Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

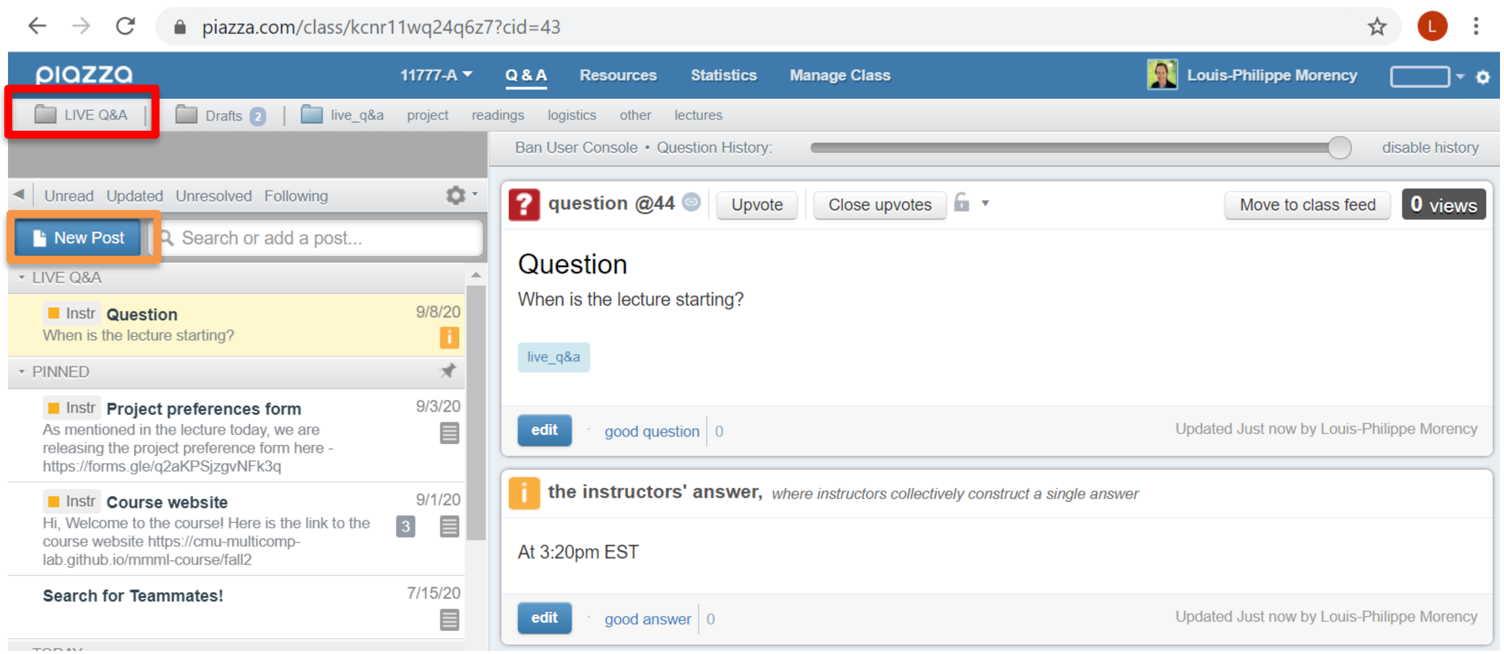
Lecture 5.1: Multimodal alignment

Louis-Philippe Morency

* Original version co-developed with Tadas Baltrusaitis

Administrative Stuff

Piazza Live Q&A – Reminder



Upcoming Schedule

First project assignment:

- **Proposal presentations (Friday 10/9)**
- **First project reports (Sunday 10/11)**

Midterm project assignment

- Midterm presentations (Friday 11/12)
- Midterm reports (Sunday 11/14)

Final project assignment

- Final presentations (Friday 12/11)
- Final reports (Sunday 12/13)

Unimodal Representation Analyses

Main goals:

- Get familiar with unimodal representations
  - Learn about tools based on CNNs, word2vec, …
- Understand the structure in your unimodal data
  - Perform some visualization of the unimodal data
- Explore qualitatively the unimodal data
  - How does it relate to your labels? Look at specific examples

Examples of unimodal analyses:

- What are the different verbs used in the VQA questions?
- What objects do not get detected? Are they important?
- Visualize face embeddings with respect of emotion labels

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 5.1: Multimodal alignment

Louis-Philippe Morency

* Original version co-developed with Tadas Baltrusaitis

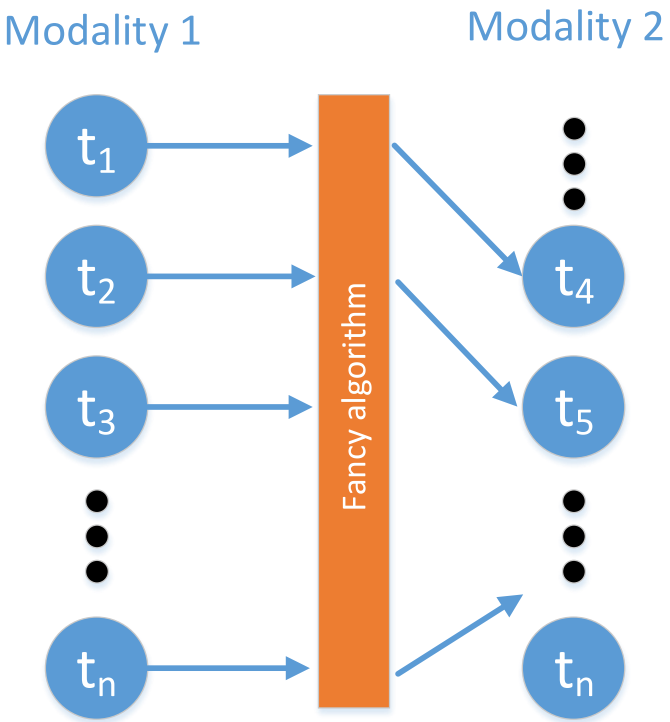
Lecture objectives

- Multimodal alignment
  - Implicit
  - Explicit
- Explicit signal alignment
  - Dynamic Time Warping
  - Canonical Time Warping
- Attention models in deep learning (implicit and explicit alignment)
  - Soft attention
  - Hard attention
  - Spatial Transformer Networks

Multimodal alignment

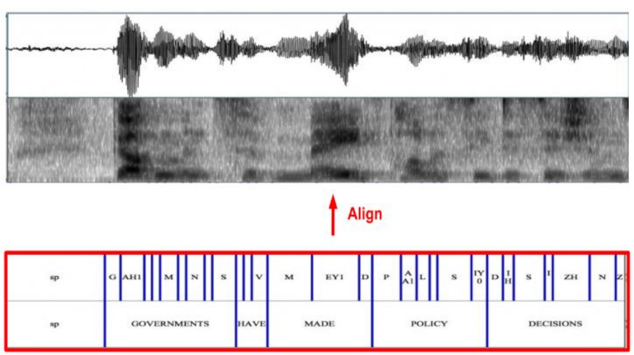
Multimodal-alignment

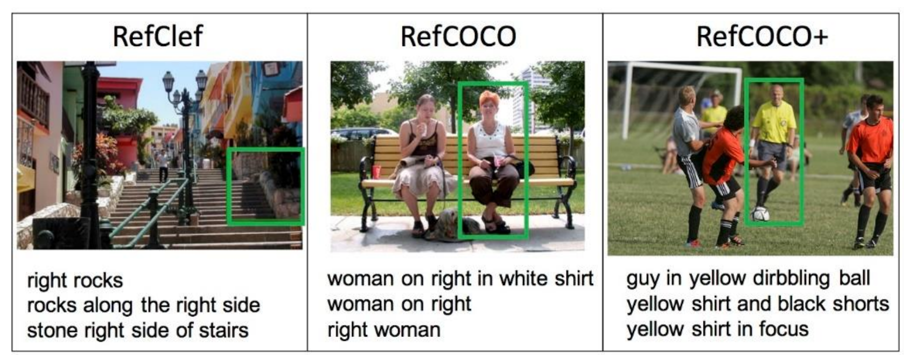
- Multimodal alignment – finding relationships and correspondences between two or more modalities
- Two types
  - **Explicit** – alignment is the task in itself
  - **Implicit / Latent** – alignment helps when solving a different task (for example "Attention" models)
- Examples ?
  - Images with captions
  - Recipe steps with a how-to video
  - Phrases/words of translated sentences



Explicit multimodal-alignment

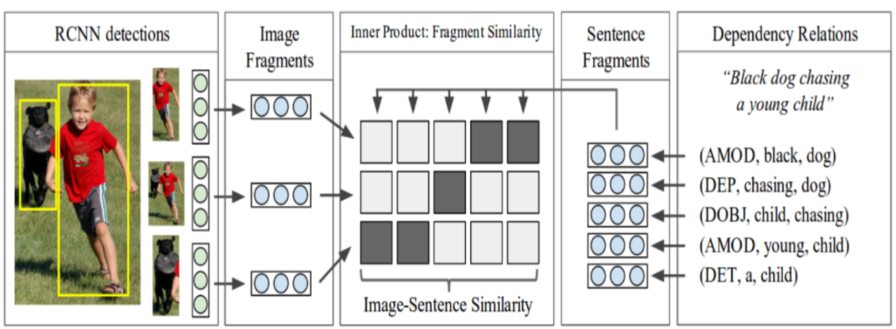
- Explicit alignment - goal is to find correspondences between modalities
  - Aligning speech signal to a transcript
  - Aligning two out-of sync sequences
  - Co-referring expressions

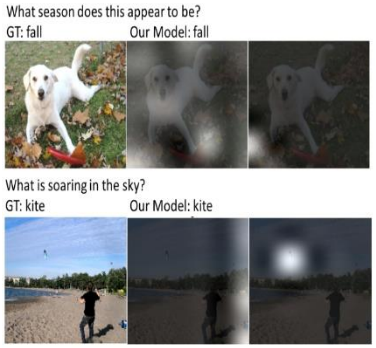




Implicit multimodal-alignment

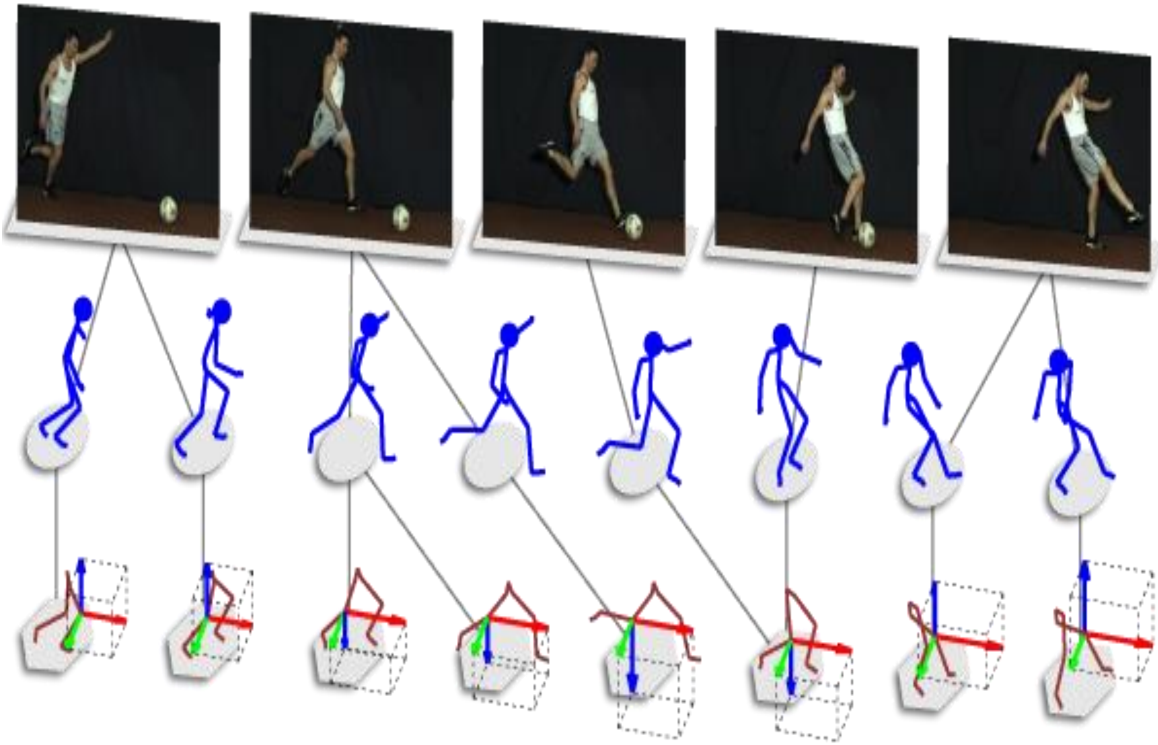
- Implicit alignment - uses internal latent alignment of modalities in order to better solve various problems
  - Machine Translation
  - Cross-modal retrieval
  - Image & Video Captioning
  - Visual Question Answering





