Language Technologies Institute

Carnegie Mellon University

Advanced Multimodal Machine Learning

Lecture 9.1: Multimodal alignment

Louis-Philippe Morency

* Original version co-developed with Tadas Baltrusaitis

Upcoming Schedule

- First project assignment:
  - Proposal presentation (10/3 and 10/5)
  - First project report (10/8)
- Midterm project assignment
  - Midterm presentations (Tuesday 11/7 & Thursday 11/9)
  - Midterm report (Sunday 11/12)
- Final project assignment
  - Final presentation (12/4 & 12/5)
  - Final report (12/10)

Midterm Presentation Instructions

- 8-9 minute presentations (max: 10 mins)
  - +2 minutes for questions
  - +3 minutes for written feedback and notes
- All team members should be involved.
- The ordering of the presentations (Tuesday vs. Thursday) will be inverted for final presentations.
- The presentations will be from 4:30pm – 6pm:
  - Tuesday November 7: DH 1112
  - Thursday November 9: GHC 6115

Midterm Presentation Instructions

- Motivation and general definition of your research problem (2-3 slides)
- Mathematical formalization of the research problem, including definition of the main variables and overall objective function (1-3 slides)
- Explain at least one multimodal baseline model for your research problem (2-4 slides)
- Present current results of this baseline model on your dataset. You should study the failure cases of the baseline model (3-5 slides)
- Describe the research directions you are planning to explore. Discuss how they will address some of the shortcoming of your baseline model. (2-3 slides)

Midterm Project Report Instructions

- PART 1
  - **Research problem:** describe and motivate the research problem you are planning to work on. Explain why this problem is important for the research community and, if possible, the society in general. Define in generic terms the main computational challenges involved in this research problem.
  - **Related Work:** Present an overview of the work happening in this research area. This section should include about 12-15 citations of prior work, grouped in similar topics. Also, you should present in more details the 3-4 research papers most related to your proposed work. The related work section should end emphasizing how your proposed approach differ from previous work.
  - **Dataset and Input Modalities:** Describe the dataset(s) you are planning to use for this project. If many options exist, please motivate your choice of dataset for this research problem. Describe the input modalities and annotations available in this dataset. Specify which subset of these modalities and annotations you are planning to use.

Midterm Project Report Instructions

- PART 2
  - **Problem statement:** formalize mathematically your research problem. This should include the mathematical definition of the variables involved in your problem.
  - **Multimodal baseline models:** Describe mathematically at least one multimodal baseline model for your research problem.
  - **Experimental methodology:** Describe your experimental methodology for evaluating the multimodal baseline model(s).
  - **Results and Discussion:** Present in tables and/or figures your experimental results. This section should include more than re-running existing baseline models.
  - **Proposed approach:** Describe what models you are planning to test for the final report experiments. Whenever possible, you should write down the loss function of these models, following the same mathematical formulation previously used.

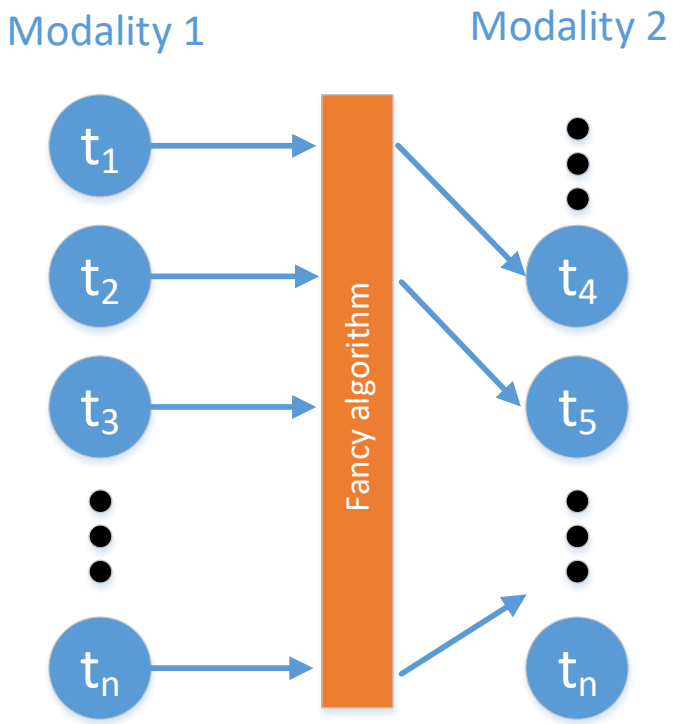
Lecture objectives

- 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

Multi-modal alignment

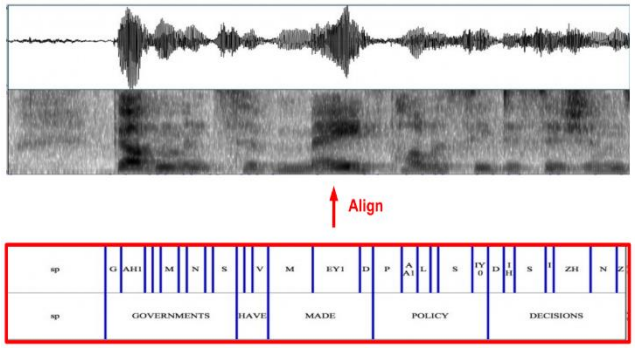
Multimodal-alignment

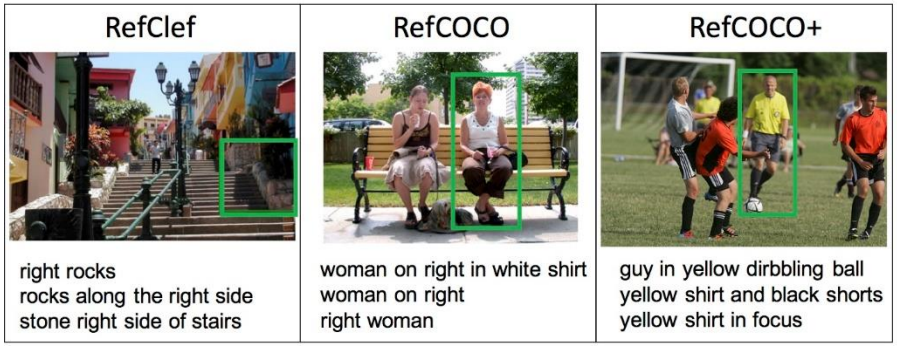
- Multimodal alignment – finding relationships and correspondences between two or more modalities
- Examples
  - Images with captions
  - Recipe steps with a how-to video
  - Phrases/words of translated sentences
- Two types
  - Explicit – alignment is the task in itself
  - Latent – alignment helps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



Explicit multimodal-alignment

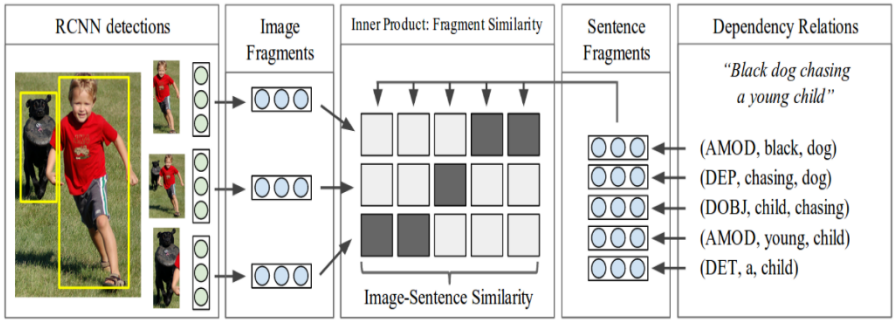
- Explicit alignment - goal is to find correspondences between modalities
  - Aligning speech signal to a transcript
  - Aligning two out-of sync sequences
  - Co-referring expressions

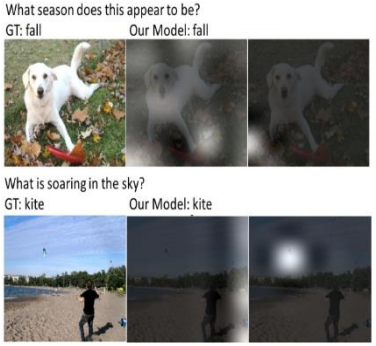




Implicit multimodal-alignment

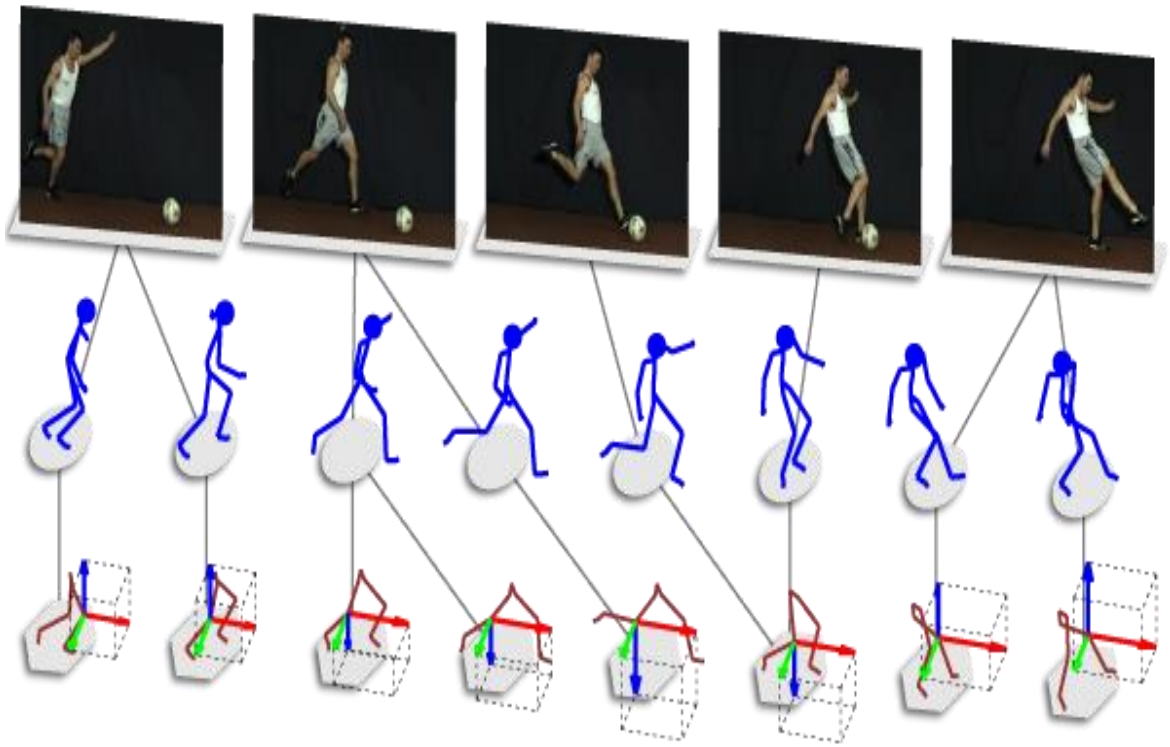
- Implicit alignment - uses internal latent alignment of modalities in order to better solve various problems
  - Machine Translation
  - Cross-modal retrieval
  - Image & Video Captioning
  - Visual Question Answering





