Language Technologies Institute

Carnegie Mellon University

Advanced Multimodal Machine Learning

Lecture 12.1: Multimodal Fusion

Louis-Philippe Morency
Lecturer: Amir Zadeh

* Original version co-developed with Tadas Baltrusaitis

Lecture Objectives

- Multimodal fusion
  - Model free approaches
  - Model based approaches
- Kernel methods for fusion
  - Support vector machines
  - Multiple kernel learning
- Recap of multimodal challenges
- New directions in multimodal machine learning

Multimodal fusion

Multimodal fusion

- Process of joining information from two or more modalities to perform a prediction
- Examples
  - Audio-visual speech recognition
  - Audio-visual emotion recognition
  - Multimodal biometrics
  - Speaker identification and diarization
  - Visual/Media Question answering

(a) answer-phone (a) get-out-car (a) fight-person

(Potential) Benefits

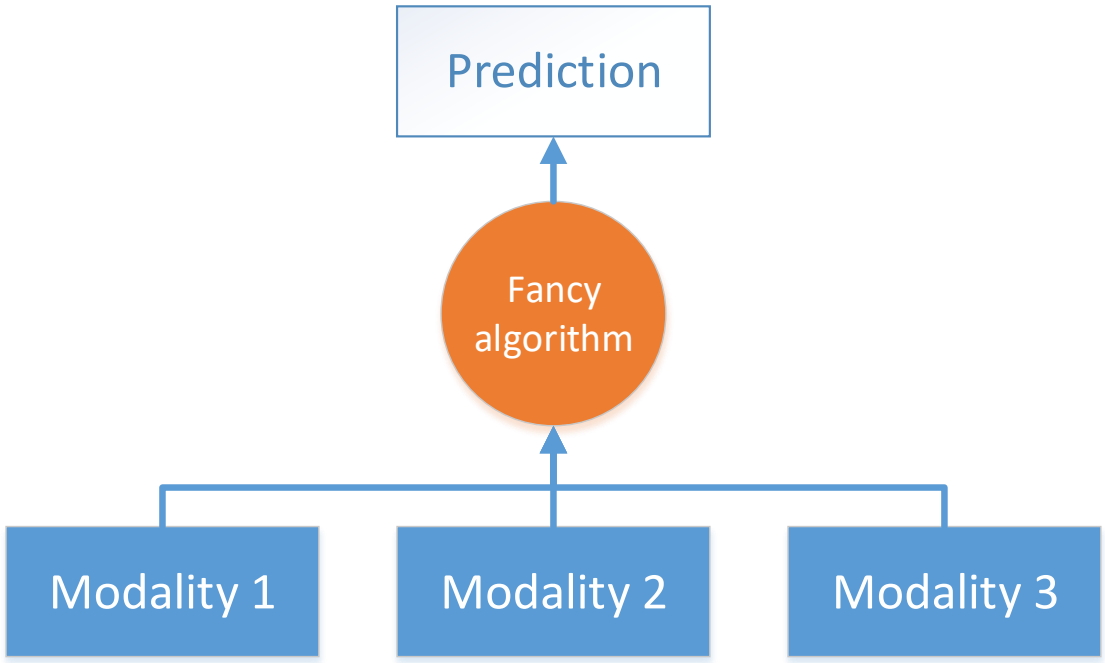
- Supplementary information - McGurk effect
  - The sum is greater than the parts
- Robustness in presence of noise in one modality
- Dealing with missing or unobserved data in one of the modalities

Challenges and pitfalls

- Different sampling rates
- Misaligned data
- Different amounts of noise in modalities
- Potentially missing data in one modality
- One of the modalities not being informative
- Different predictive power of each modality
- Modalities only providing redundancy
  - Complementary and not supplementary