Explicit alignment

Temporal sequence alignment



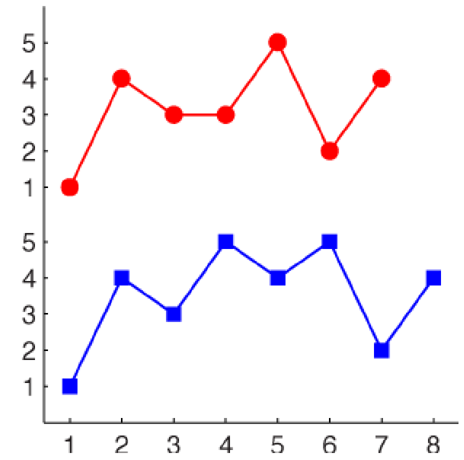
Applications:
- Re-aligning asynchronous data
- Finding similar data across modalities (we can estimate the aligned cost)
- Event reconstruction from multiple sources

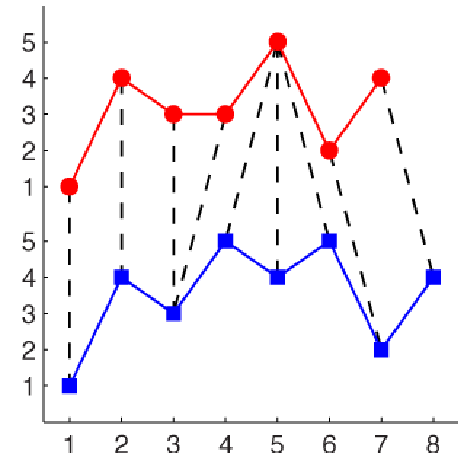
Let's start unimodal – Dynamic Time Warping

- 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:

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

- Where $\boldsymbol{p}^x$ and $\boldsymbol{p}^y$ are index vectors of same length
- Dynamic Time Warping is designed to find these index vectors

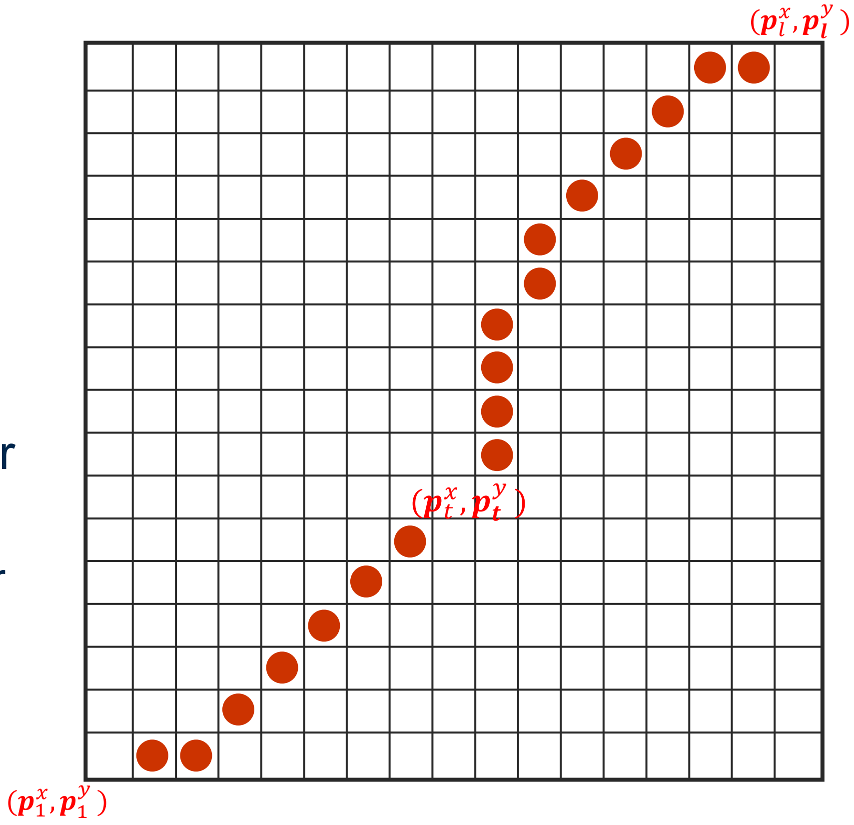




Dynamic Time Warping continued

Lowest cost path in a cost matrix

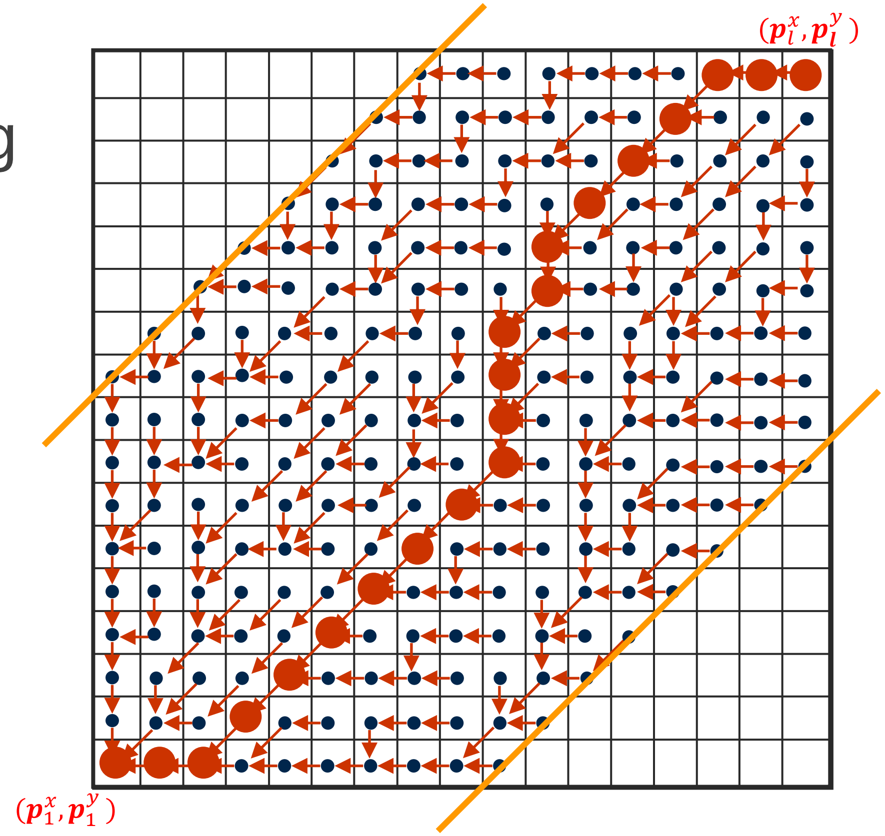
- Restrictions?
    - Monotonicity – no going back in time
    - Continuity - no gaps
    - Boundary conditions - start and end at the same points
    - Warping window - don't get too far from diagonal
    - Slope constraint – do not insert or skip too much



Dynamic Time Warping continued

Lowest cost path in a cost matrix

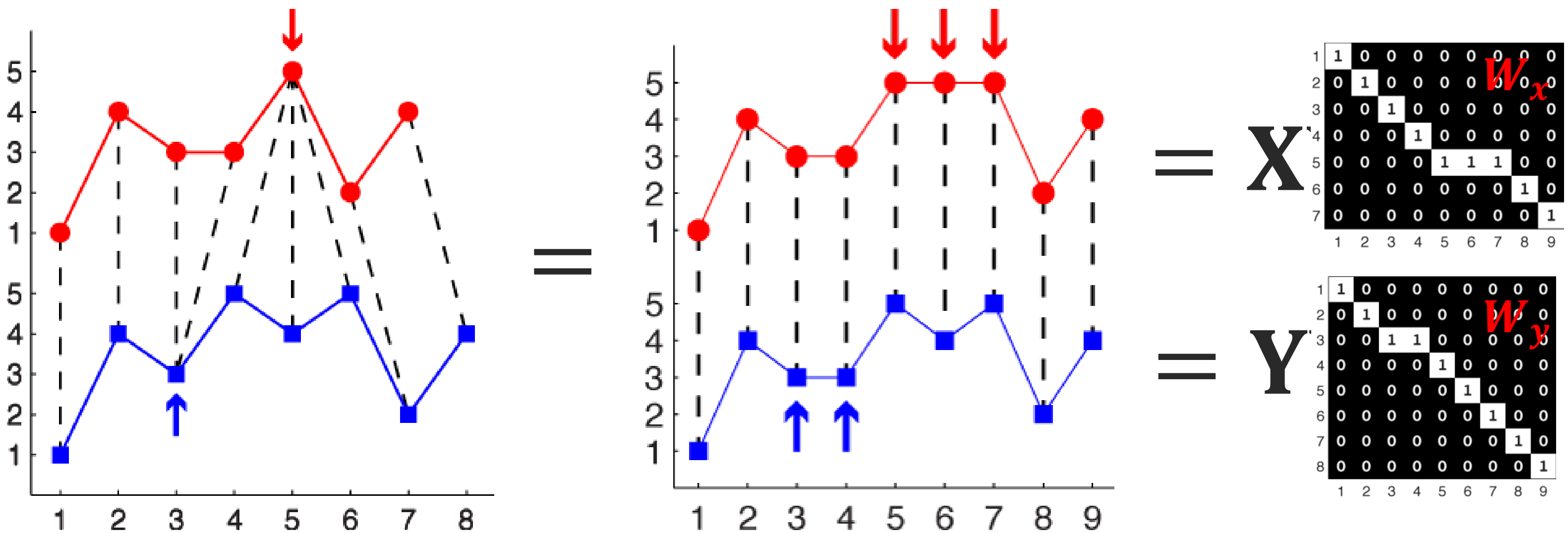
- Solved using dynamic programming while respecting the restrictions



