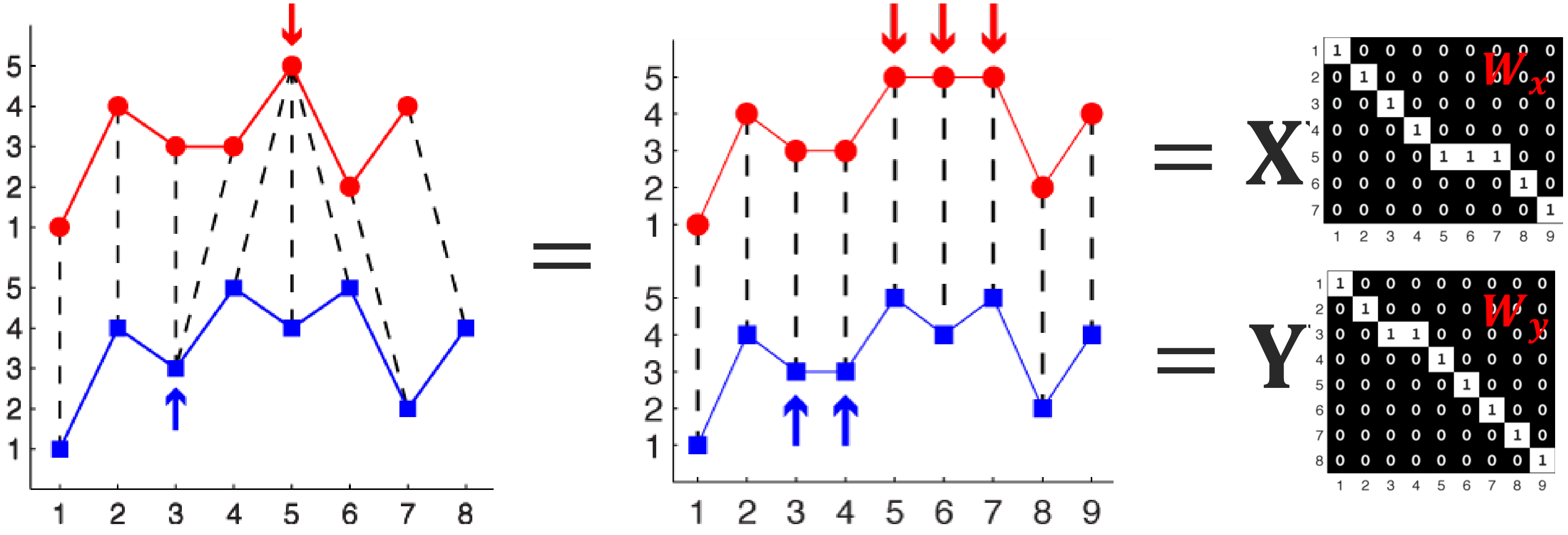
Solved using dynamic programming while respecting the restrictions



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

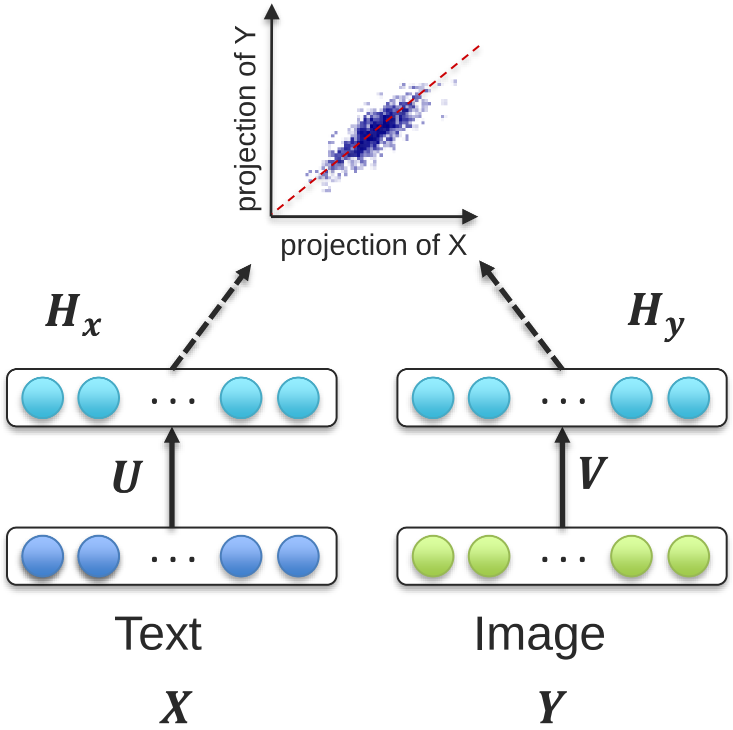
Canonical Correlation Analysis – Reminder