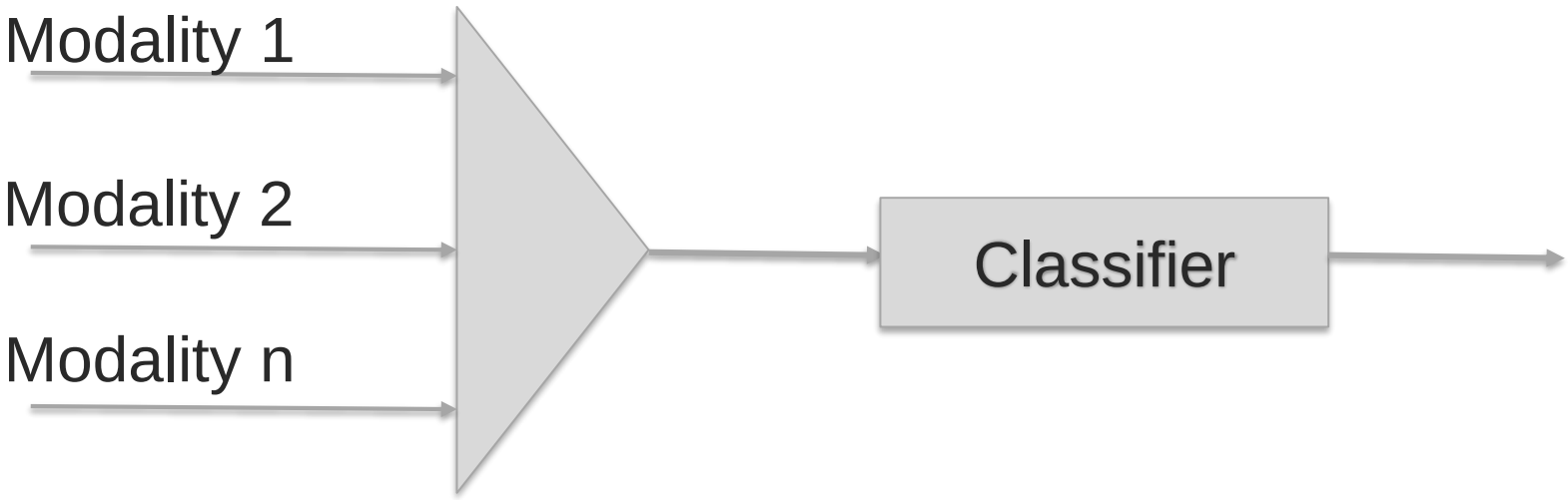
Multimodal Fusion

- Two major types
- Model Free
  - Early, late, hybrid
- Model Based
  - Neural Networks
  - Kernel Methods
  - Graphical models
  - Attention Based
  - Memory based