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:

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

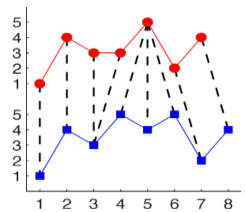
Frobenius norm $\|\boldsymbol{A}\|_F^2 = \sum_i \sum_j |a_{i,j}|^2$

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

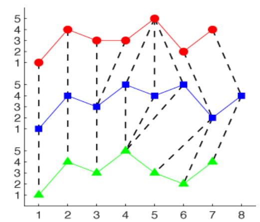
$\boldsymbol{W}_x, \boldsymbol{W}_y$ - alignment matrices

DTW – Some Limitations

- Computationally complex



$O(n_x n_y)$



$O(n_x n_y n_z)$

m sequences

$$O(\prod_{i=1}^{m} n_i)$$

- Sensitive to outliers

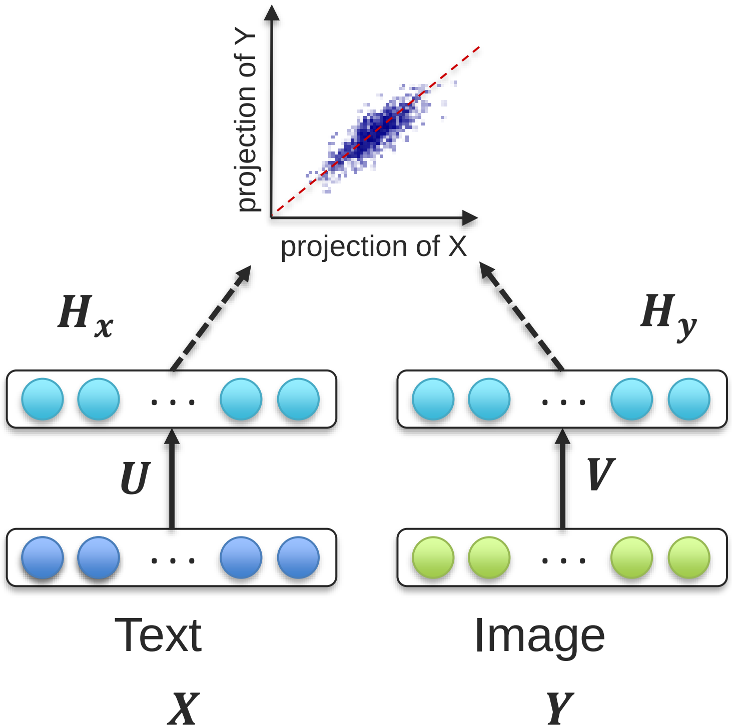
- Unimodal!

Canonical Correlation Analysis reminder

maximize: $tr(\boldsymbol{U}^T\boldsymbol{\Sigma}_{XY}\boldsymbol{V})$

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

1. Linear projections maximizing correlation
2. Orthogonal projections
3. Unit variance of the projection vectors

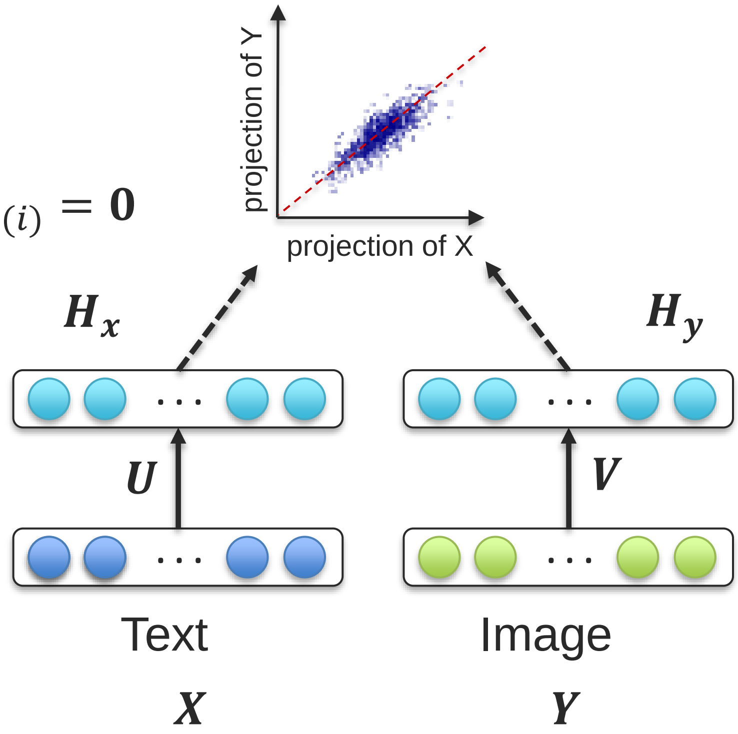


Canonical Correlation Analysis reminder

- When data is normalized it is actually equivalent to smallest RMSE reconstruction
- 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

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

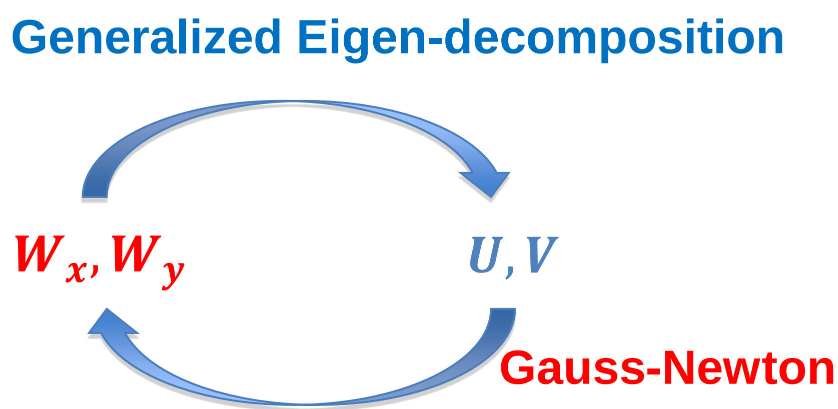
- Allows to align multi-modal or multi-view (same modality but from a different point of view)
- $\boldsymbol{W}_x, \boldsymbol{W}_y$ – temporal alignment
- $\boldsymbol{U}, \boldsymbol{V}$ – cross-modal (spatial) alignment

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

Canonical Time Warping

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

Optimized by Coordinate-descent – fix one set of parameters, optimize another



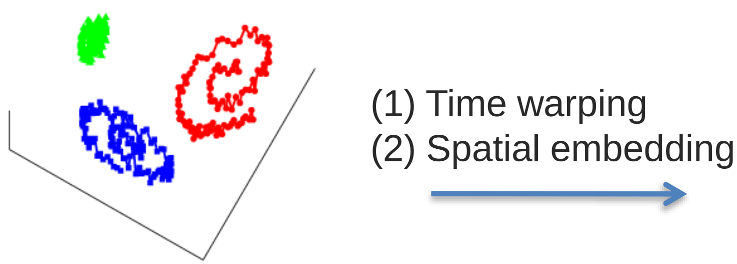
[Canonical Time Warping for Alignment of Human Behavior, Zhou and De la Tore, 2009, NIPS]

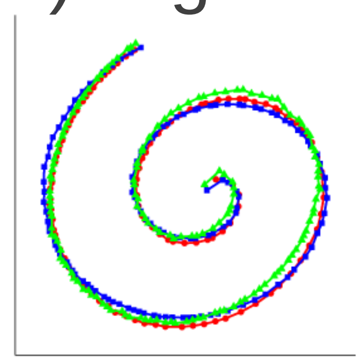
Generalized Time warping

- Generalize to multiple sequences all of different modality

$$L(\boldsymbol{U_i}, \boldsymbol{W_i}) = \sum_{i=1} \sum_{j=1} \left\| \mathbf{U}_i^T \mathbf{X}_\mathrm{i} \mathbf{W}_\mathrm{i} - \mathbf{U}_j^T \mathbf{X}_\mathrm{j} \mathbf{W}_j \right\|_F^2$$

- $\boldsymbol{W_i}$ – set of temporal alignments
- $\boldsymbol{U_i}$ – set of cross-modal (spatial) alignments



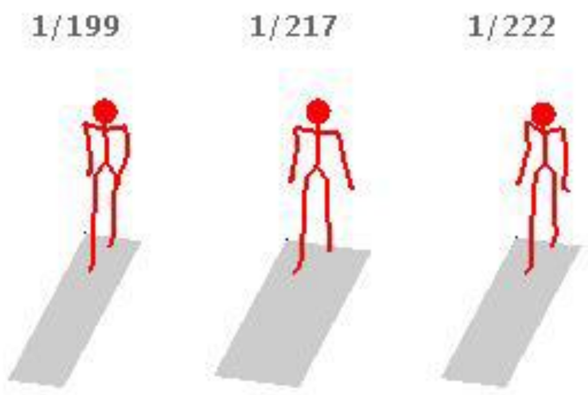


[Generalized Canonical Time Warping, Zhou and De la Tore, 2016, TPAMI]

Alignment examples (unimodal)

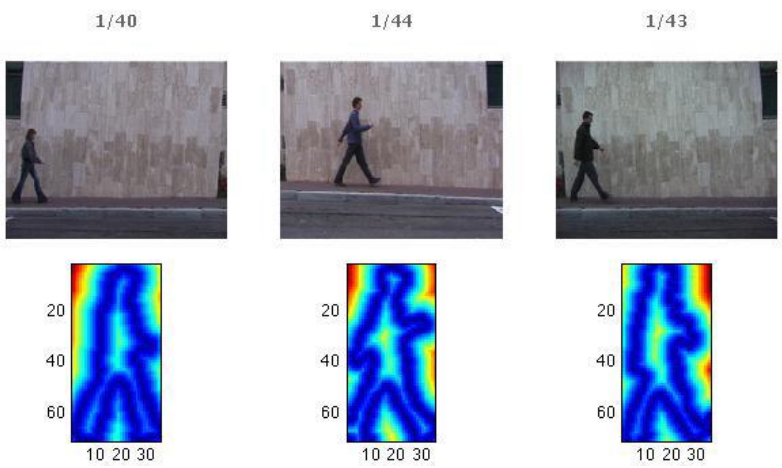
CMU Motion Capture

- Subject 1: 199 frames
- Subject 2: 217 frames
- Subject 3: 222 frames



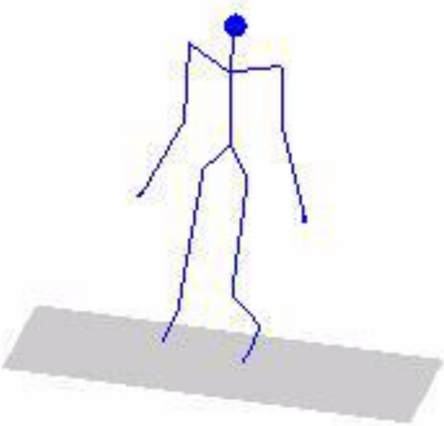
Weizmann

- Subject 1: 40 frames
- Subject 2: 44 frames
- Subject 3: 43 frames



Alignment examples (multimodal)