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:

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



Canonical Time Warping

Dynamic Time Warping + Canonical Correlation Analysis = Canonical Time Warping

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

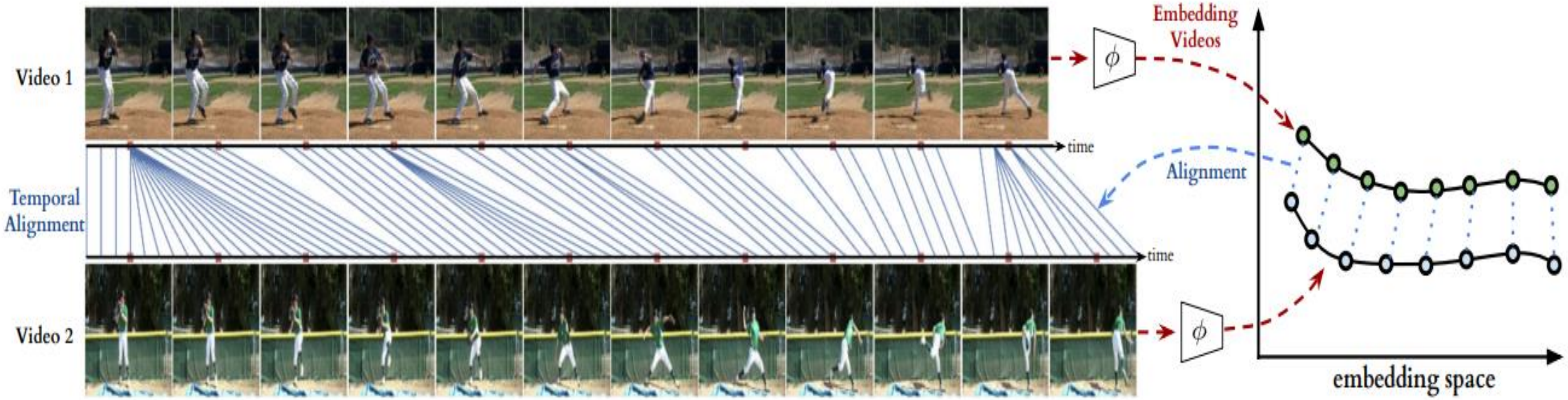
Allows to align multi-modal or multi-view (same modality but from a different point of view)

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

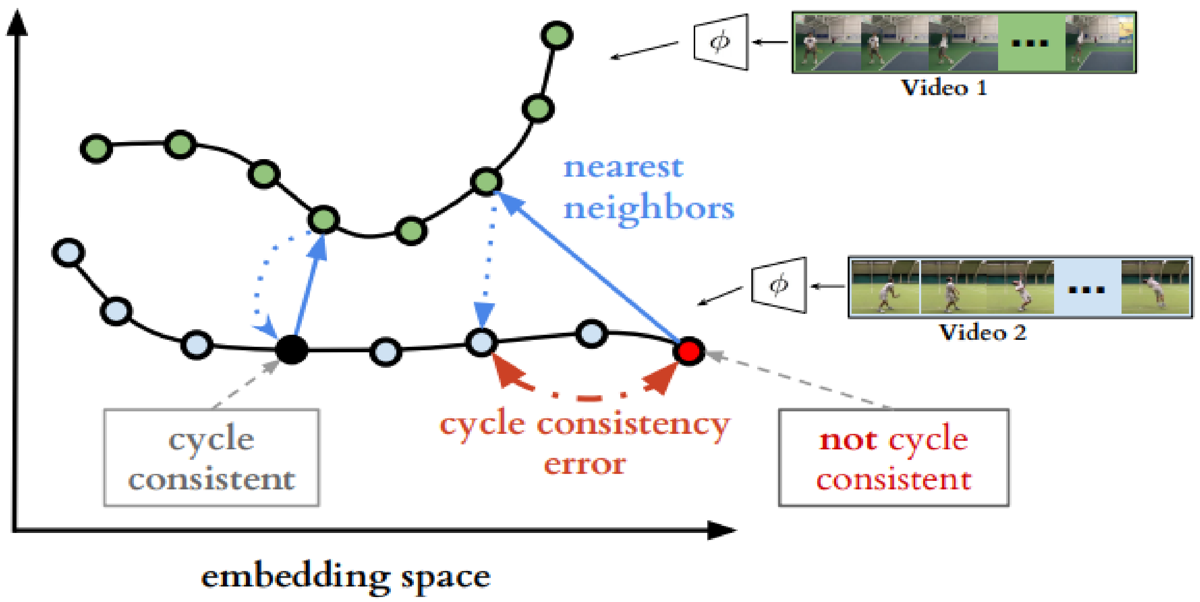
Premise: we have paired video sequences that can be be temporally aligned