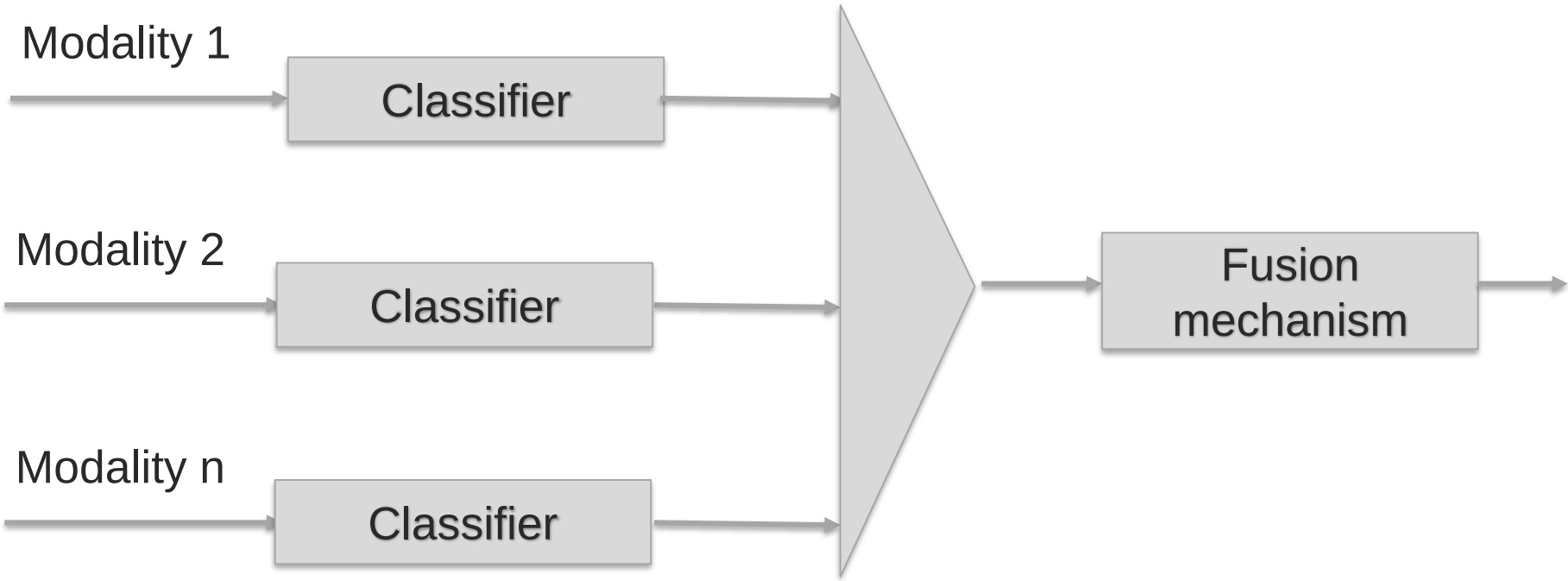
Model free approaches

Model free approaches – early fusion



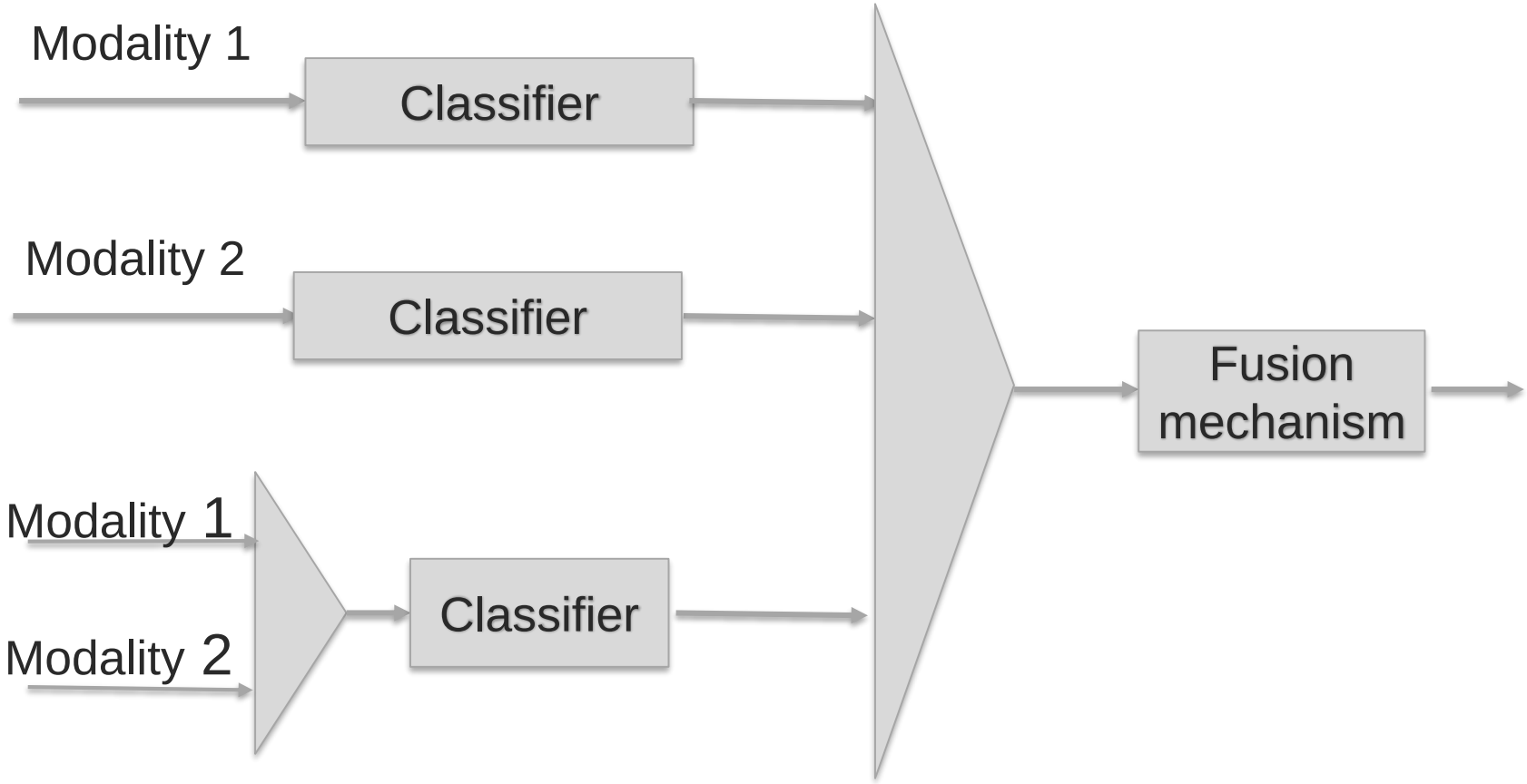
- Easy to implement – just concatenate the features
- Exploit dependencies between features
- Can end up very high dimensional
- More difficult to use if features have different framerates