Explicit alignment

Temporal sequence alignment



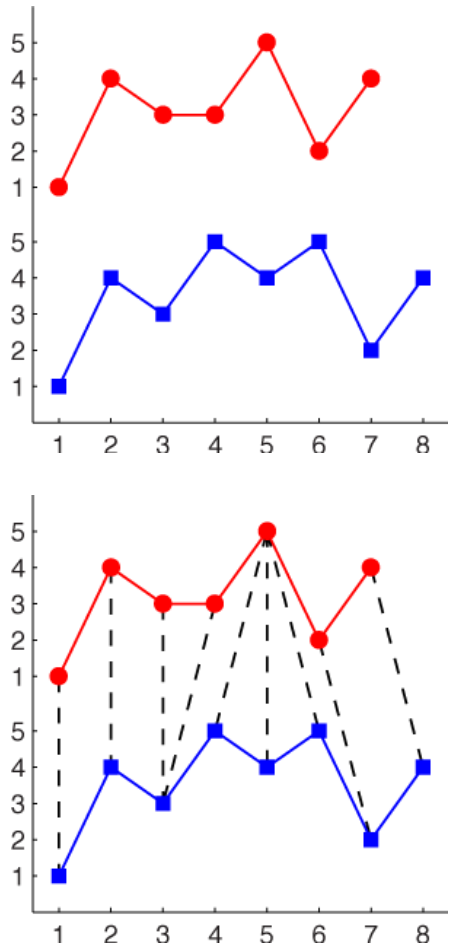
Applications:
- Re-aligning asynchronous data
- Finding similar data across modalities (we can estimate the aligned cost)
- Event reconstruction from multiple sources

Let's start unimodal – Dynamic Time Warping

- We have two unaligned temporal unimodal signals
  - $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_x}] \in \mathbb{R}^{d \times n_x}$
  - $\mathbf{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{n_y}] \in \mathbb{R}^{d \times n_y}$
- Find set of indices to minimize the alignment difference:

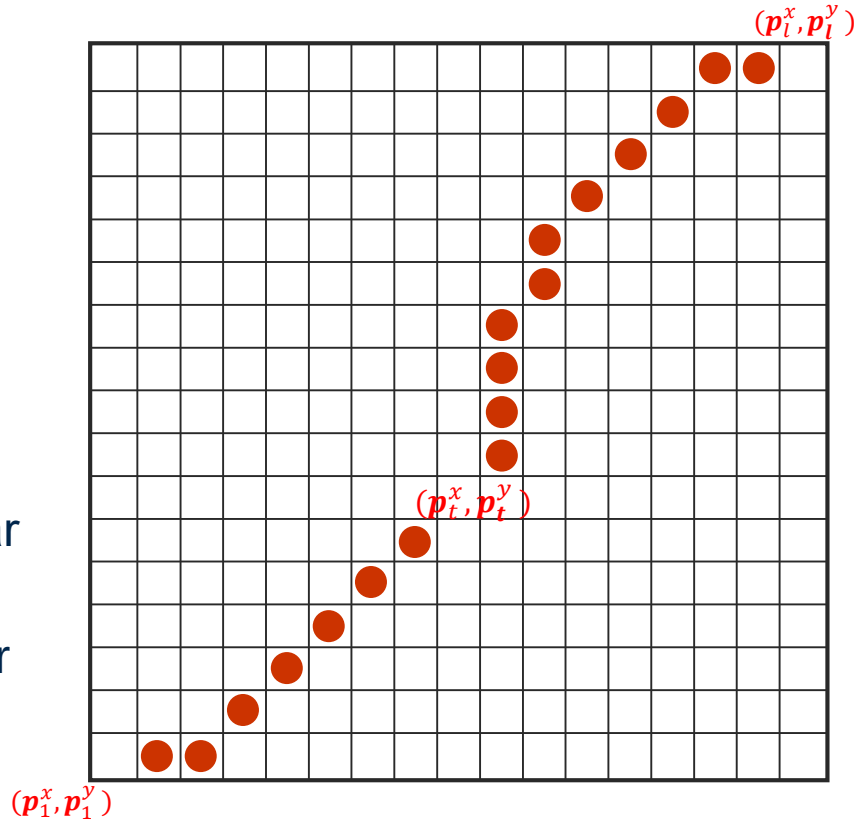
$$L(\boldsymbol{p}_t^x, \boldsymbol{p}_t^y) = \sum_{t=1}^{l} \left\| \boldsymbol{x}_{\boldsymbol{p}_t^x} - \boldsymbol{y}_{\boldsymbol{p}_t^y} \right\|_2^2$$

- Where $\boldsymbol{p}^x$ and $\boldsymbol{p}^y$ are index vectors of same length
- Finding these indices is called Dynamic Time Warping



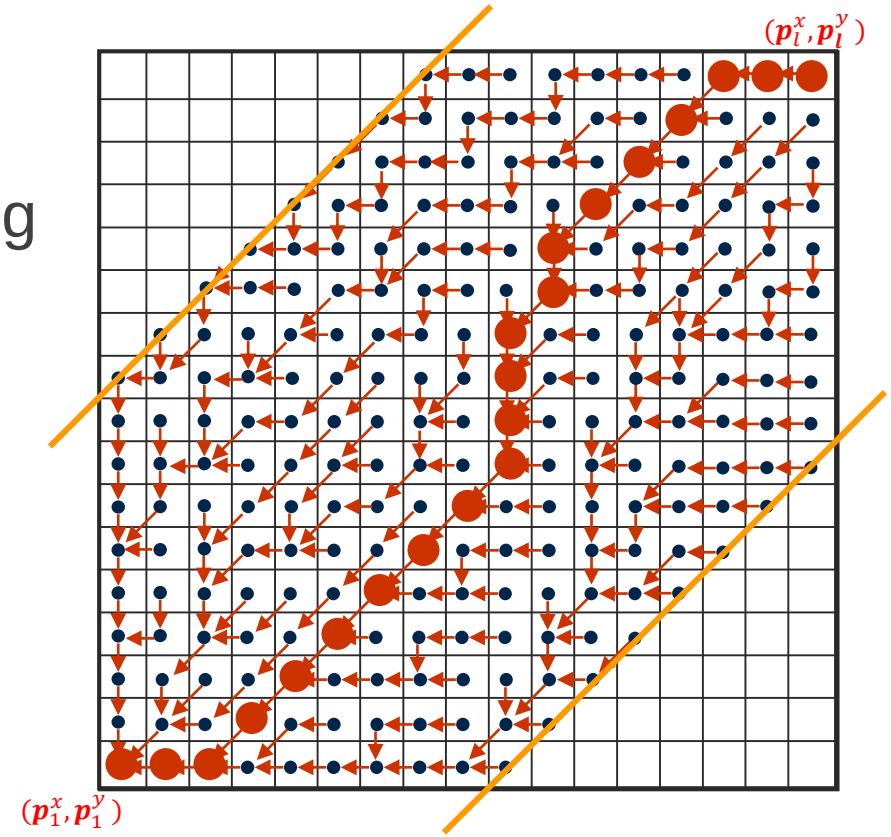
Dynamic Time Warping continued

- 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

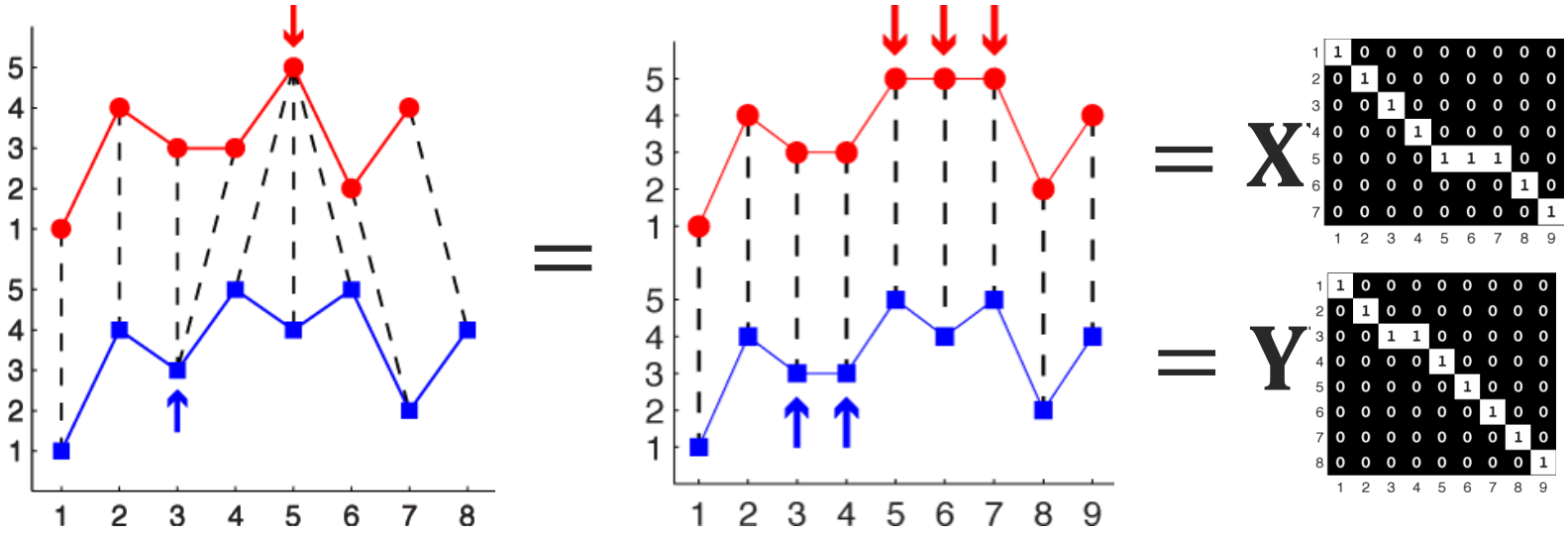
- Lowest cost path in a cost matrix
- Solved using dynamic programming whilst respecting the restrictions



DTW alternative formulation

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

Replication doesn't change the objective!