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

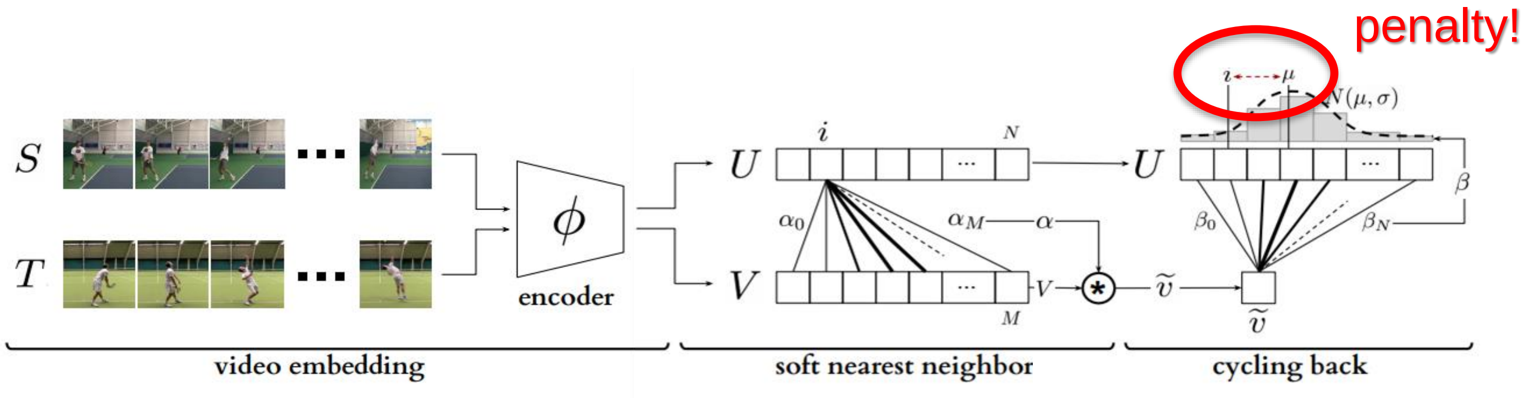
Temporal Cycle-Consistency Learning

<u>Solution:</u> 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: $$\alpha_j = \frac{e^{-||u_i - v_j||^2}}{\sum_k^M e^{-||u_i - v_k||^2}}$$

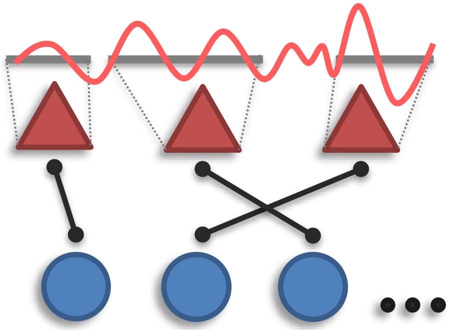
Soft nearest neighbor: $$\tilde{v} = \sum_j^M \alpha_j v_j,$$