Model free approaches – late fusion



- Train a unimodal predictor and a multimodal fusion one
- Requires multiple training stages
- Do not model low level interactions between modalities
- Fusion mechanism can be voting, weighted sum or an ML approach

Model free approaches – hybrid fusion



- Combine benefits of both early and late fusion mechanisms

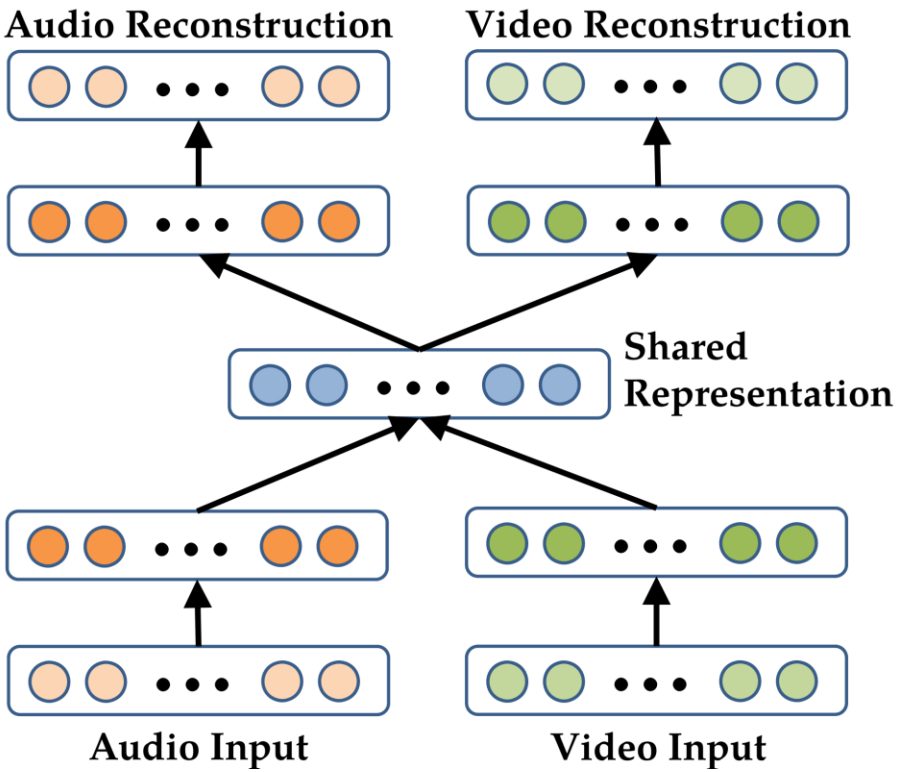
Model based fusion: Neural networks

Multimodal fusion using neural networks

- The fusion happens at some point in the neural network in an intermediate stage
- Line between representation and fusion is fuzzy