Alternative objective:

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

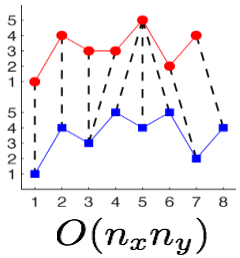
$\boldsymbol{X}, \boldsymbol{Y}$ – original signals (same #rows, possibly different #columns)

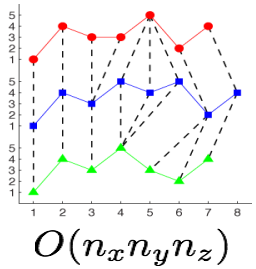
$\boldsymbol{W}_x, \boldsymbol{W}_y$ - alignment matrices

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

DTW - limitations

- Computationally complex





m sequences

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

- Sensitive to outliers

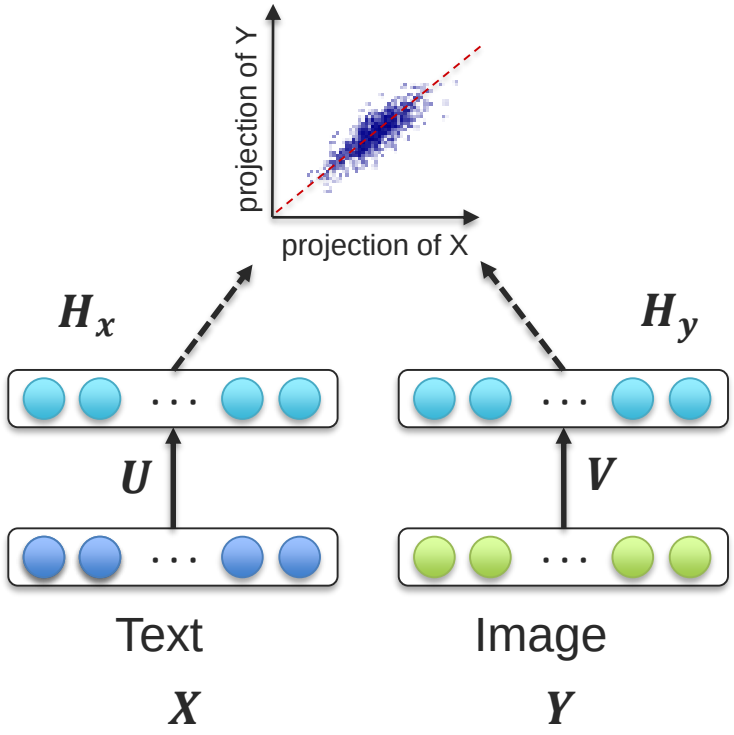
- Unimodal!

Canonical Correlation Analysis reminder

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

subject to: $\boldsymbol{U^T \Sigma_{YY} U = V^T \Sigma_{YY} V = I}$

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

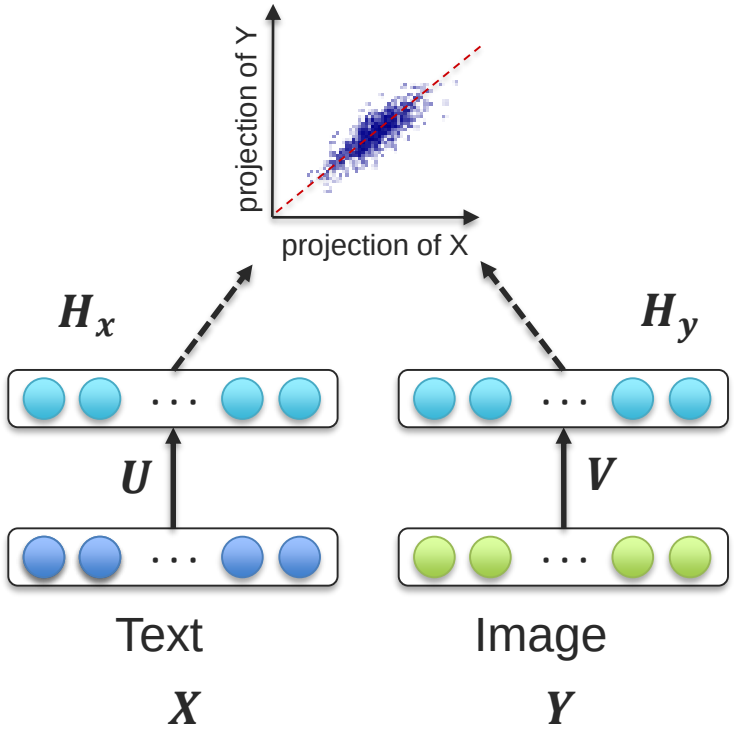


Canonical Correlation Analysis reminder

- When data is normalized it is actually equivalent to smallest RMSE reconstruction
- CCA loss can also be re-written as:

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

subject to: $\boldsymbol{U^T\Sigma_{YY}U = V^T\Sigma_{YY}V = I}$



Canonical Time Warping

- Dynamic Time Warping + Canonical Correlation Analysis = Canonical Time Warping

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

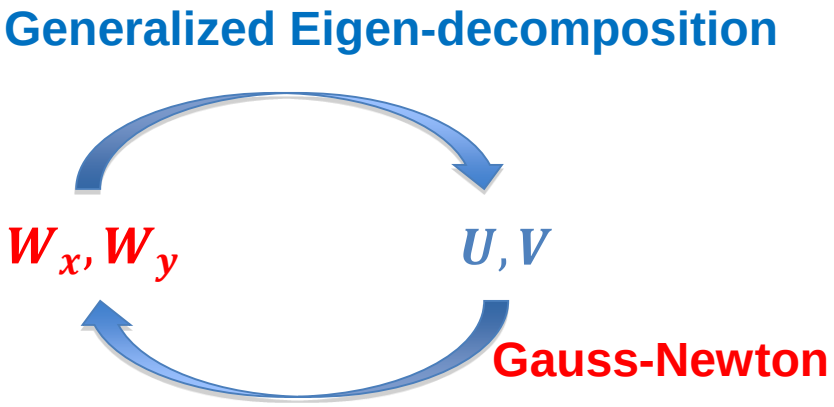
- Allows to align multi-modal or multi-view (same modality but from a different point of view)
- $\boldsymbol{W_x}, \boldsymbol{W_y}$ – temporal alignment
- $\boldsymbol{U}, \boldsymbol{V}$ – cross-modal (spatial) alignment

[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



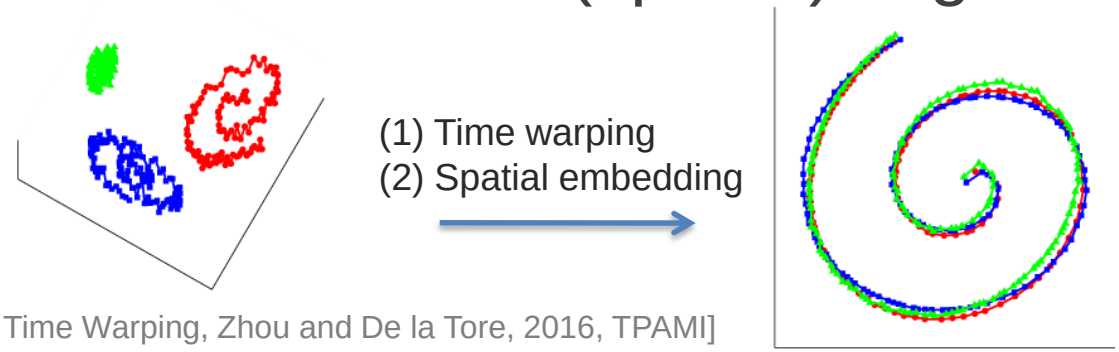
[Canonical Time Warping for Alignment of Human Behavior, Zhou and De la Tore, 2009, NIPS]

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

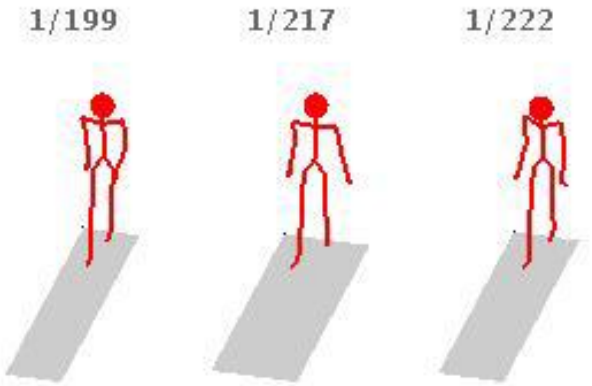


[Generalized Canonical Time Warping, Zhou and De la Tore, 2016, TPAMI]

Alignment examples (unimodal)

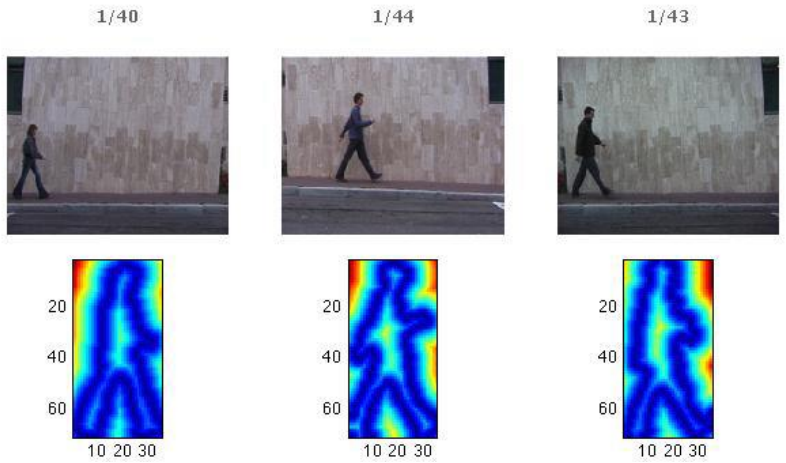
CMU Motion Capture

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



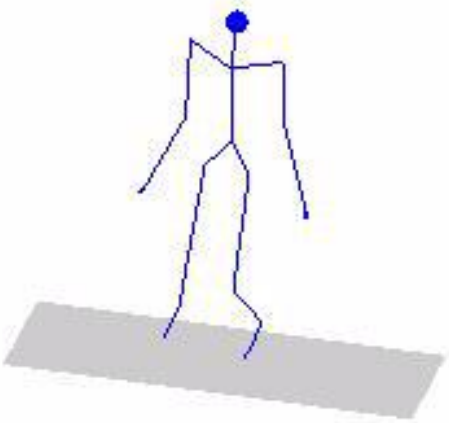
Weizmann

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



Alignment examples (multimodal)

1/273



1/51

1/127

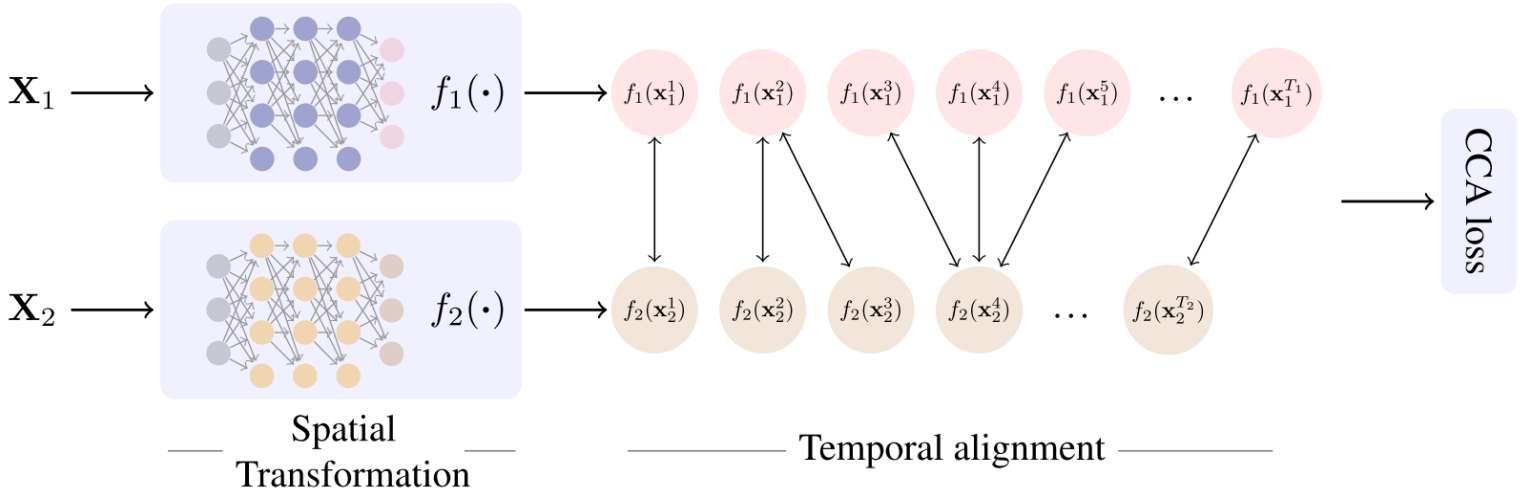
Canonical time warping - limitations

- Linear transform between modalities
- How to address this?