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

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

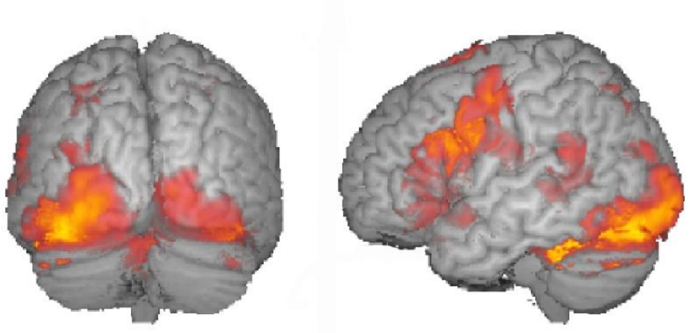
$$L_{cbr} = \frac{|i - \mu|^2}{\sigma^2} + \lambda \log(\sigma)$$

Discretization (aka Segmentation)

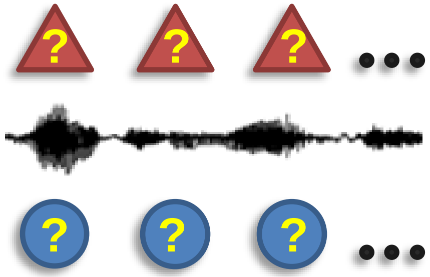


Common assumptions: ① Segmented elements

Examples:



Medical imaging



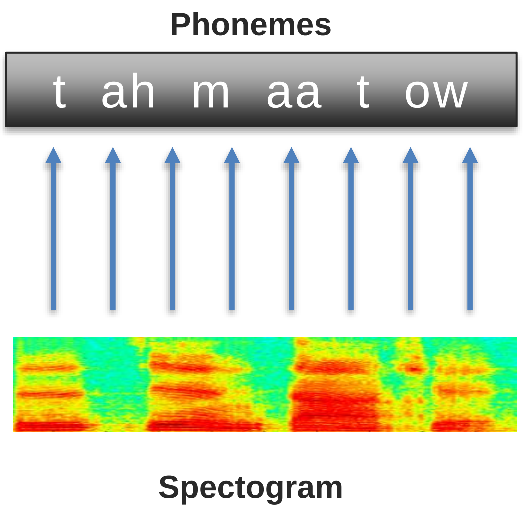
Signals

objects

Images

Discretization – Example

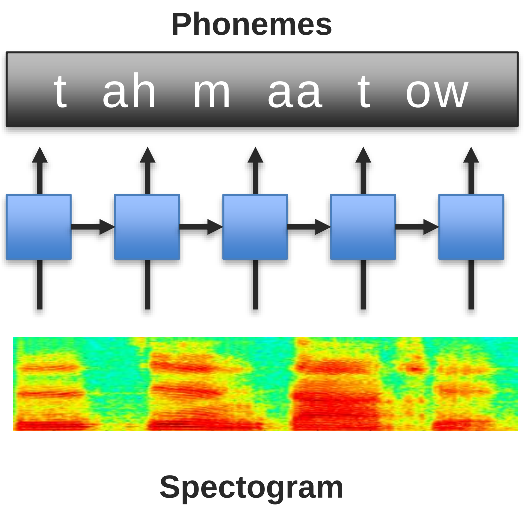
Sequence Labeling and Alignment



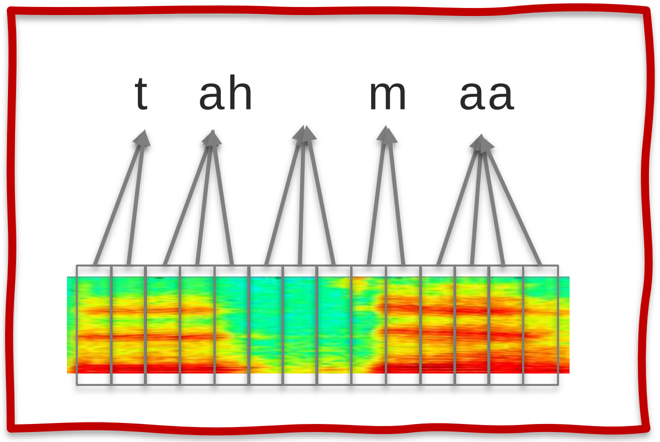
How can we predict the sequence of phoneme labels?

Discretization – Example

Sequence Labeling and Alignment



Challenge: many-to-1 alignment



How can we predict the sequence of phoneme labels?