1/273



1/51

1/127

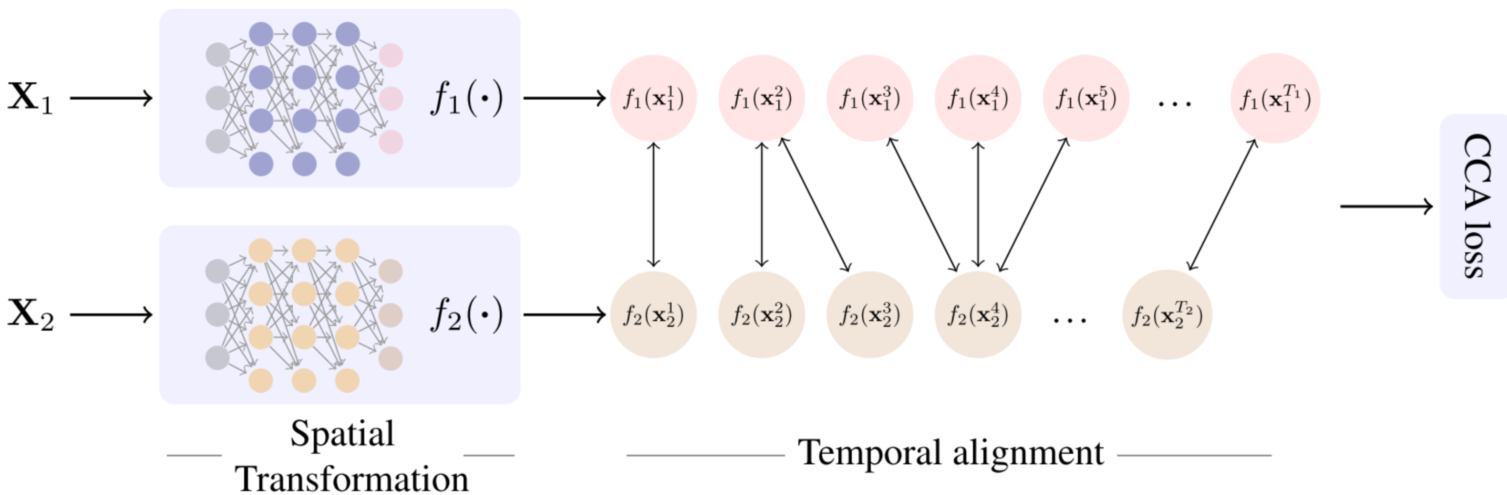
Canonical time warping - limitations

- Linear transform between modalities
- How to address this?

Deep Canonical Time Warping

$$L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{W}_x, \boldsymbol{W}_y) = \left\| f_{\boldsymbol{\theta}_1}(\mathbf{X})\mathbf{W_x} - f_{\boldsymbol{\theta}_1}(\mathbf{Y})\mathbf{W_y} \right\|_F^2$$

- Could be seen as generalization of DCCA and GTW



[Deep Canonical Time Warping, Trigeorgis et al., 2016, CVPR]

Deep Canonical Time Warping

$$L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{W}_x, \boldsymbol{W}_y) = \left\| f_{\boldsymbol{\theta}_1}(\mathbf{X})\mathbf{W_x} - f_{\boldsymbol{\theta}_1}(\mathbf{Y})\mathbf{W_y} \right\|_F^2$$

- The projections are orthogonal (like in DCCA)
- Optimization is again iterative:
  - Solve for alignment ($\boldsymbol{W}_x, \boldsymbol{W}_y$) with fixed projections ($\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$)
    - Eigen decomposition
  - Solve for projections ($\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$) with fixed alignment ($\boldsymbol{W}_x, \boldsymbol{W}_y$)
    - Gradient descent
  - Repeat till convergence

[Deep Canonical Time Warping, Trigeorgis et al., 2016, CVPR]

Implicit alignment

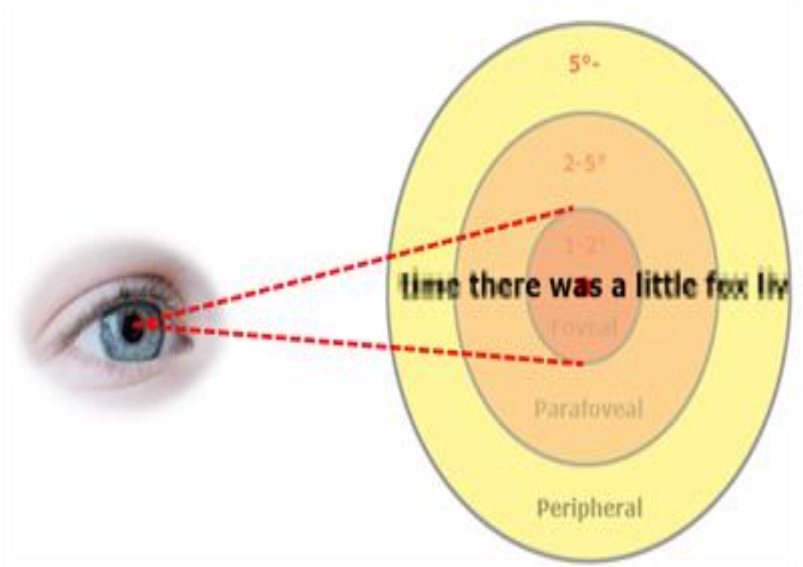
Implicit alignment

- We looked how to explicitly align temporal data
- Could use that as an internal (hidden) step in our models?
- Can we instead encourage the model to align data when solving a different problem?
- Yes!
  - Graphical models
  - Neural attention models (focus of today's lecture)

Attention models

Attention in humans

- Foveal vision – we only see in “high resolution” in 2 degrees of vision
- We focus our attention selectively to certain words (for example our names)
- We attend to relevant speech in a noisy room



Attention models in deep learning

- Many examples of attention models in recent years!
- Why:
  - Allows for implicit data alignment
  - Good results empirically
  - In some cases faster (don’t need to focus on all the image)
  - Better Interpretability

Types of Attention Models

- Recent attention models can be roughly split into three major categories
  1. Soft attention
     - Acts like a gate function. Deterministic inference.
  2. Transform network
     - Warp the input to better align with canonical view
  3. Hard attention
     - Includes stochastic processes. Related to reinforcement learning.

Soft attention

Machine Translation

Given a sentence in one language translate it to another

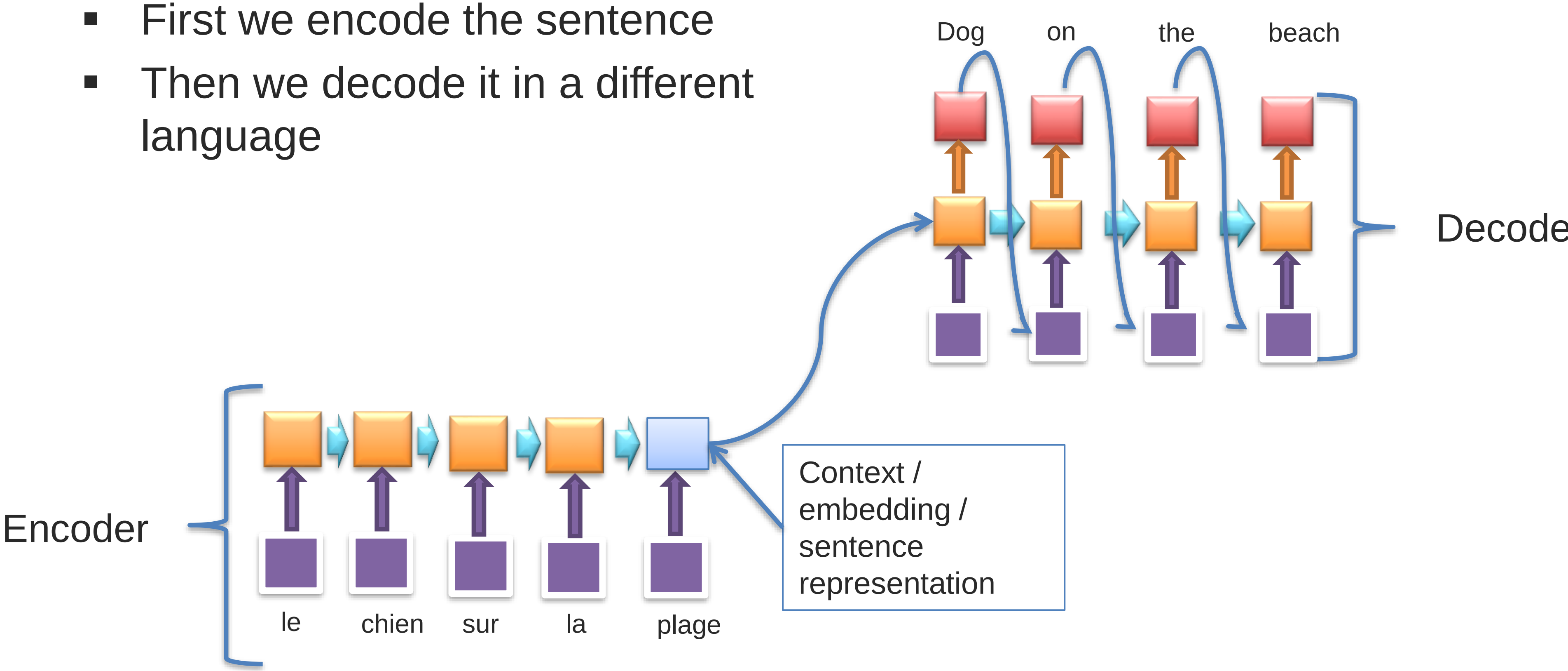
Dog on the beach → le chien sur la plage

- Not exactly multimodal task – but a good start! Each language can be seen almost as a modality.

Machine Translation with RNNs

A quick reminder about encoder decoder frameworks

- First we encode the sentence
- Then we decode it in a different language



Machine Translation with RNNs

What is the problem with this?

What happens when the sentences are very long?

- We expect the encoders hidden state to capture everything in a sentence, a very complex state in a single vector, such as

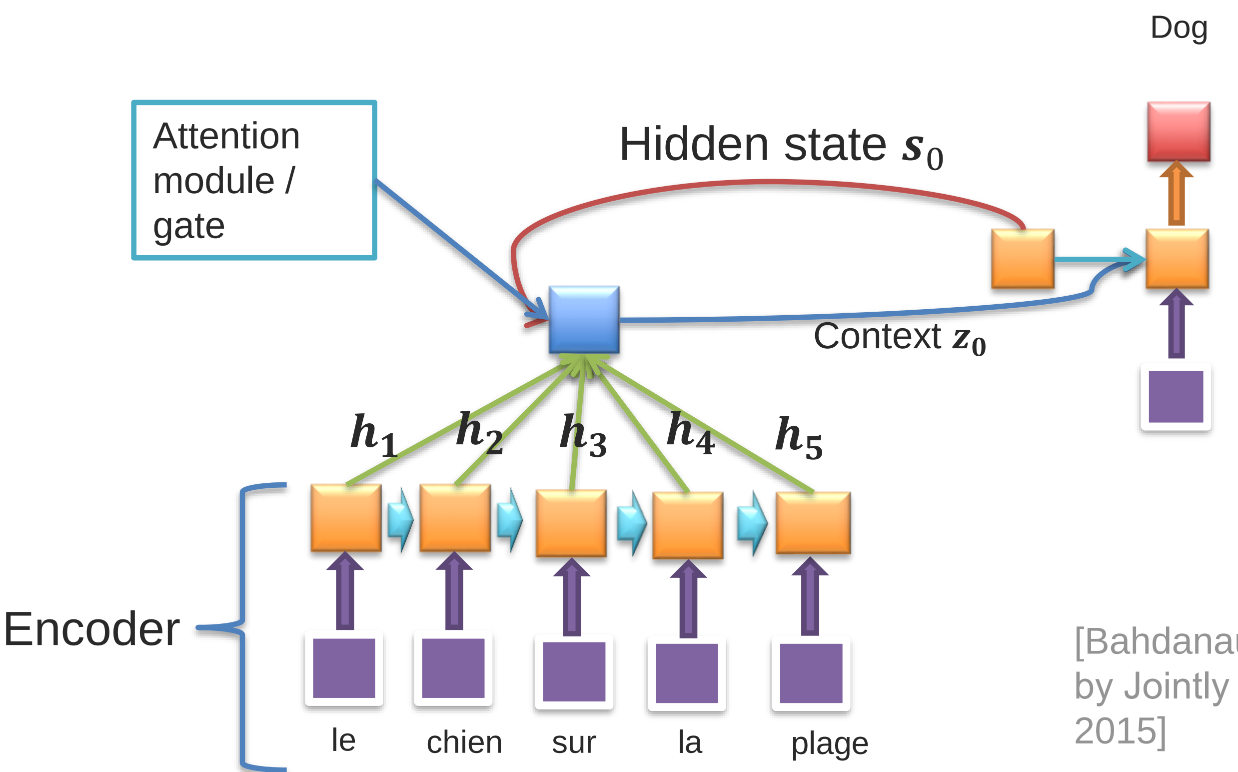
The agreement on the European Economic Area was signed in August 1992.