Deep Canonical Time Warping

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

- Could be seen as generalization of DCCA and GTW



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

Deep Canonical Time Warping

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

- 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

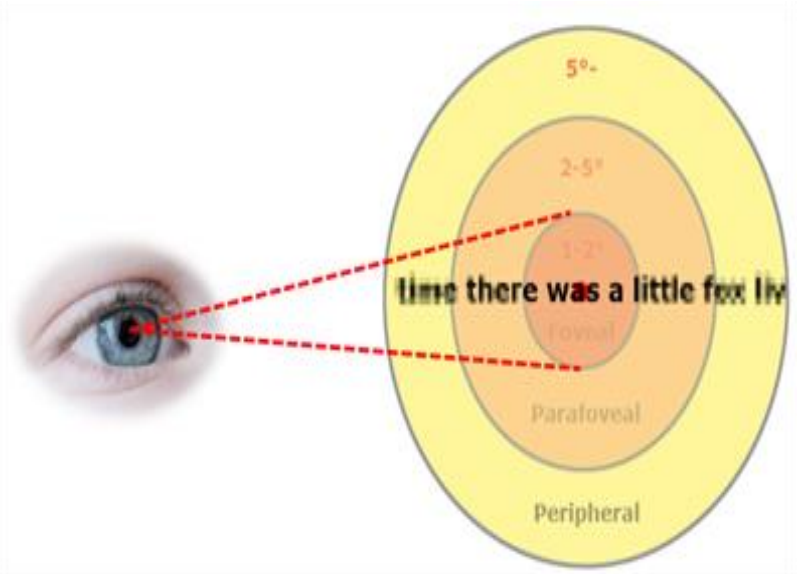
Implicit alignment

- We looked how to explicitly align temporal data
- Could use that as a pre-processing step in our pipelines
- Can we instead allow/encourage the model to align data when solving a particular 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

- Lots of attention
- 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
- Soft attention
  - Acts like a gate function. Deterministic inference.
- Transform network
  - Warp the input to better align with canonical view
- Hard attention
  - Includes stochastic processes. Similar to reinforcement learning.

Soft attention

Machine Translation

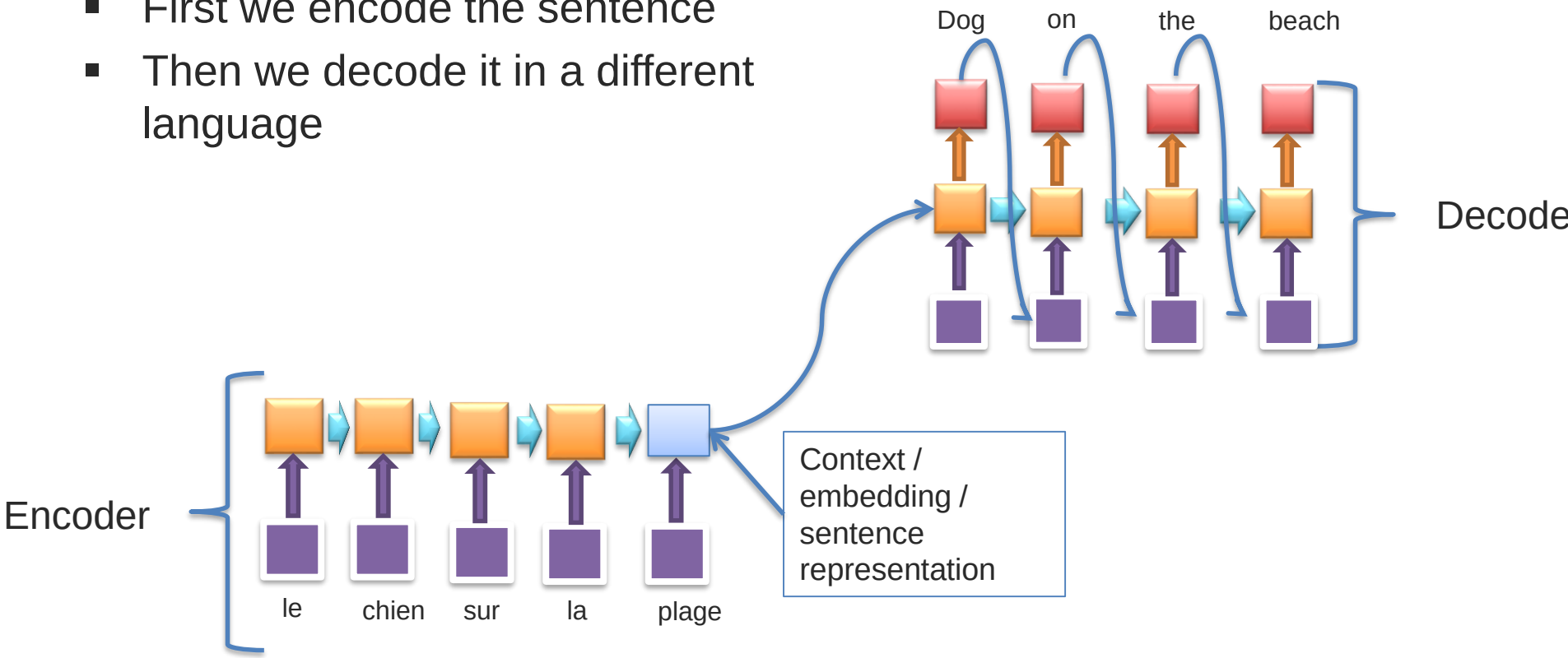
- Given a sentence in one language translate it to another

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

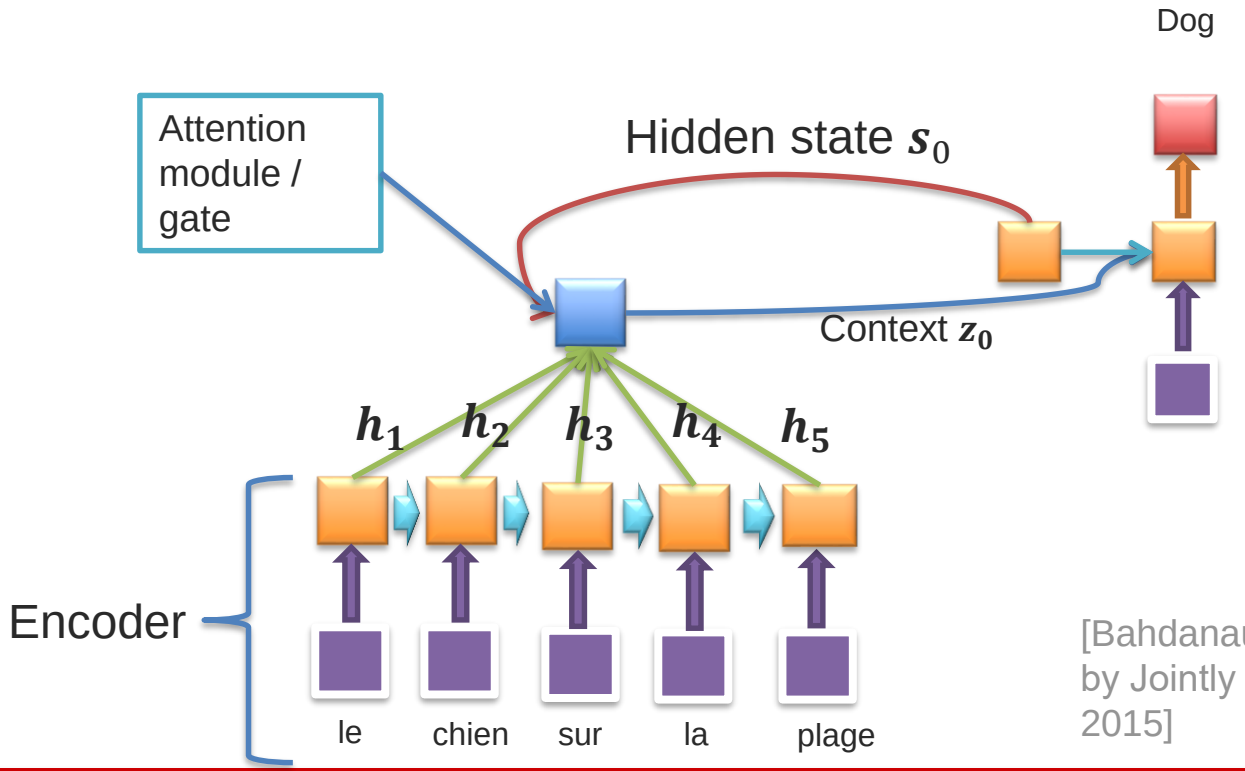
- What is the problem with this?
- What happens when the sentences are very long?
- We expect the encoders hidden state to capture everything in a sentence, a very complex state in a single vector, such as