AVSR using neural networks

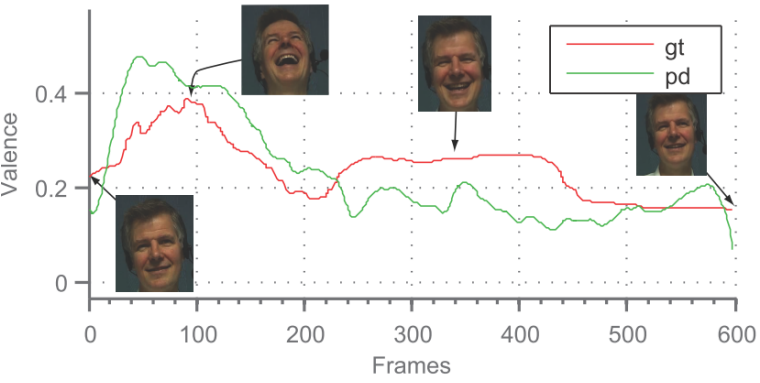
- Interestingly some late 80s early 90s work on this
- A more modern approach multimodal autoencoder
- Fine-tuning an autoencoder learned representation using an AVSR task
- Where does the fusion actually happen?

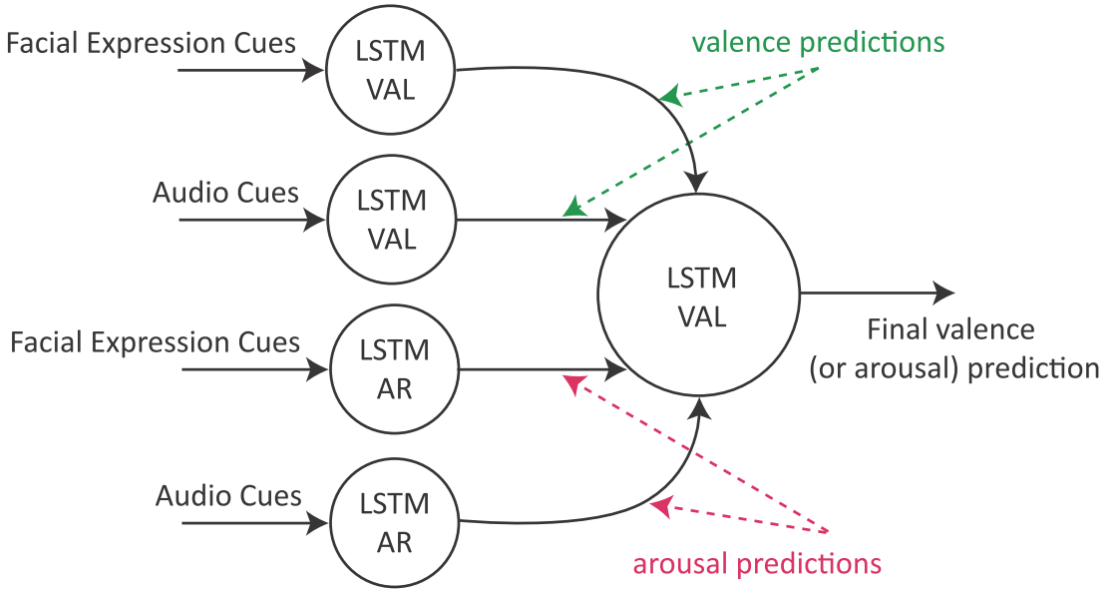


[Ngiam et al., Multimodal Deep Learning, 2011]

Emotion recognition using LSTMs

- More obvious fusion
- Using LSTM based fusion for audio-visual emotion recognition at each time step [Nicolaou 2011]