Discretization – A Classification Approach

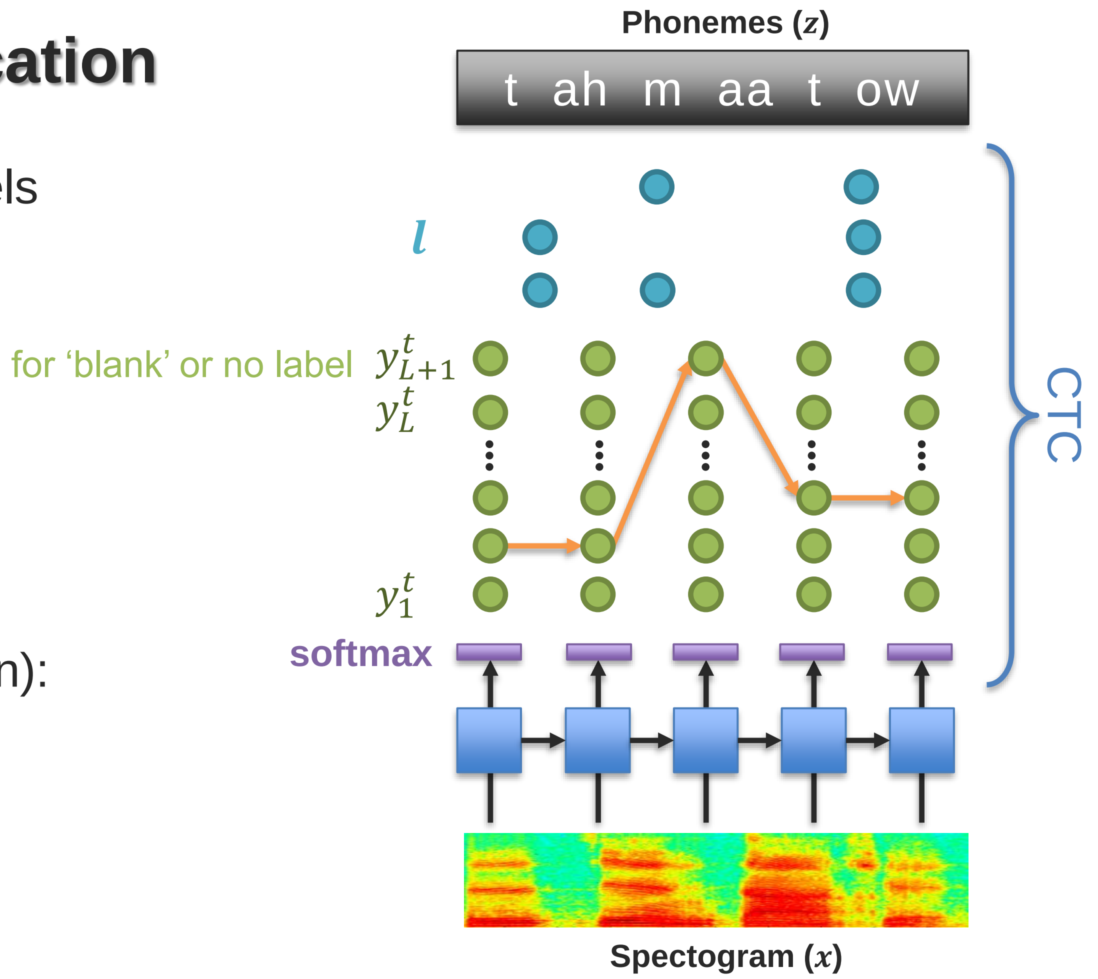
Connectionist Temporal Classification

4 Most probable sequence labels

3 Predicted labels $\boldsymbol{l}$

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

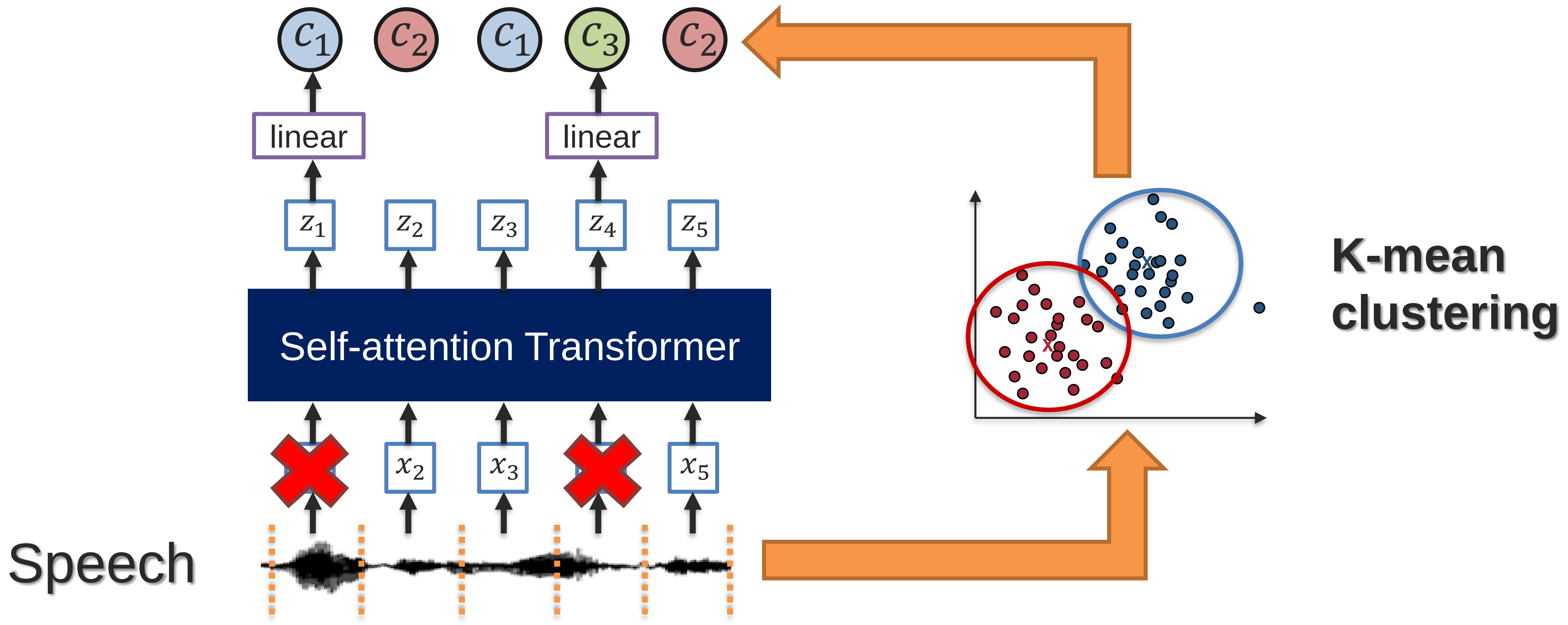
1 Output activations (distribution):



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