↓

L' accord sur la zone économique européenne a été signé en août 1992.

Decoder – attention model

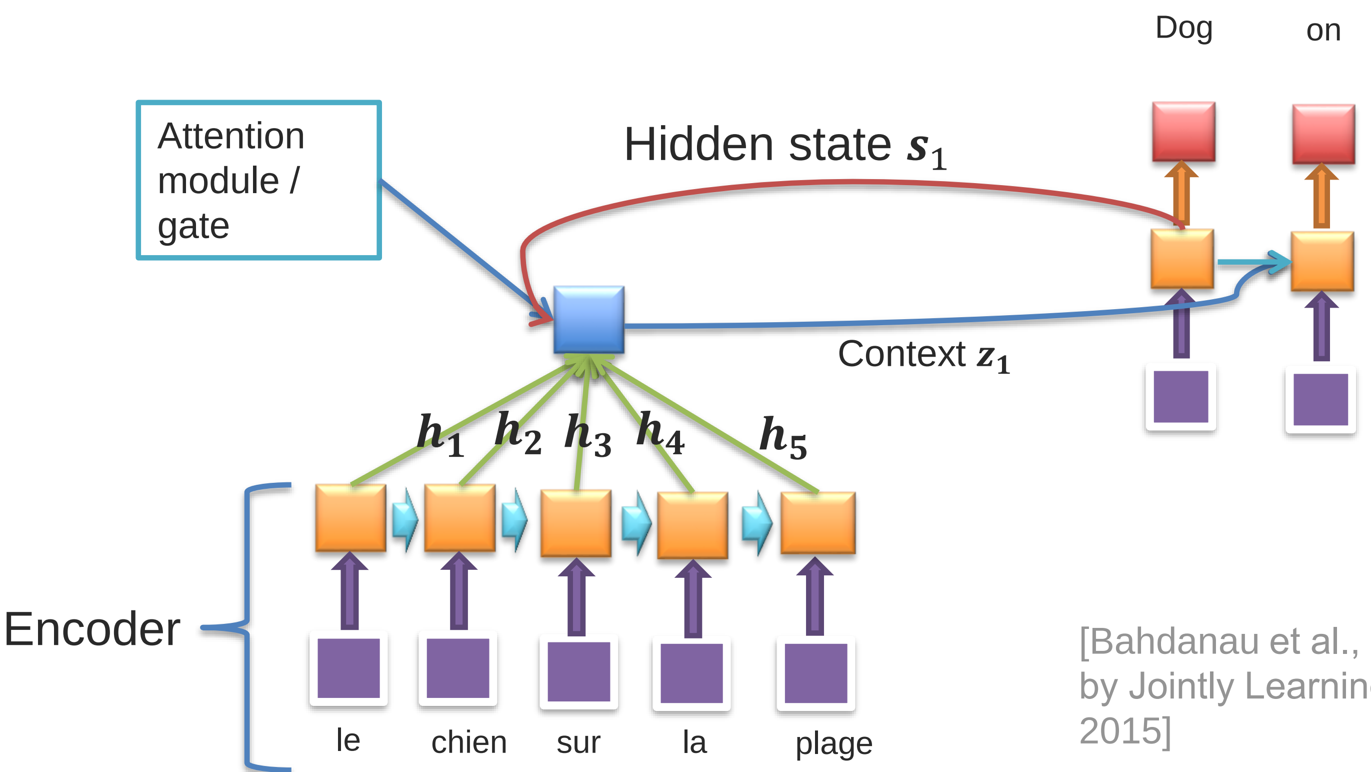
- Before encoder would just take the final hidden state, now we actually care about the intermediate hidden states



[Bahdanau et al., "Neural Machine Translation by Jointly Learning to Align and Translate", ICLR 2015]

Decoder – attention model

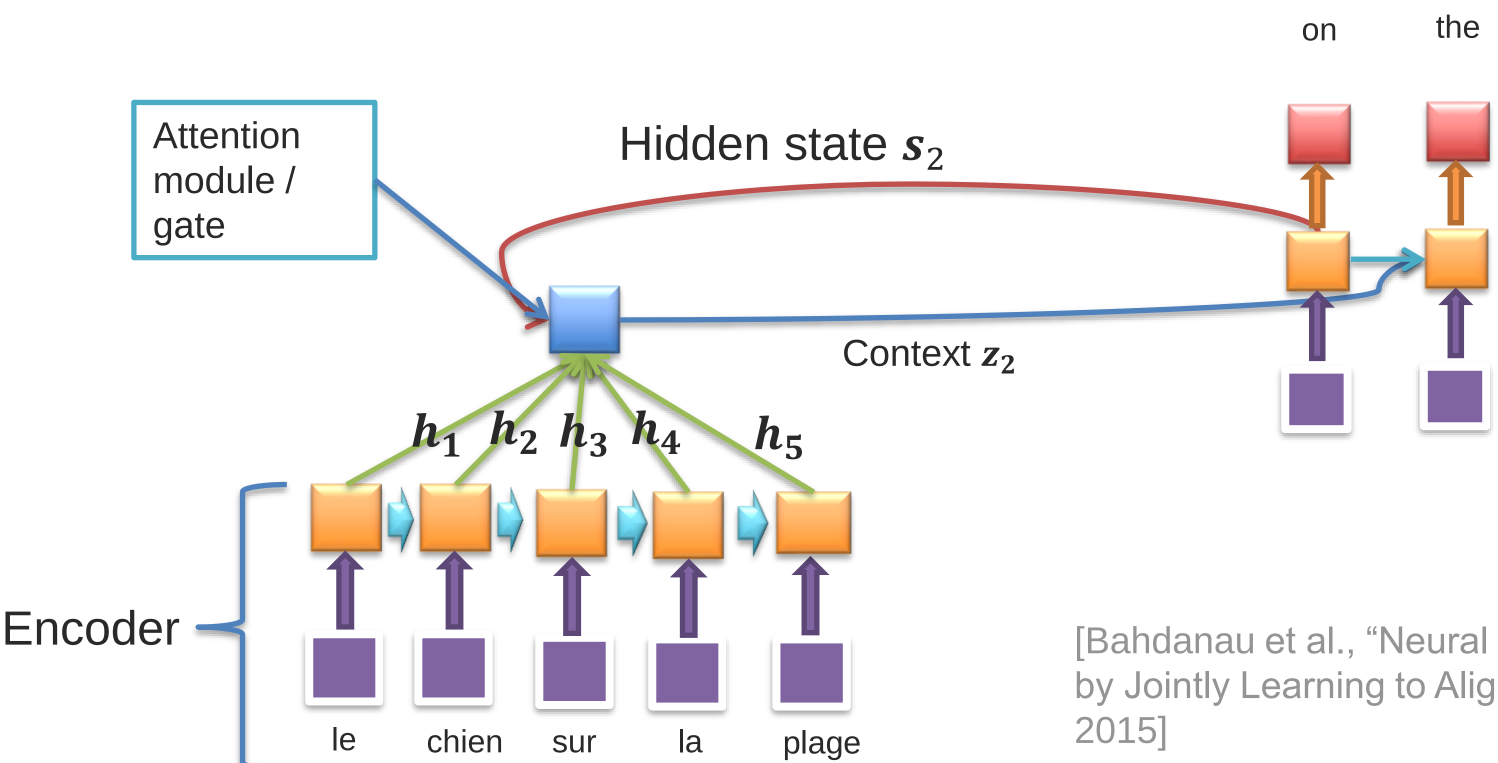
- Before encoder would just take the final hidden state, now we actually care about the intermediate hidden states



[Bahdanau et al., “Neural Machine Translation by Jointly Learning to Align and Translate”, ICLR 2015]

Decoder – attention model

- Before encoder would just take the final hidden state, now we actually care about the intermediate hidden states



[Bahdanau et al., "Neural Machine Translation by Jointly Learning to Align and Translate", ICLR 2015]

How do we encode attention?

Before:

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

MT with attention

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 given the current state of decoder how important the current encoding is now

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

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

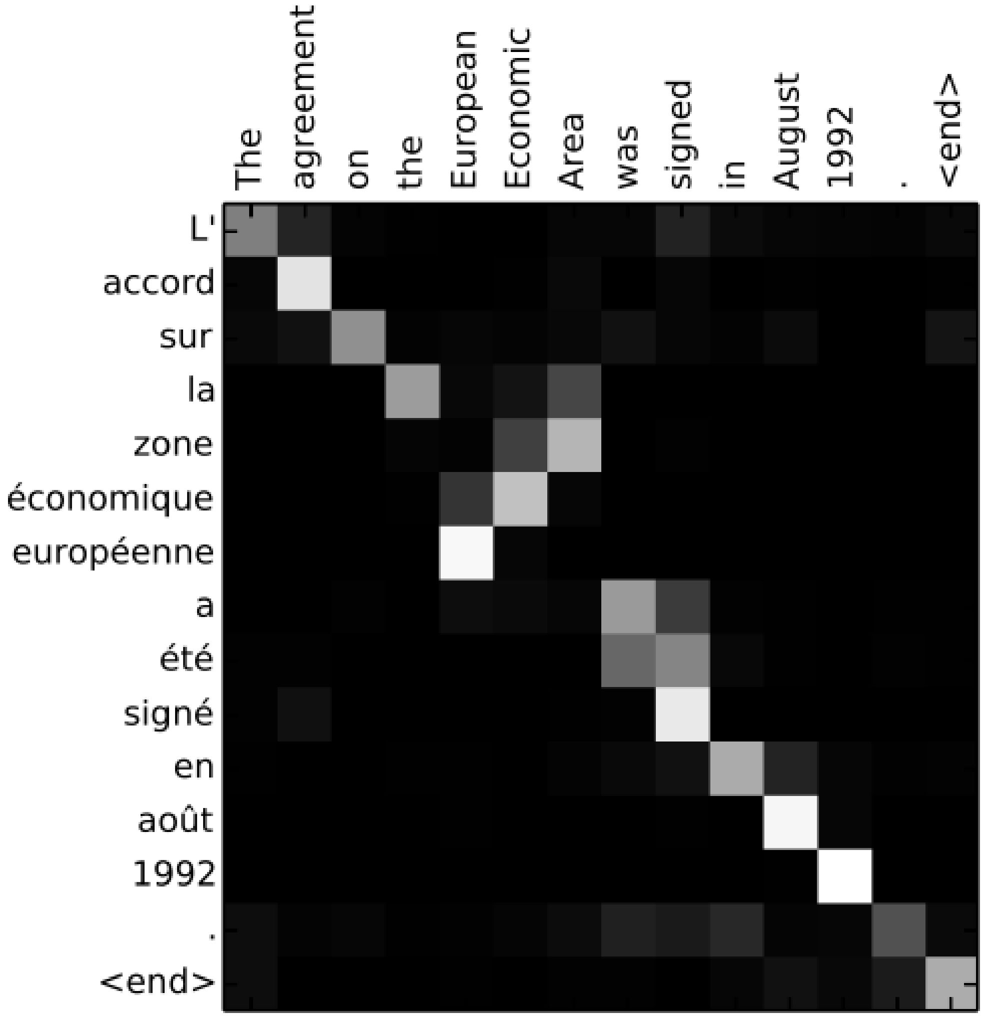
expectation of the context (a fancy way to say it's a weighted average)

MT with attention

Basically we are using a neural network to tell us where a neural network should be looking!