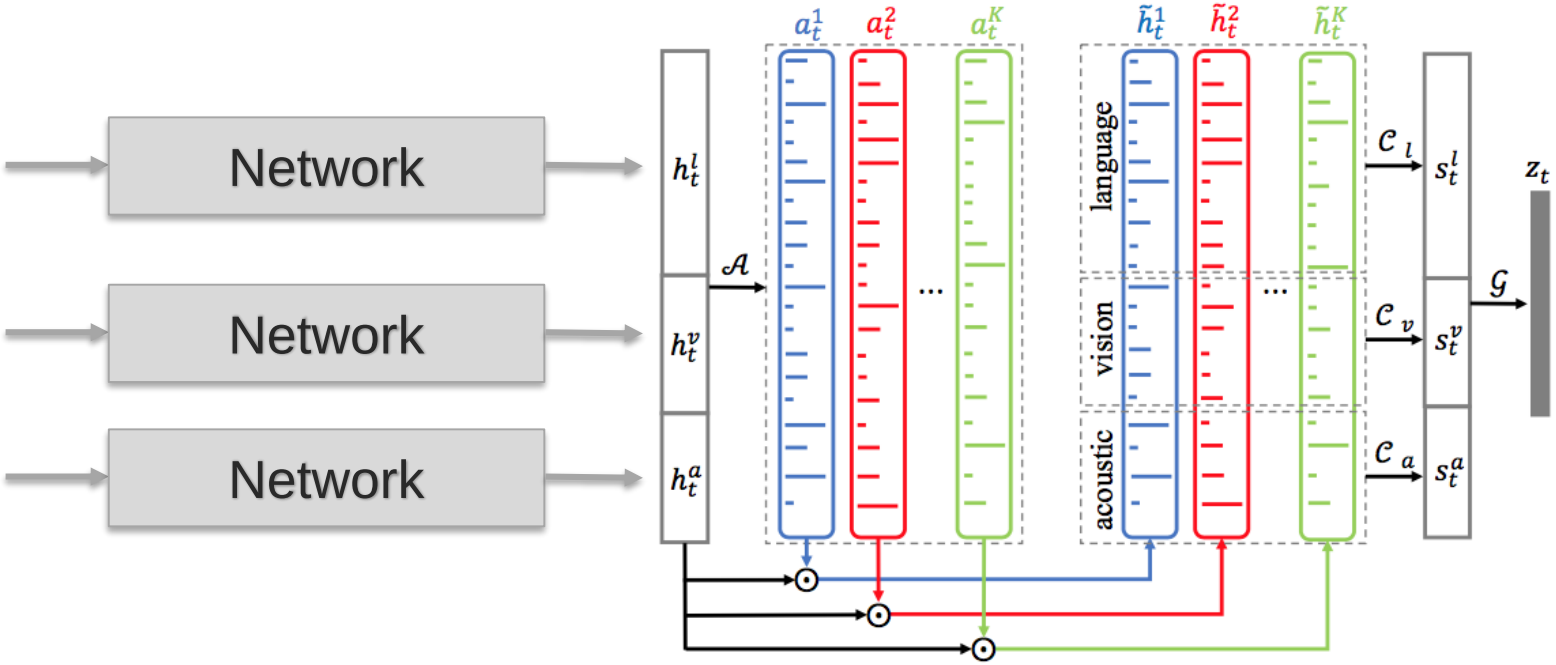




[Nicolaou et al., Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence–Arousal Space, 2011]