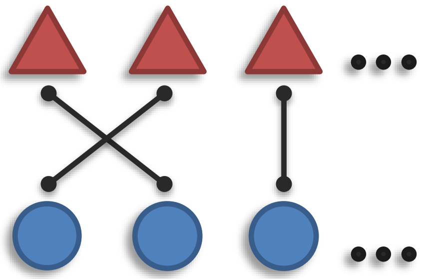
Challenge 2: Alignment

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

Sub-challenges:

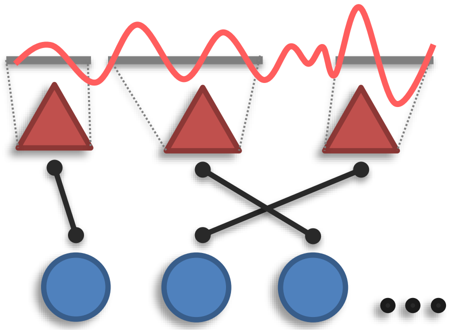
Next week!

Discrete Alignment



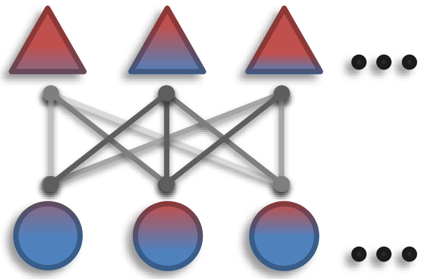
Discrete elements and connections

Continuous Alignment



Segmentation and continuous warping

Contextualized Representation



Implicit alignment + representation