- We can use with RNN, LSTM or GRU
- Encoder being used is the same structure as before
  - Can use uni-directional
  - Can use bi-directional
- Model can be trained using our regular back-propagation through time, all of the modules are differentiable

Does it work?


The
agreement
on
the
European
Economic
Area
was
signed
in
August
1992
.
<end>
L'
accord
sur
la
zone
économique
européenne
a
été
signé
en
août
1992
.
<end>

MT with attention recap

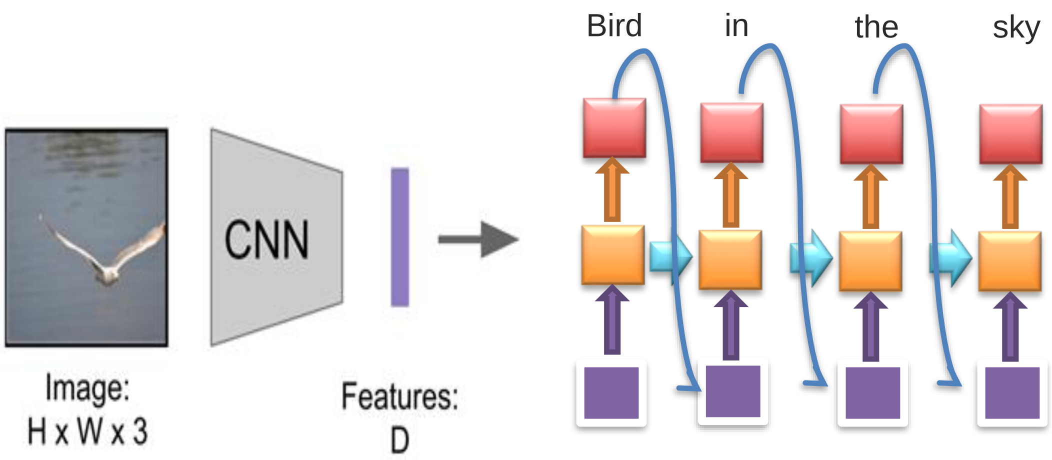
- It gives good translation results (especially for long sentences)
- Also get a (soft) alignment of sentences in different languages
  - Extra interpretability of method functioning

How do we move to multimodal?

Visual captioning with soft attention

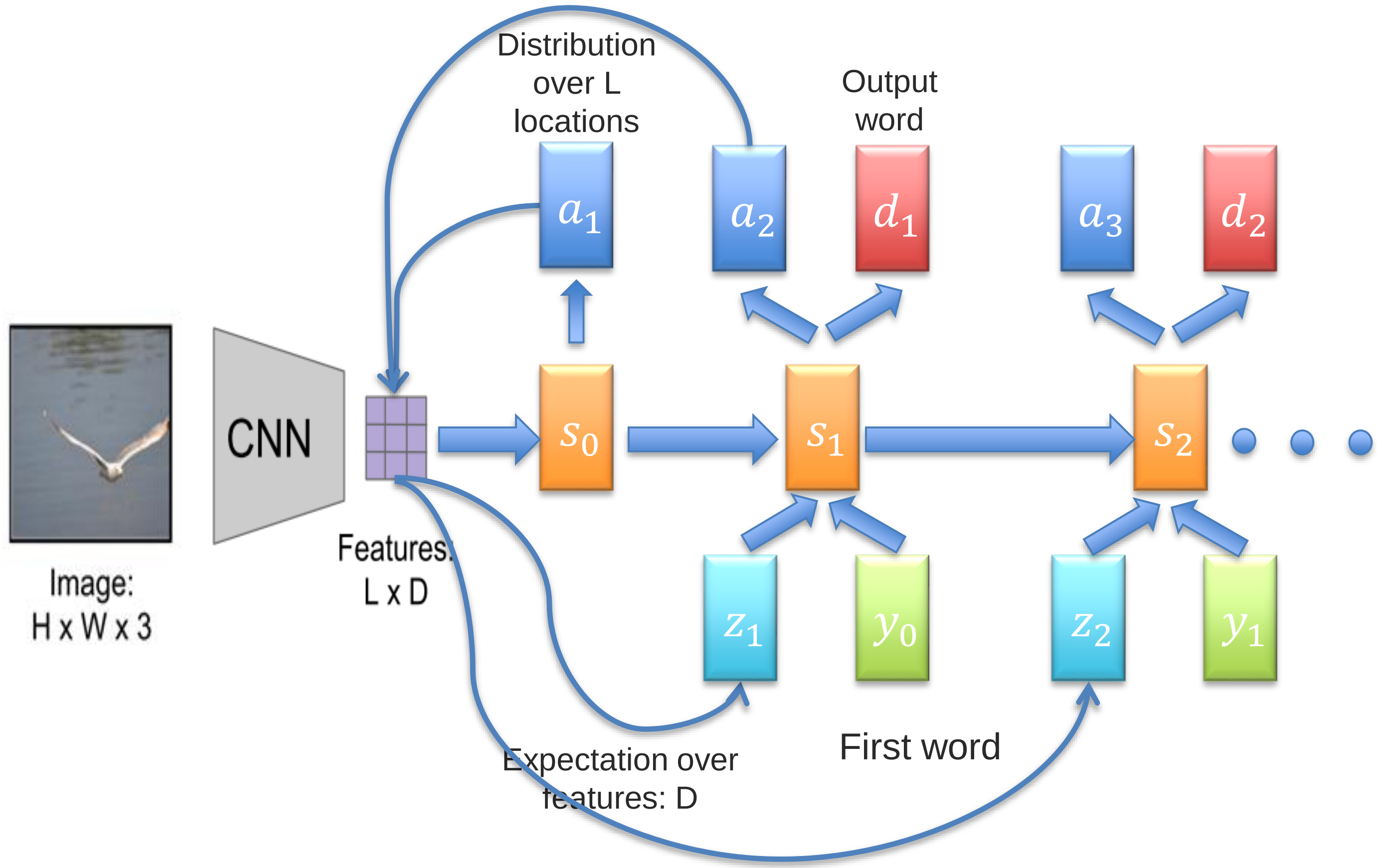
[Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, Xu et al., 2015]

Recap RNN for Captioning



Why not using final layer of the CNN?

Looking at more fine grained features



Soft attention

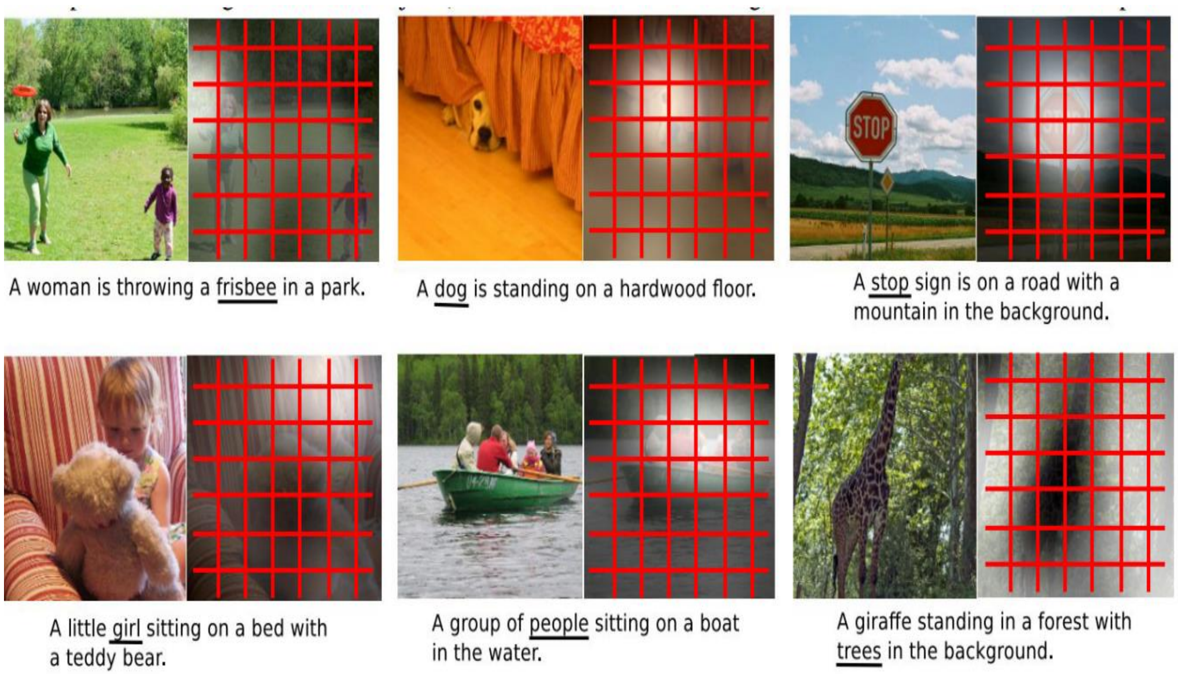
- Allows for latent data alignment
- Allows us to get an idea of what the network “sees”
- Can be optimized using back propagation

- Good at paper naming!
    - Show, Attend and Tell (extension of Show and Tell)
    - Listen, Attend and Walk
    - Listen, Attend and Spell
    - Ask, Attend and Answer

Spatial Transformer networks

Some limitations of grid-based attention

Can we fixate on small parts of image but still have easy end-to-end training?



Spatial Transformer Networks

Input image:
H x W x 3

Box Coordinates:
(xc, yc, w, h)

Can we make this function differentiable?