Using [Multiple] Attentions

- Modeling Human Communication – Sentiment, Emotions, Speaker Traits

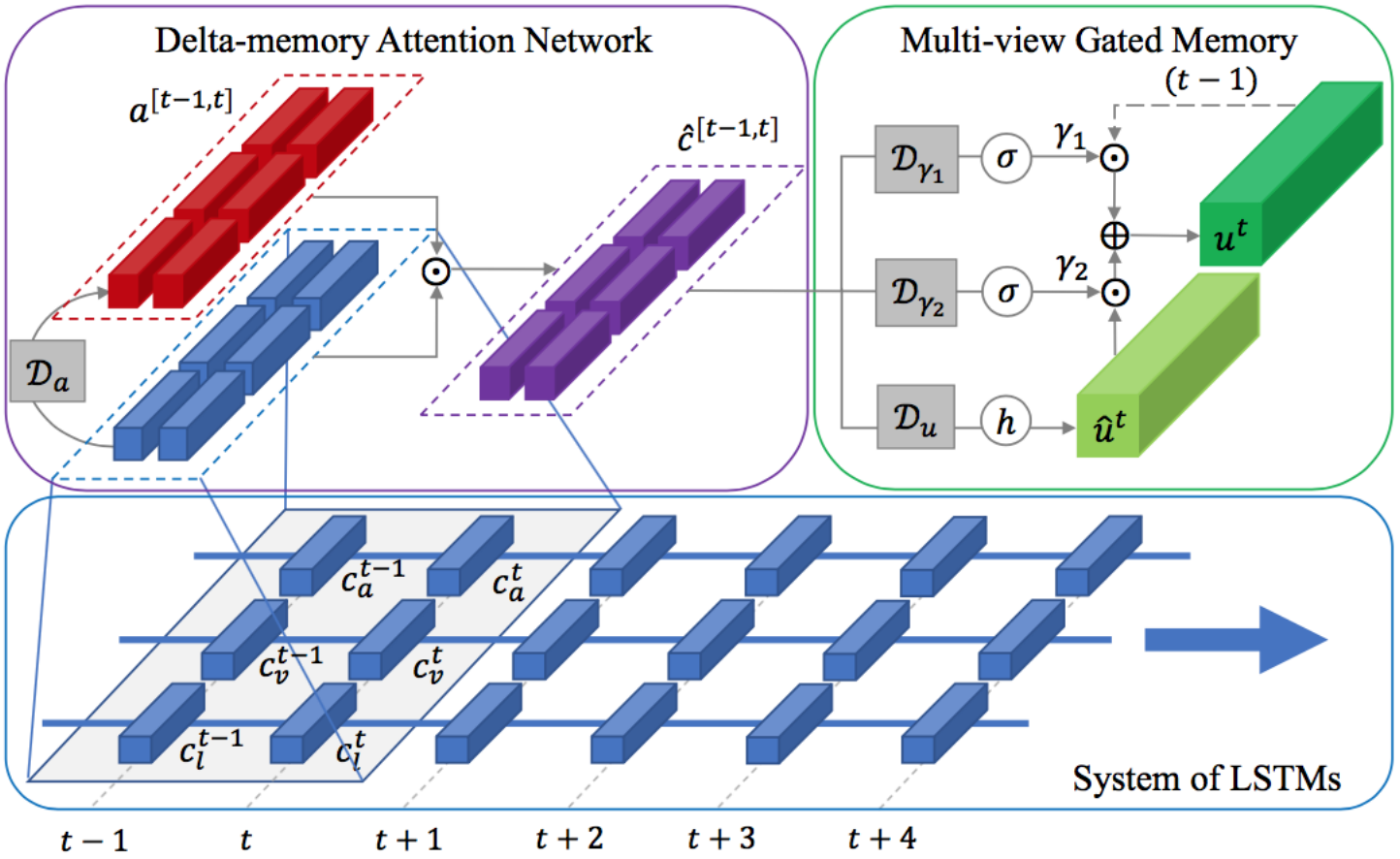


[Zadeh et al., Human Communication Decoder Network for Human Communication Comprehension, AAAI 2018]

Memory Based

- A memory accumulates multimodal information over time.
- From the representations throughout a source network.
- No need to modify the structure of the source network, only attached the memory.