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

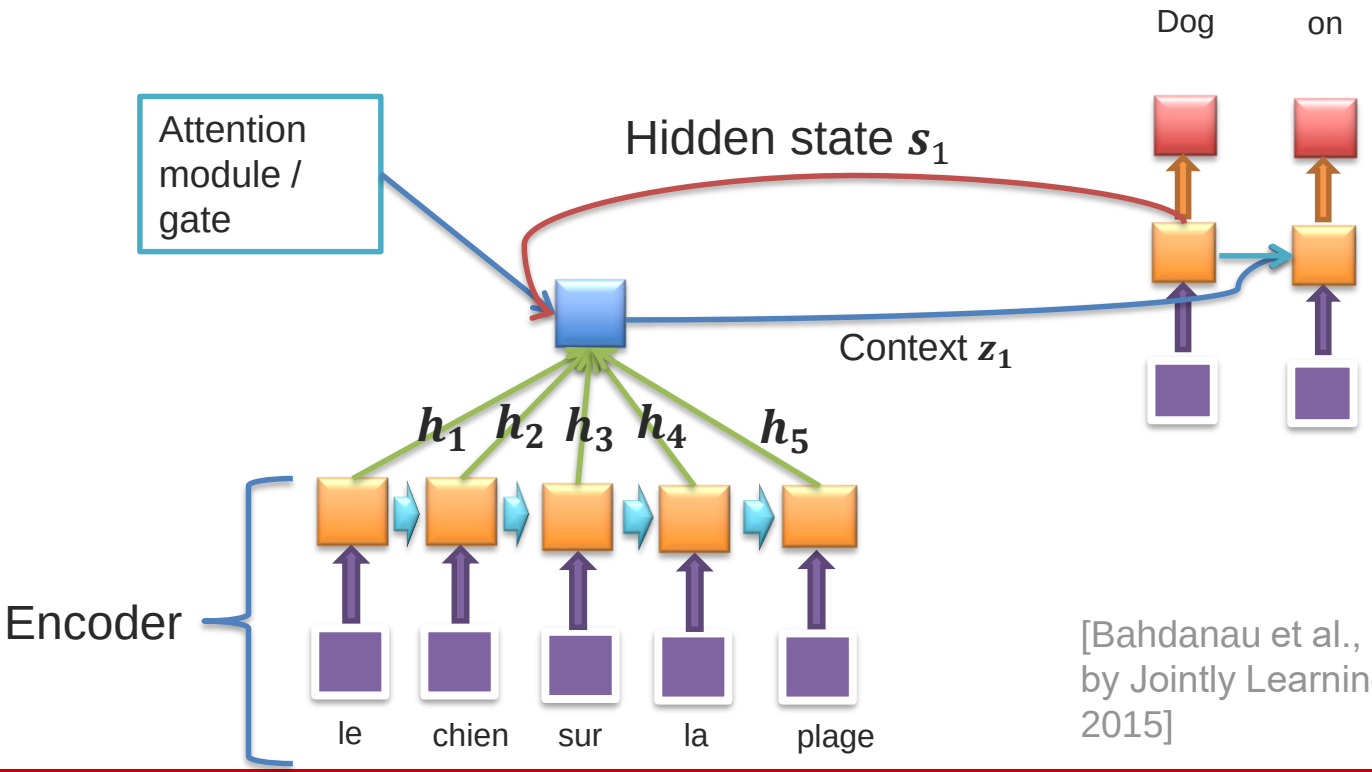
- Before encoder would just take the final hidden state, now we actually care about the intermediate hidden states



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

Decoder – attention model

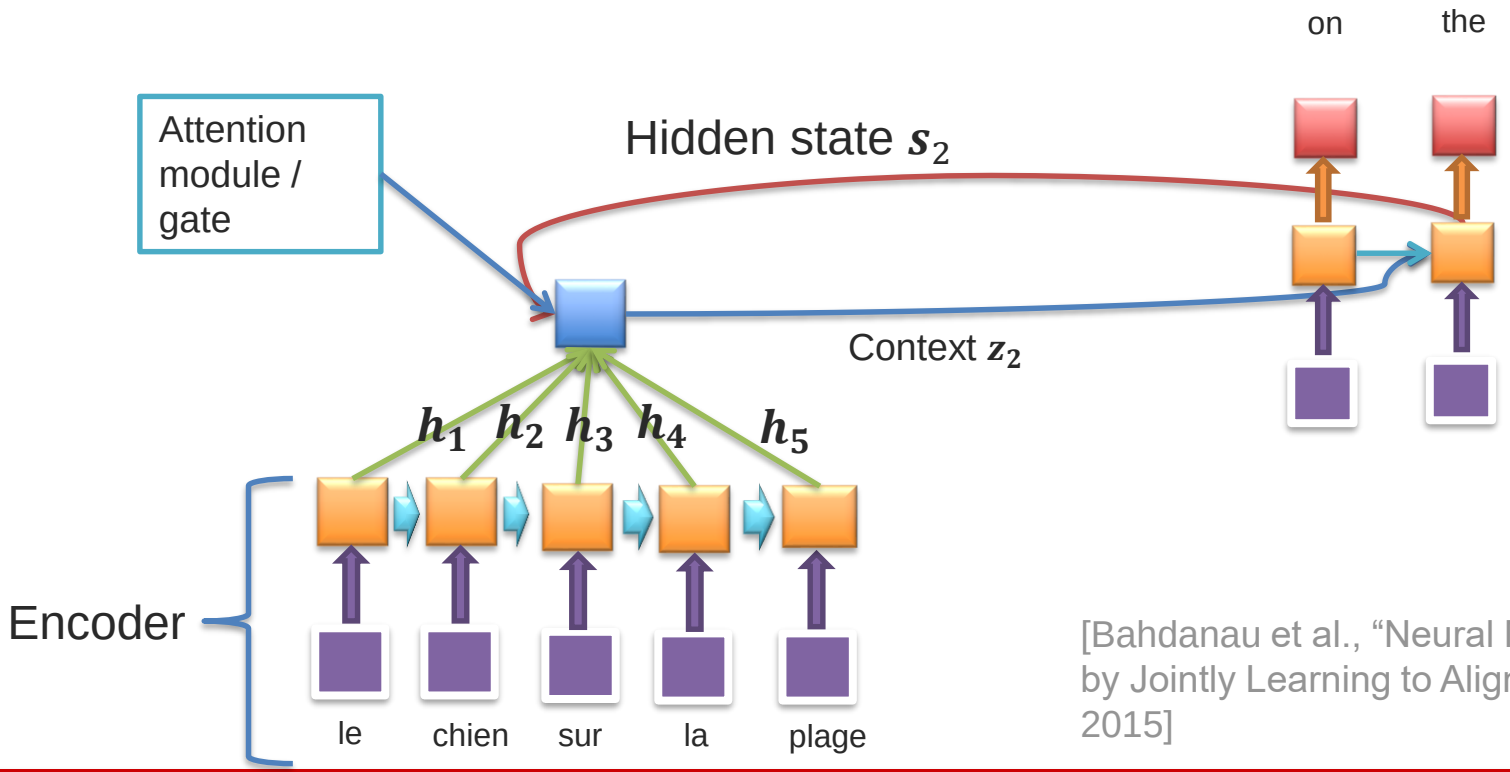
- Before encoder would just take the final hidden state, now we actually care about the intermediate hidden states



[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:
  - $p(y_i|y_1, \ldots, y_{i-1}, \boldsymbol{x}) = g(y_{i-1}, \boldsymbol{s}_i, \boldsymbol{z})$, where $\boldsymbol{z} = \boldsymbol{h}_T$, and $\boldsymbol{s}_i$ - the current state of the decoder
- Now:
  - $p(y_i|y_1, \ldots, y_{i-1}, \boldsymbol{x}) = g(y_{i-1}, \boldsymbol{s}_i, \boldsymbol{z}_i)$
- 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} - the (scalar) attention for word j at generation step i

MT with attention

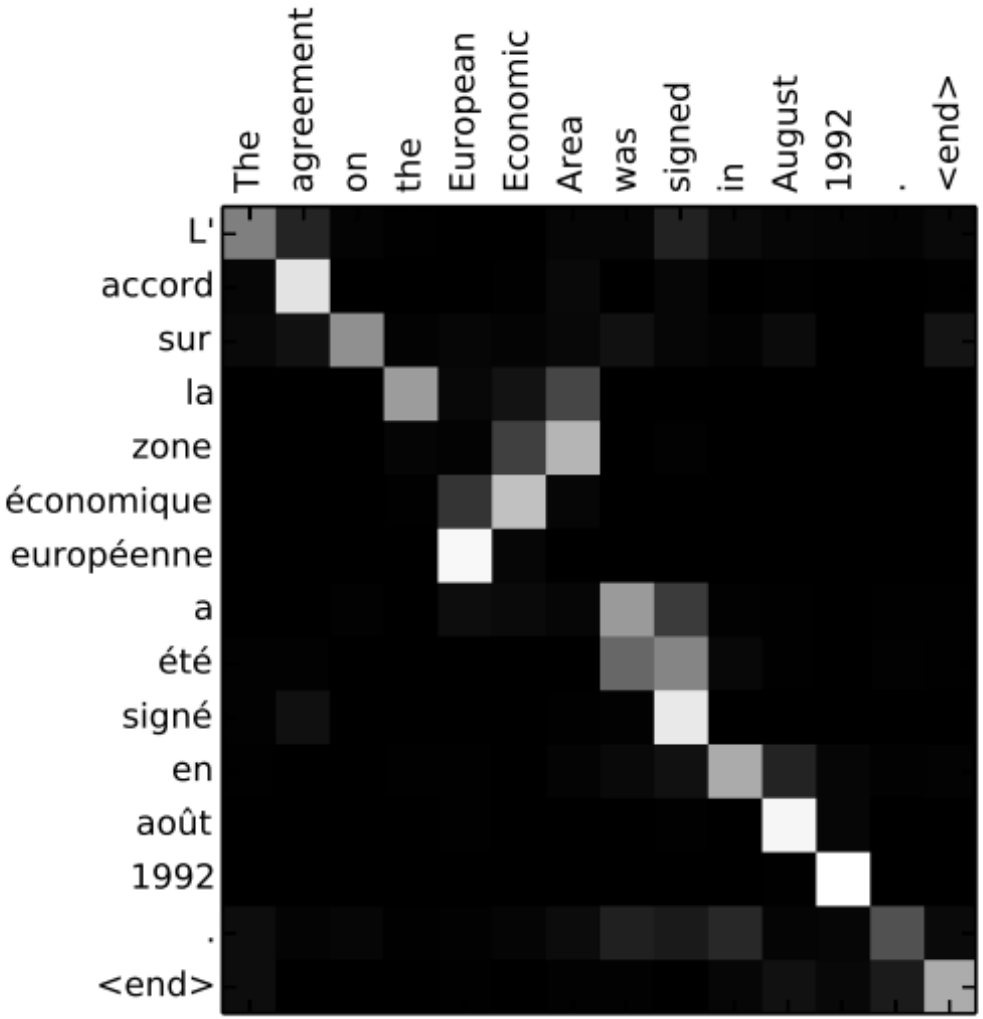
- So how do we determine α_{ij},
  - $\alpha_{i,j} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$ - softmax, making sure they sum to 1
- Where:
  - $e_{ij} = \boldsymbol{v}^T \sigma\left(W s_{i-1} + U h_j\right)$
  - a feedforward network that can tell us given the current state of decoder how 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!
- 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>