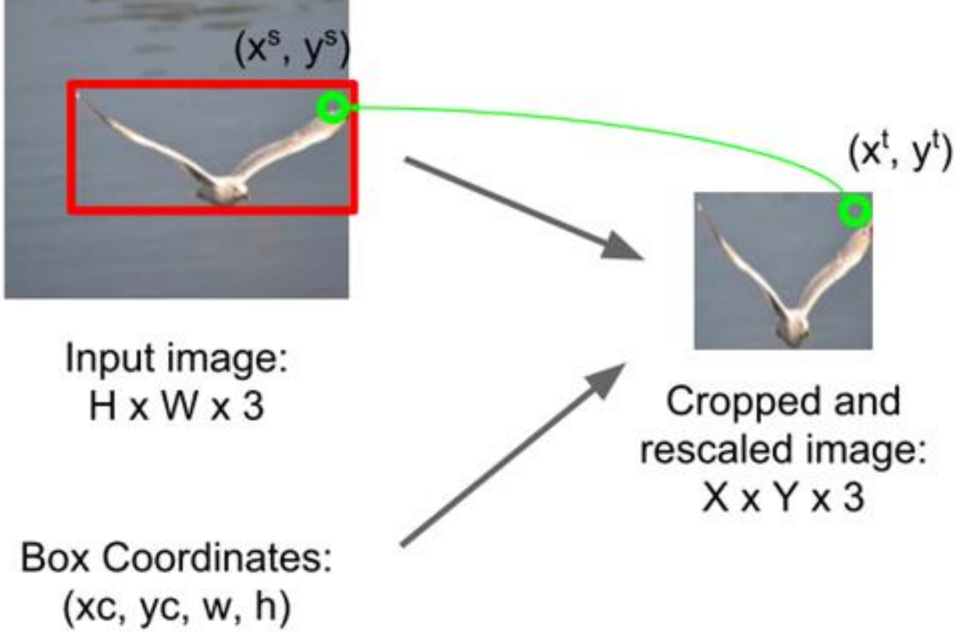
Cropped and rescaled image:
X x Y x 3

Spatial Transformer Networks

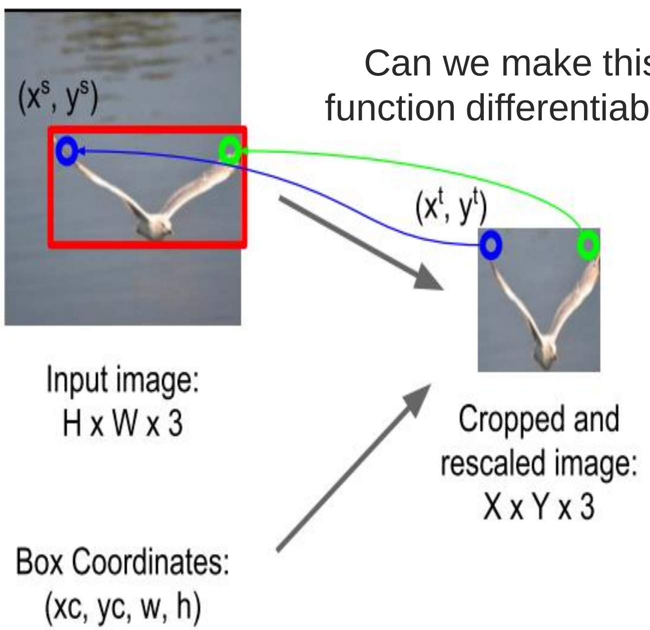
Idea: Function mapping pixel coordinates (x^t, y^t) of output to pixel coordinates (x^s, y^s) of input

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \theta_{1,3} \\ \theta_{2,1} & \theta_{2,2} & \theta_{2,3} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}$$

Can we make this function differentiable?



Spatial Transformer Networks

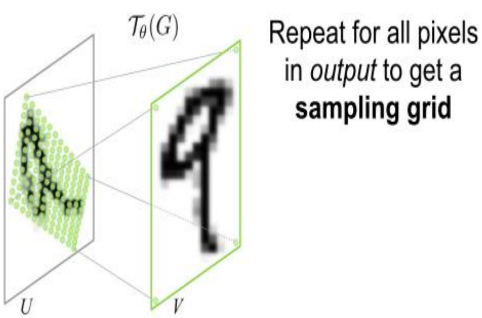


Can we make this function differentiable?

Idea: Function mapping pixel coordinates (x^t, y^t) of output to pixel coordinates (x^s, y^s) of input

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \theta_{1,3} \\ \theta_{2,1} & \theta_{2,2} & \theta_{2,3} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}$$

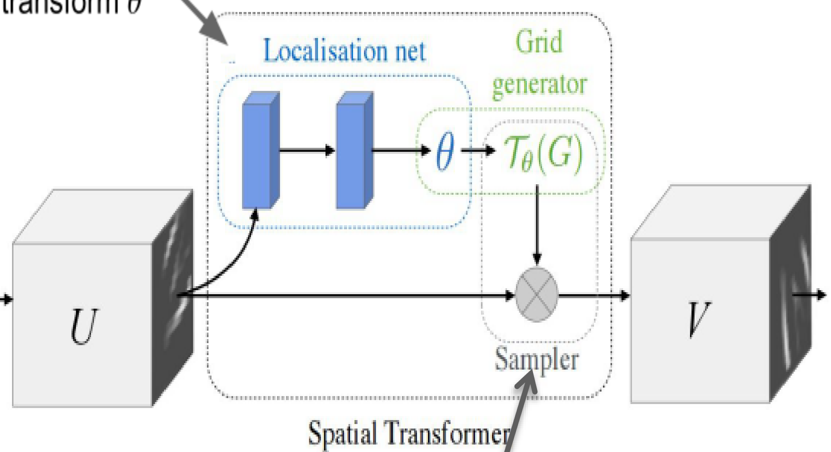
Network "attends" to input by predicting θ



Spatial Transformer Networks

A small **Localization network** predicts transform θ

Input: Full image



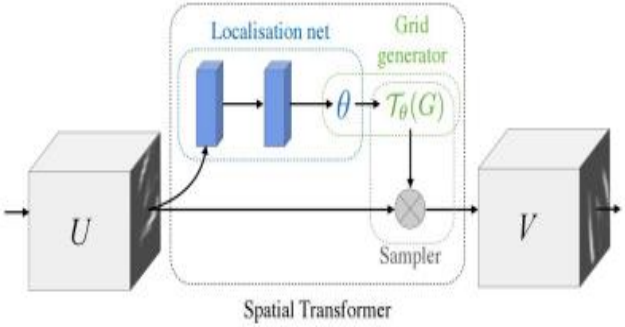
Output: Region of interest from input

Sampler uses bilinear interpolation to produce output

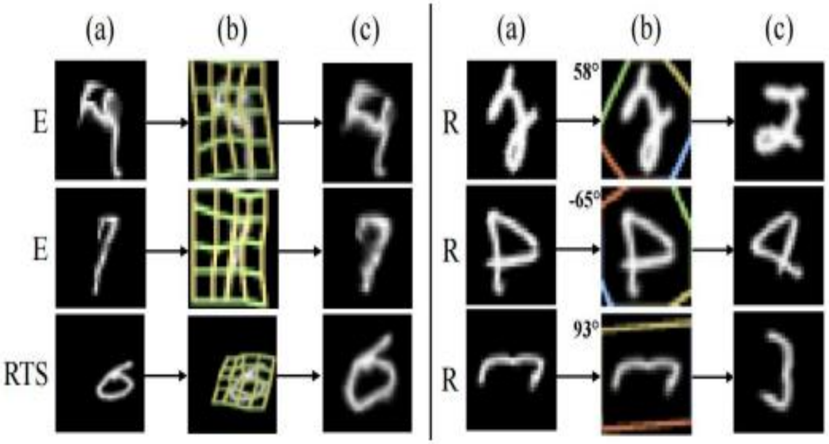
$$V_i^c = \sum_{n}^{H} \sum_{m}^{W} U_{nm}^c \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|)$$

Spatial Transformer Networks

Differentiable "attention / transformation" module

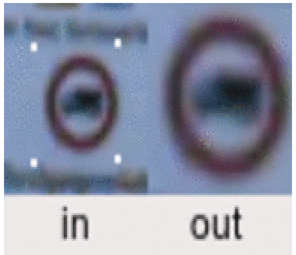


Insert spatial transformers into a classification network and it learns to attend and transform the input



Examples on real world data

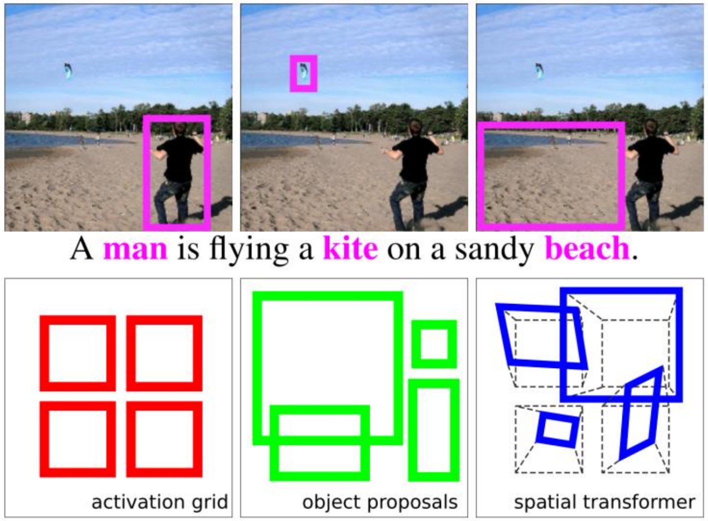
Results on traffic sign recognition



Code available http://torch.ch/blog/2015/09/07/spatial_transformers.html

Recap on Spatial Transformer Networks

- Differentiable so we can just use back-prop for training end-to-end
- Can be used with complex transformations to focus on an image
  - Affine and Piece-Wise Affine, Perspective, This Plate Splines
- We can use it instead of grid based soft and hard attention for multi-modal tasks