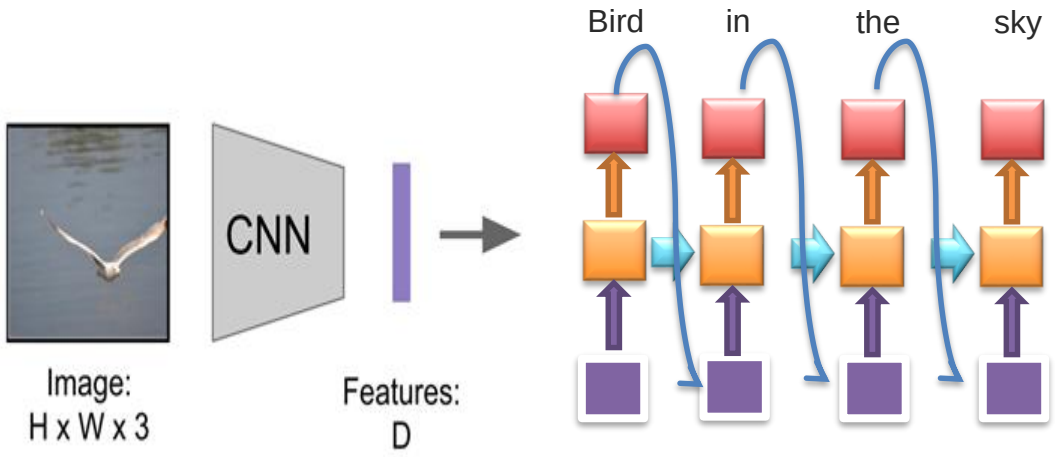
MT with attention recap

- Get 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

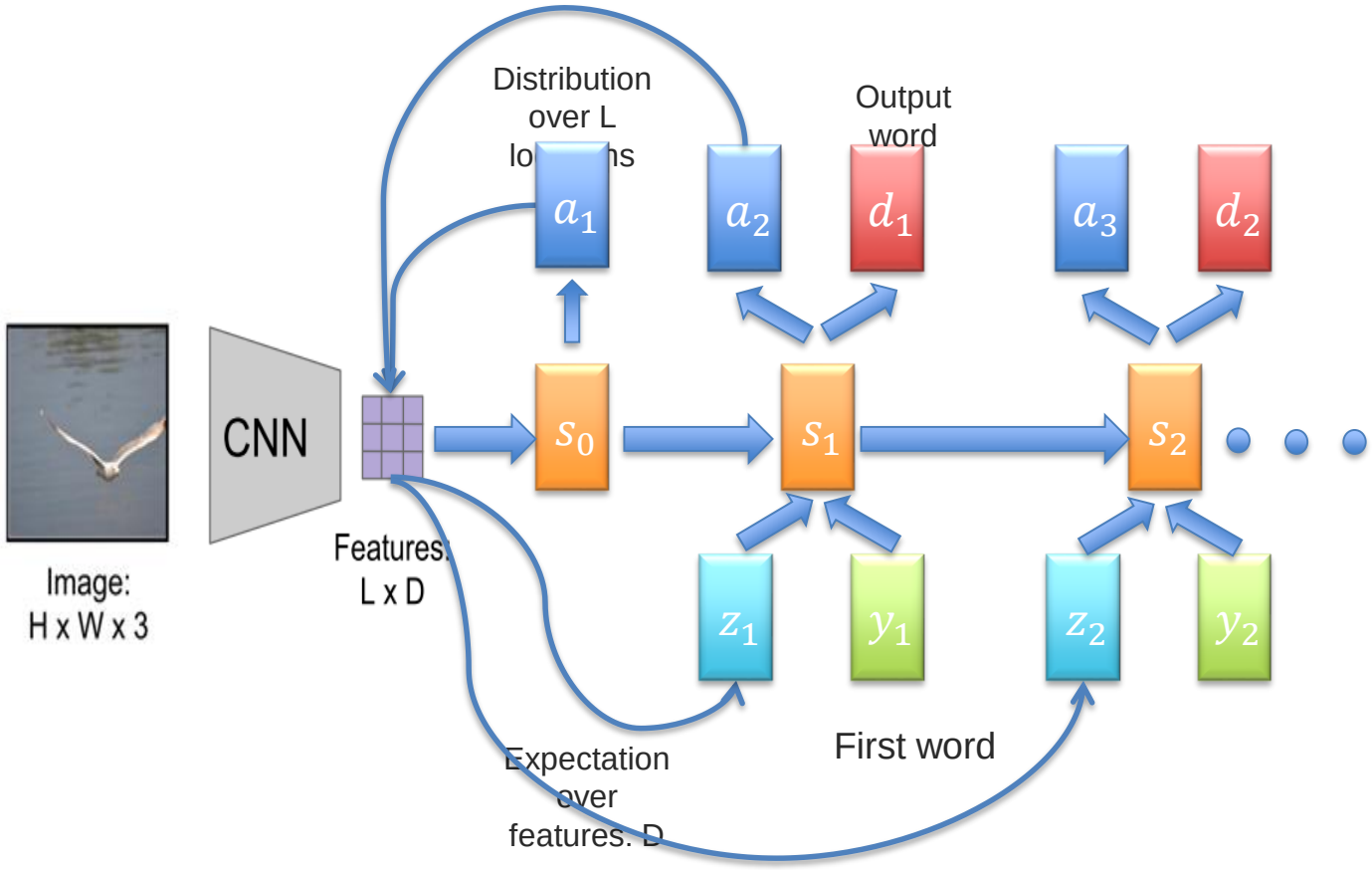
[Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, Xu et al., 2015]

Recap RNN for Captioning



Why might we not want to focus on the final layer?

Looking at more fine grained features



Soft attention

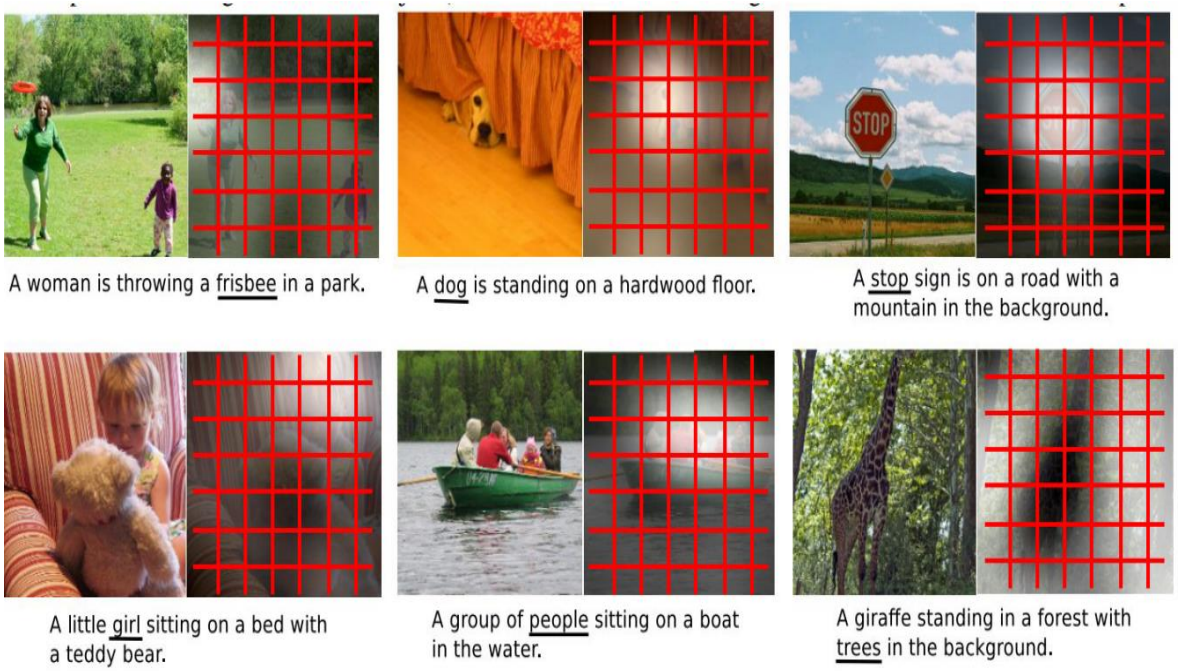
- Allows for latent data alignment
- Allows us to get an idea of what the network “sees”
- Can be optimized using back propagation

- 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

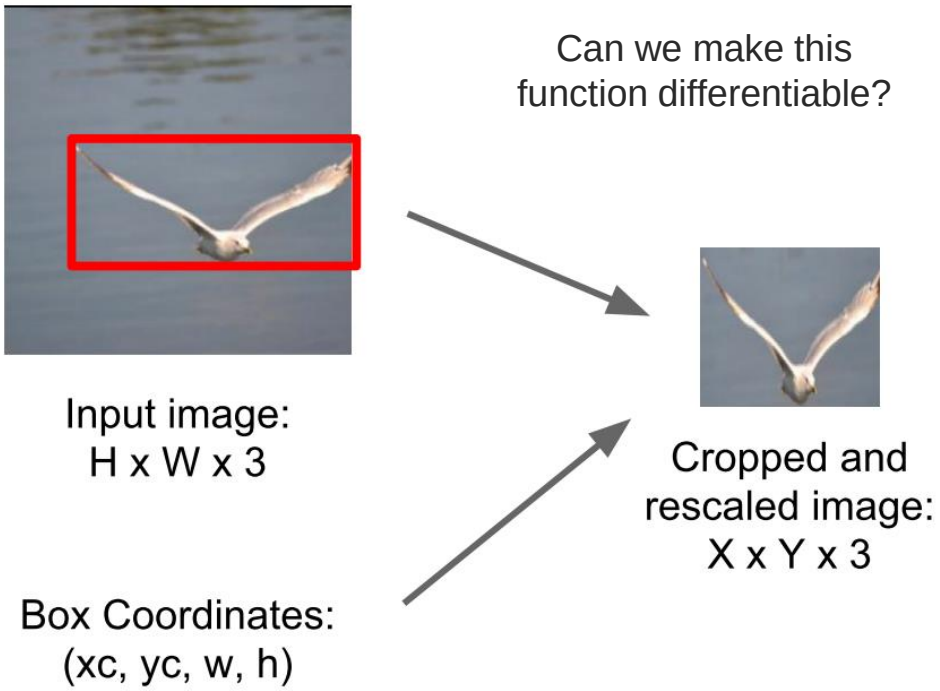
Spatial Transformer networks

Some limitations of grid based attention

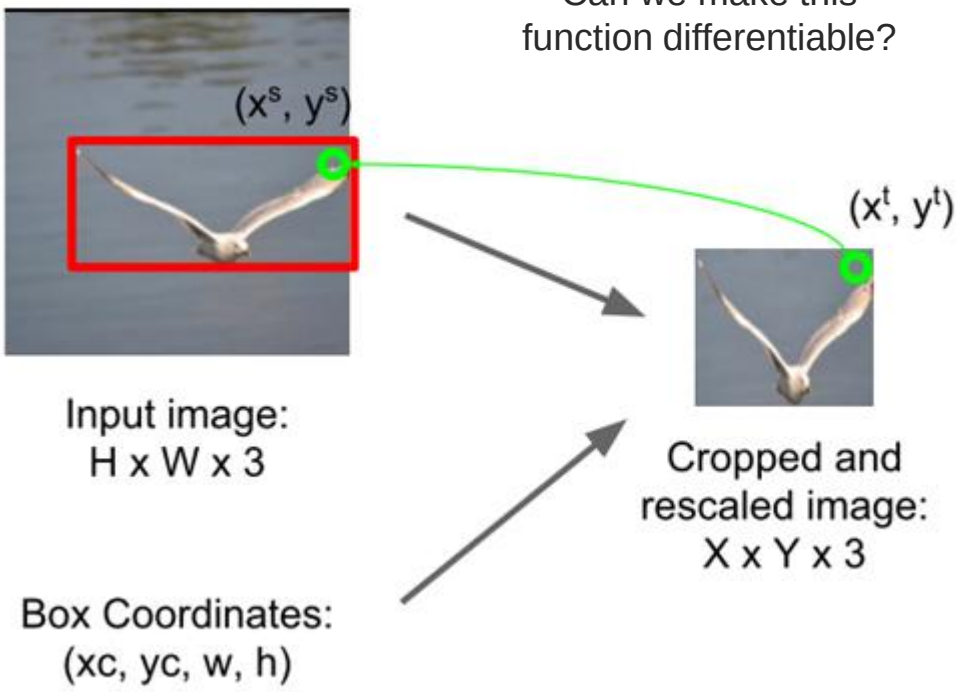
- Limited to a fixed grid analysis, glimpses solve this a bit but it's difficult to train, can we fixate on small parts of image but still have easy end-to-end training?



Spatial Transformer Networks



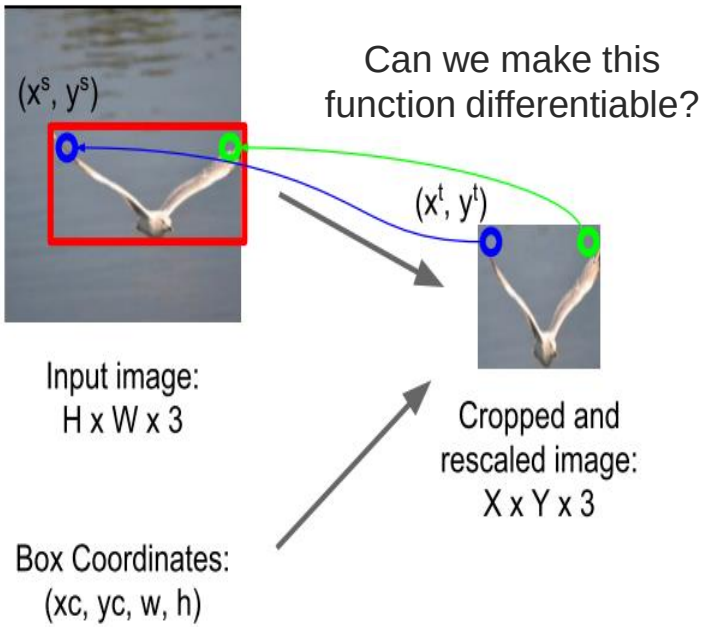
Spatial Transformer Networks



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

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

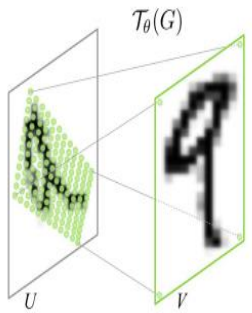
Spatial Transformer Networks



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

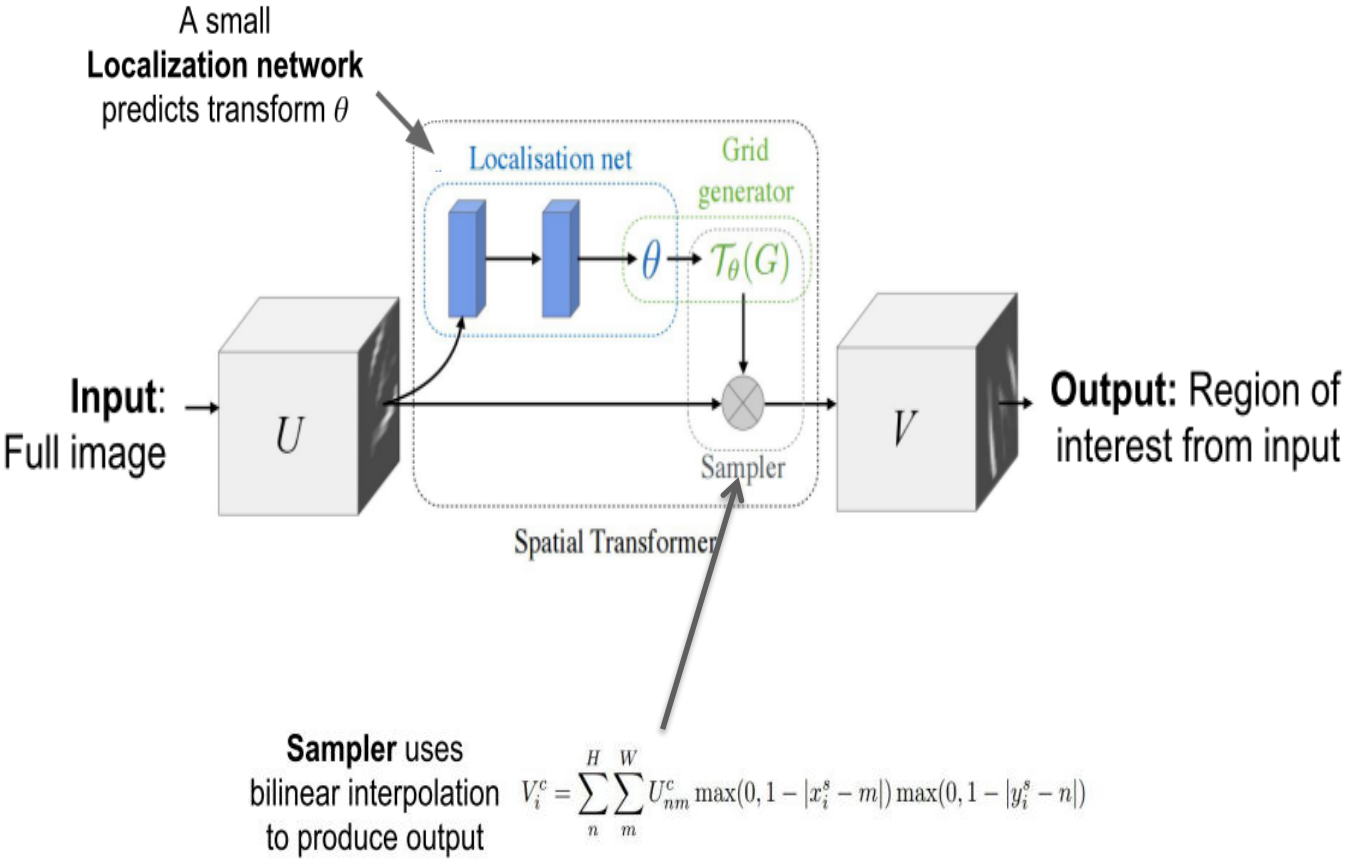
$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \theta_{1,3} \\ \theta_{2,1} & \theta_{2,2} & \theta_{2,3} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}$$

Network "attends" to input by predicting θ



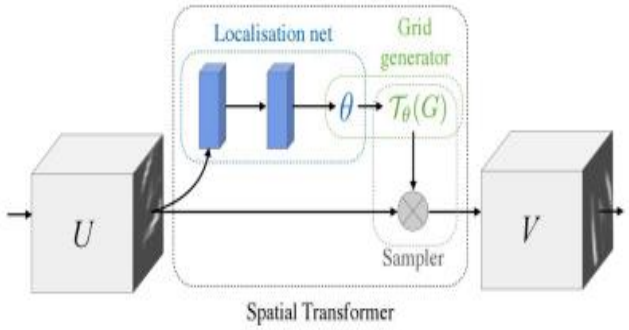
Repeat for all pixels in *output* to get a **sampling grid**

Spatial Transformer Networks



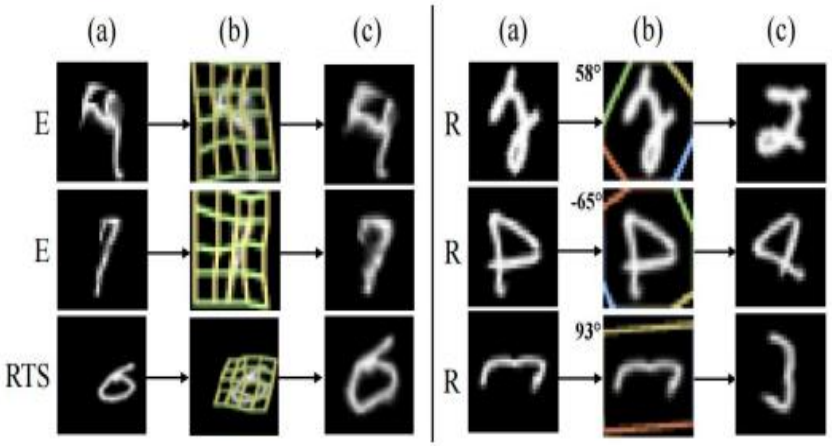
Spatial Transformer Networks

Differentiable "attention / transformation" module



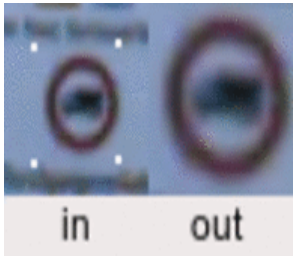
Spatial Transformer

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



Examples on real world data

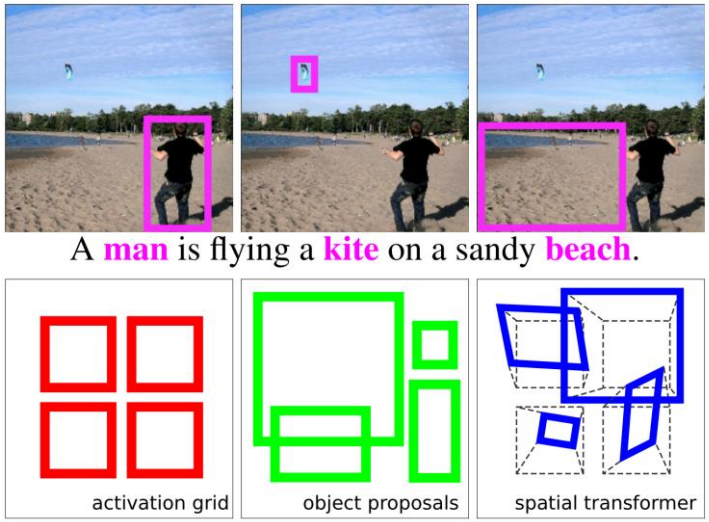
- State-of-the-art results on traffic sign recognition



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 use complex models for focusing on an image
  - Affine and Piece-Wise Affine, Perspective, This Plate Splines
- Can use to focus on certain parts of an image
- We can use it instead of grid based soft and hard attention for multi-modal tasks