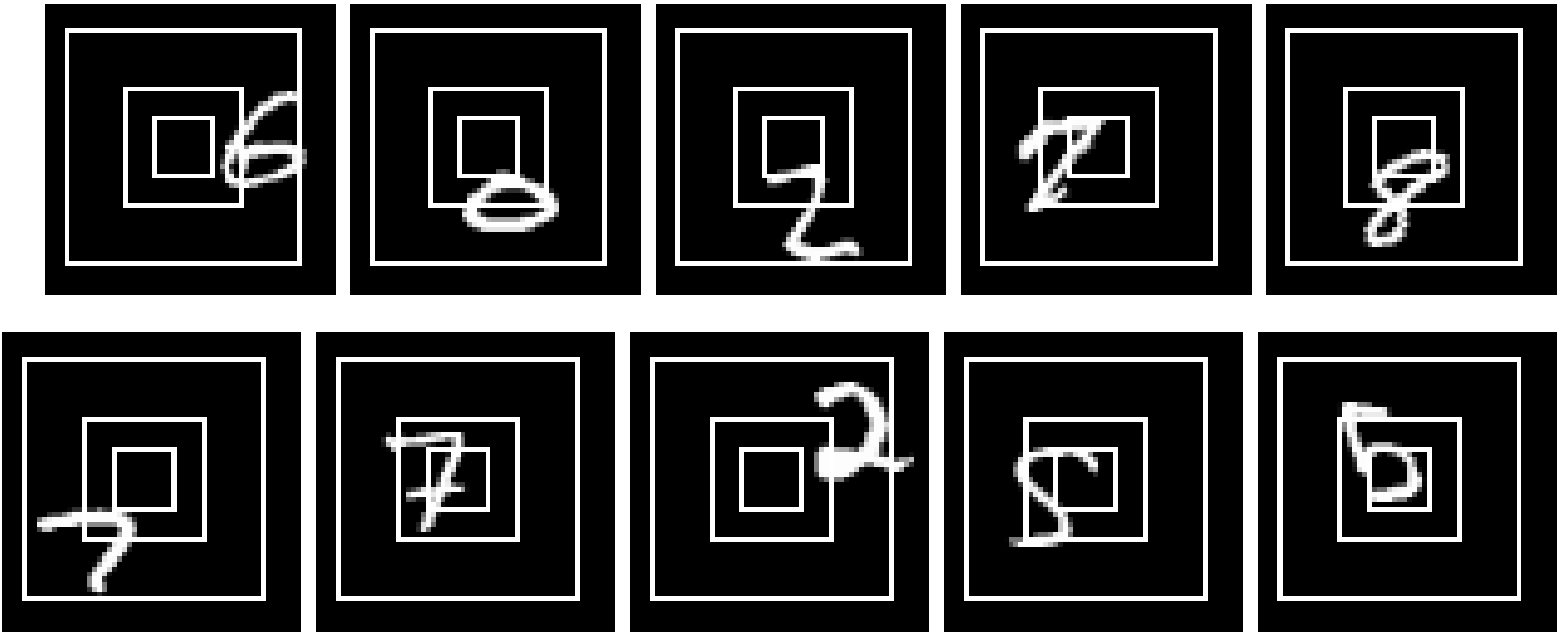
Glimpse Network (Hard Attention)

Hard attention

- Soft attention requires computing a representation for the whole image or sentence
- Hard attention on the other hand forces looking only at one part
- Main motivation was reduced computational cost rather than improved accuracy (although that happens a bit as well)
- **Saccade followed by a glimpse – how human visual system works**

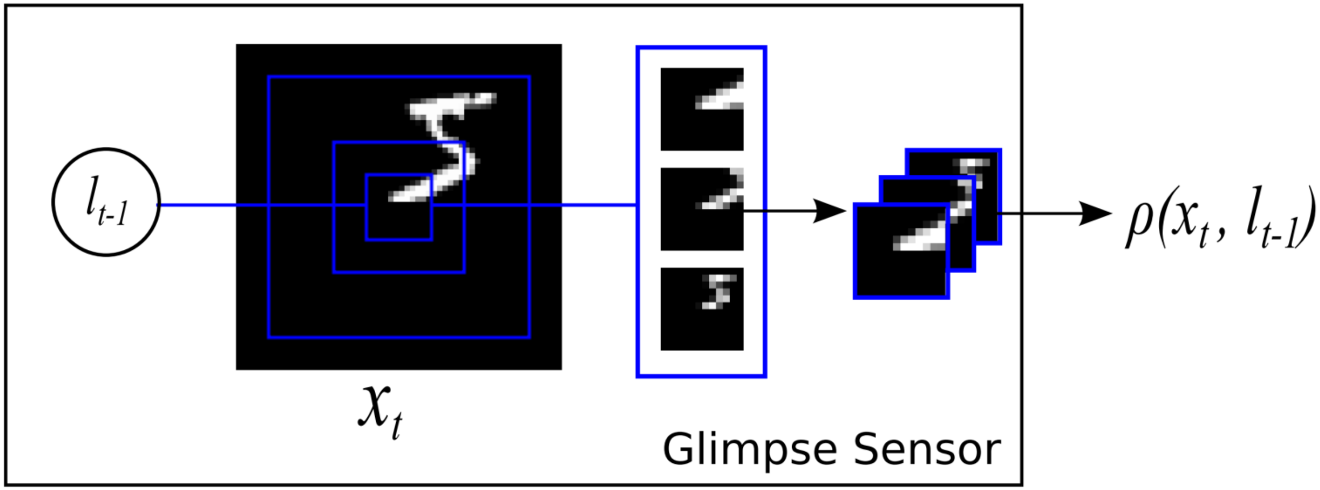
[Recurrent Models of Visual Attention, Mnih, 2014]
[Multiple Object Recognition with Visual Attention, Ba, 2015]

Hard attention examples



Glimpse Sensor

Looking at a part of an image at different scales

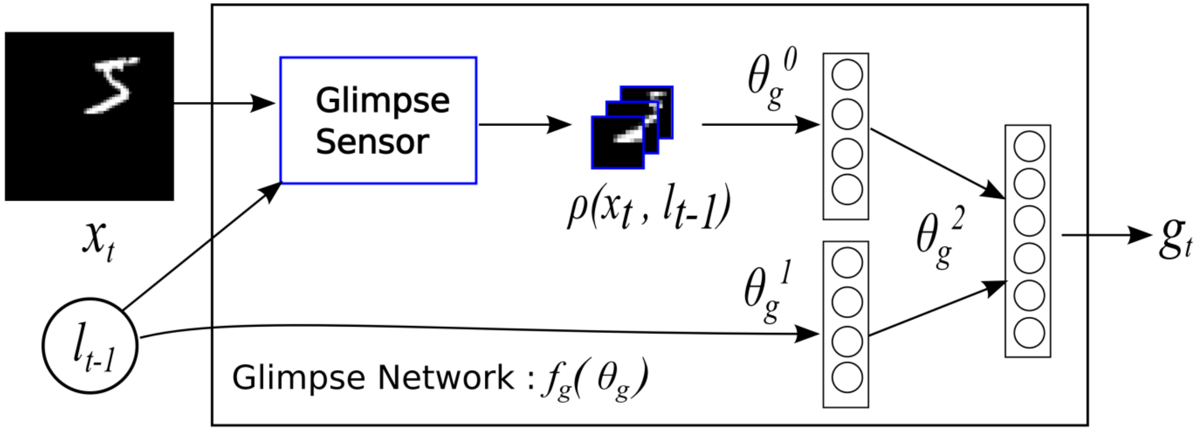


- At a number of different scales combined to a single multichannel image (human retina like representation)
- Given a location l_t output an image summary at that location

[Recurrent Models of Visual Attention, Mnih, 2014]

Glimpse network

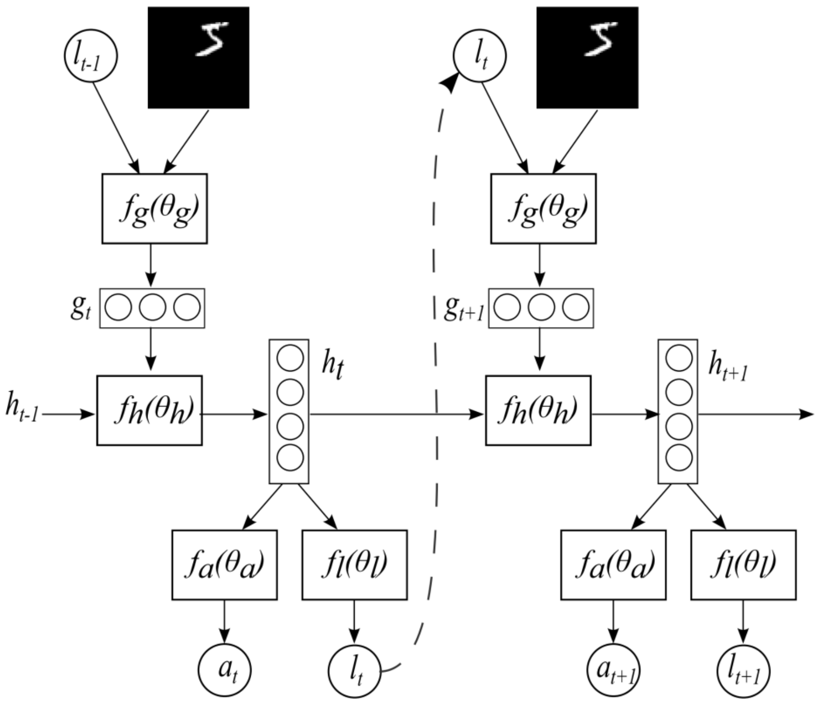
- Combining the Glimpse and the location of the glimpse into a joint network



- The glimpse is followed by a feedforward network (CNN or a DNN)
- The exact formulation of how the location and appearance are combined varies, the important thing is combining **what** and **where**
- Differentiable with respect to glimpse parameters but not the location

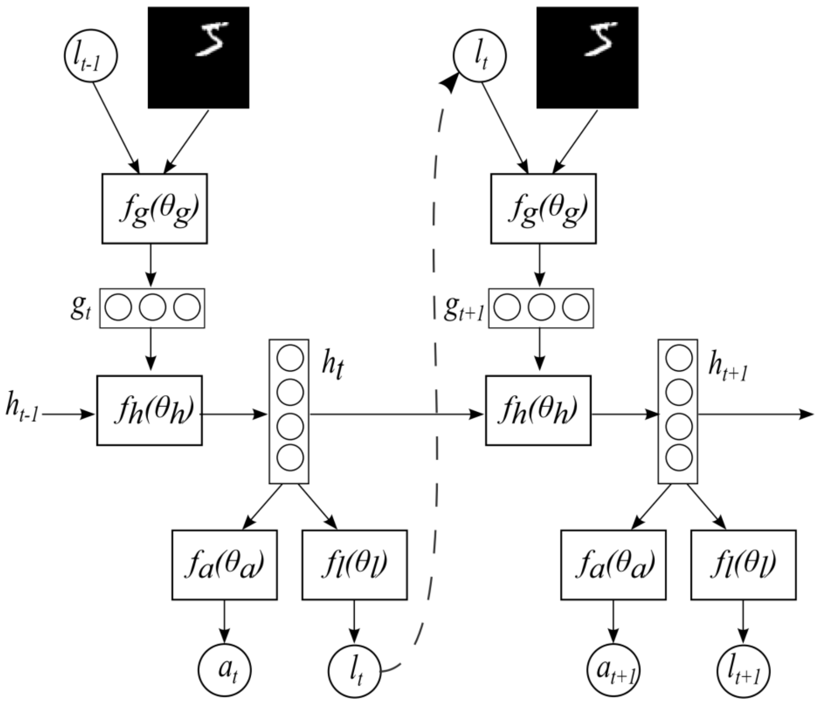
Overall Architecture - Emission network

- Given an image a glimpse location l_t, and optionally an action a_t
- Action can be:
  - Some action in a dynamic system – press a button etc.
  - Classification of an object
  - Word output
- This is an RNN with two output gates and a slightly more complex input gate!



Recurrent model of Visual Attention (RAM)

- Sample locations of glimpses leading to updates in the network
- Use gradient descent to update the weights (the glimpse network weights are differentiable)
- The emission network is an RNN
- Not as simple as backprop but doable
- Turns out this is very similar and in some cases equivalent to reinforcement learning using the REINFORCE learning rule [Williams, 1992]



Multi-modal alignment recap

Multimodal-alignment recap

- Explicit alignment - aligns two or more modalities (or views) as an actual task. The goal is to find correspondences between modalities
  - Dynamic Time Warping
  - Canonical Time Warping
  - Deep Canonical Time Warping
- Implicit alignment - uses internal latent alignment of modalities in order to better solve various problems
  - Attention models
  - Soft attention
  - Spatial transformer networks
  - Hard attention