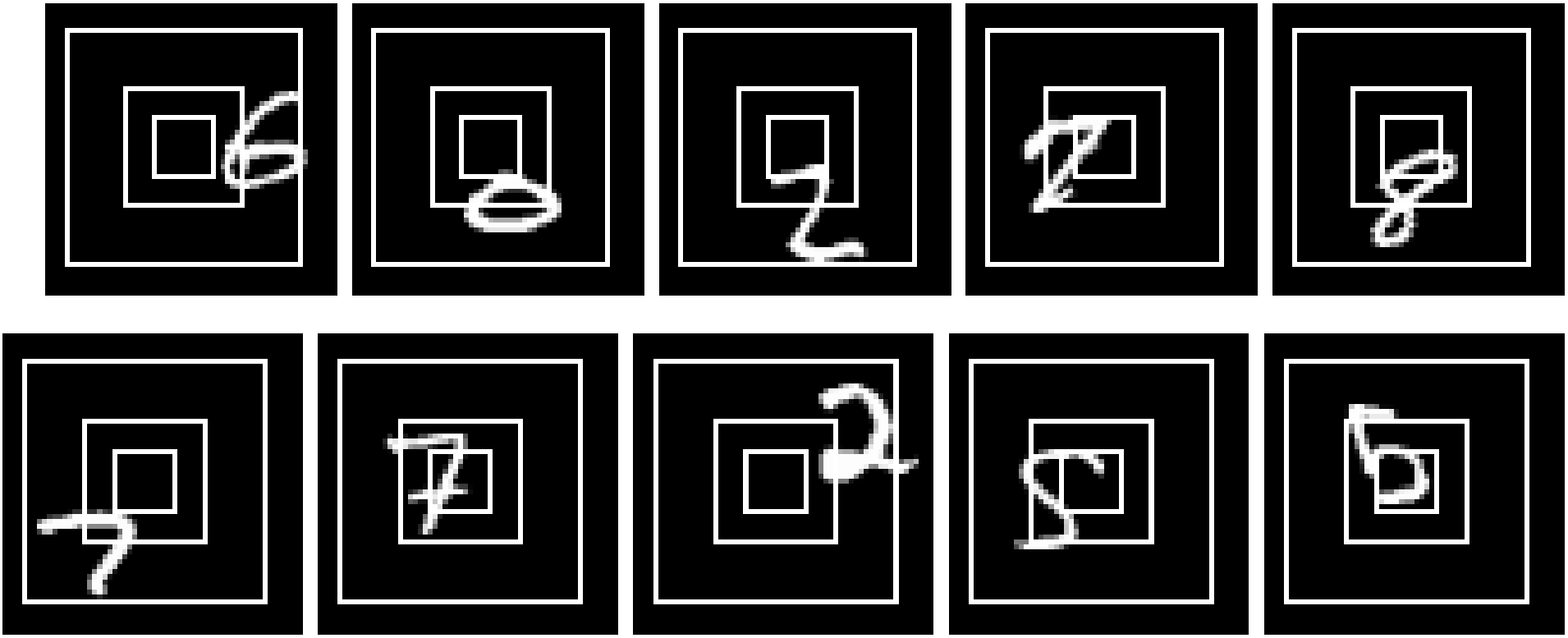
Glimpse Network (Hard Attention)

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**

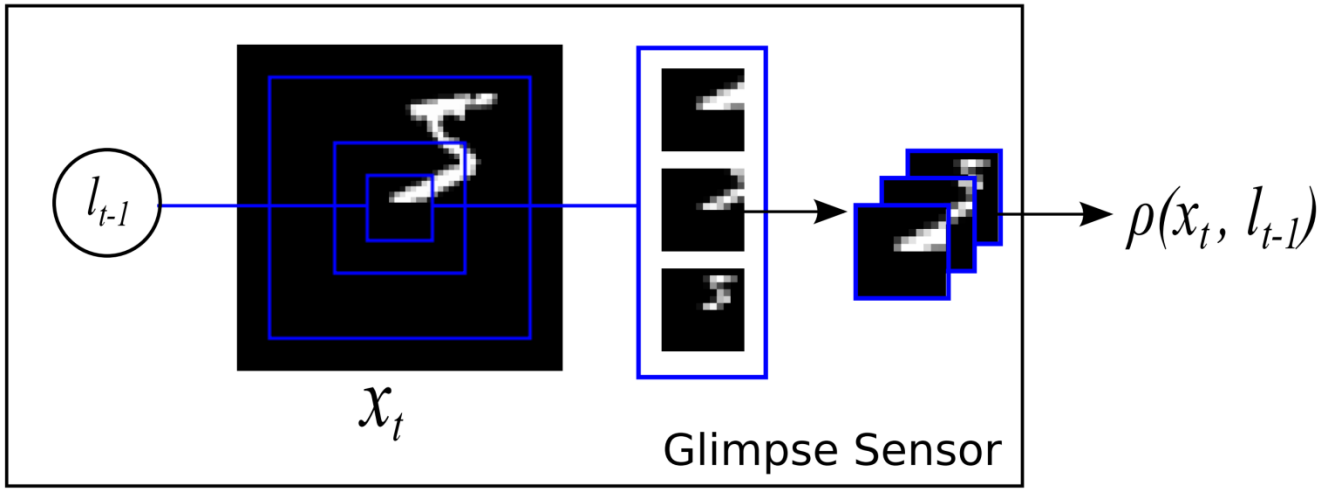
[Recurrent Models of Visual Attention, Mnih, 2014]
[Multiple Object Recognition with Visual Attention, Ba, 2015]

Hard attention examples



Glimpse Sensor

- Looking at a part of an image at different scales

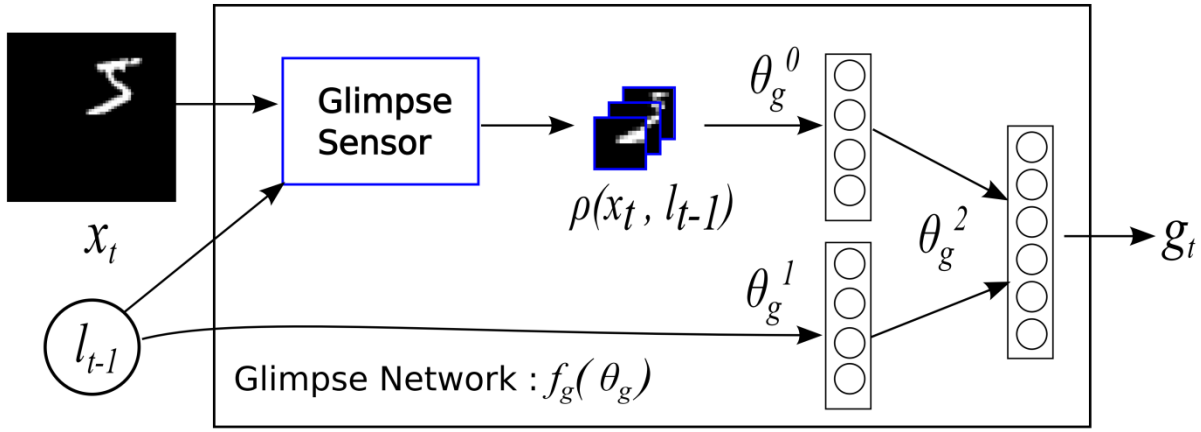


- At a number of different scales combined to a single multichannel image (human retina like representation)
- Given a location l_t output an image summary at that location

[Recurrent Models of Visual Attention, Mnih, 2014]

Glimpse network

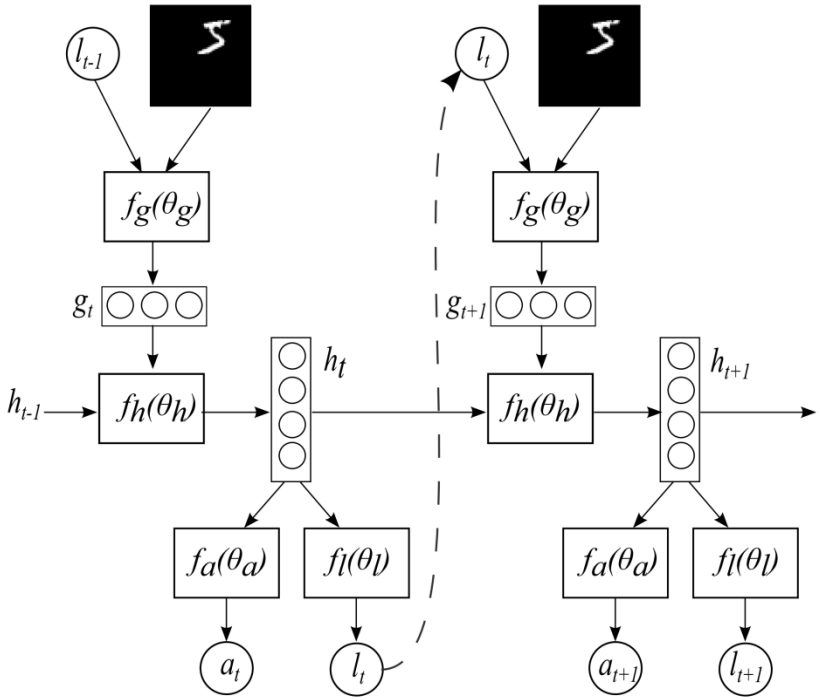
- Combining the Glimpse and the location of the glimpse into a joint network



- 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)

- Model definition

$$L = \log[p(y|X, W)] = \log \sum_{l} p(l|X, W)p(y|l, X, W)$$

W – set of parameters of the RNN $(\theta_g, \theta_l, \theta_a)$

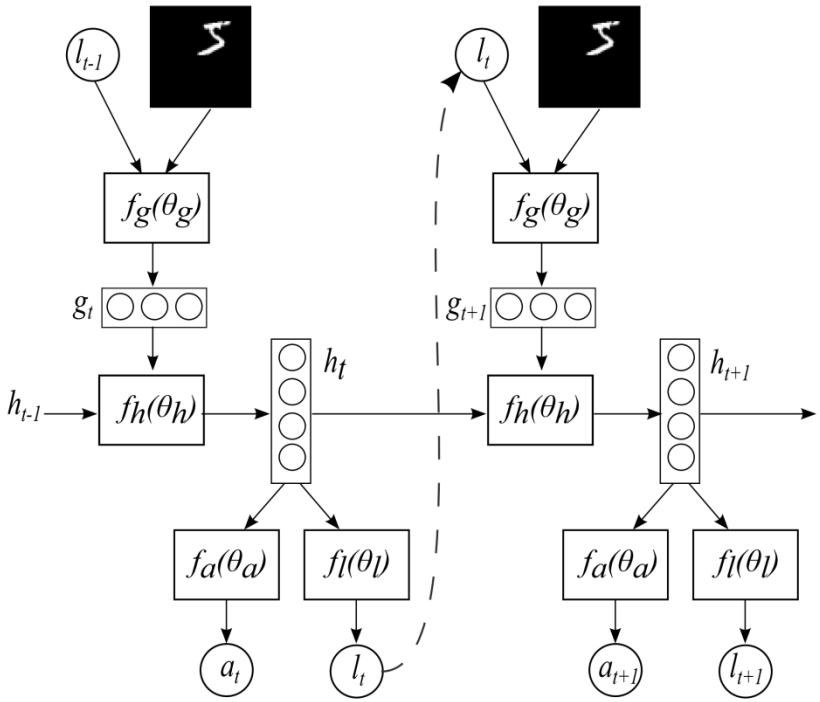
X the input (image, frame etc.)

l – the set of actions and locations

y – correct output (digit classification), correct word prediction etc.

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
  - Hard attention
  - Spatial transformer networks