Memory Based

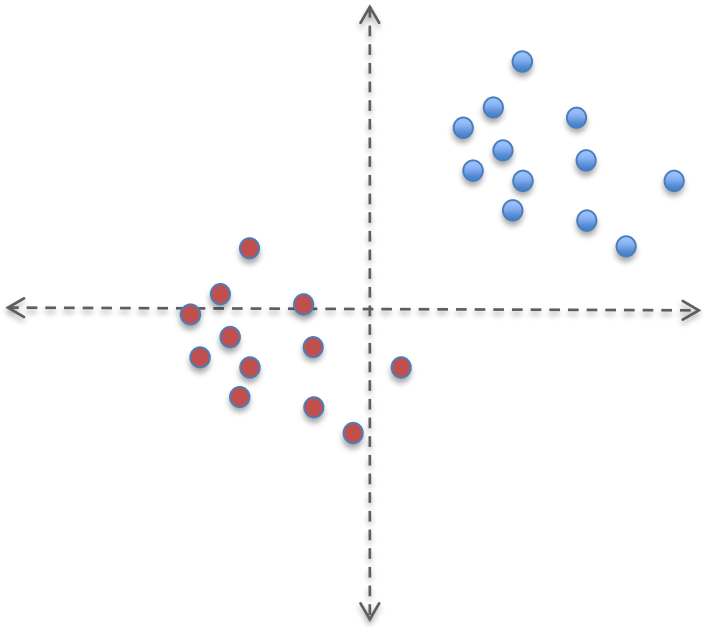


[Zadeh et al., Memory Fusion Network for Multi-view Sequential Learning, AAAI 2018]

Model based fusion: Multiple Kernel Learning

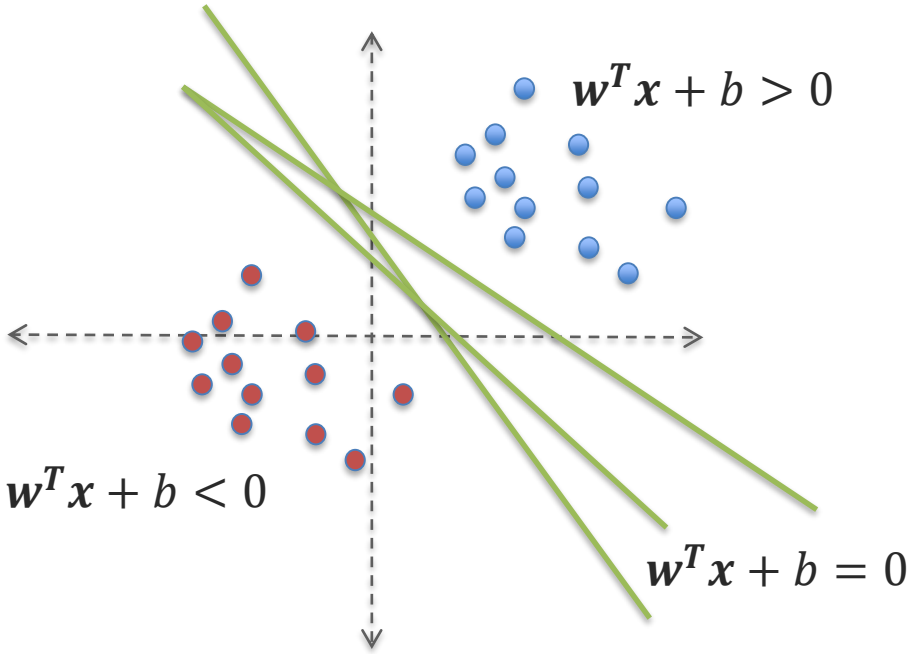
Crash course on SVMs!

- Want to build a model to classify a set of samples
- First of all lets define $\boldsymbol{x}_i \in \mathbb{R}^D$ and $y_i \in \{-1,1\}$, a binary classification problem



Basic idea behind SVM

- How to separate the data?
- Can use a line, our decision function becomes $y = \text{sign}(\boldsymbol{w}^T\boldsymbol{x} + b)$



$$\boldsymbol{x}_i \in \mathbb{R}^D$$

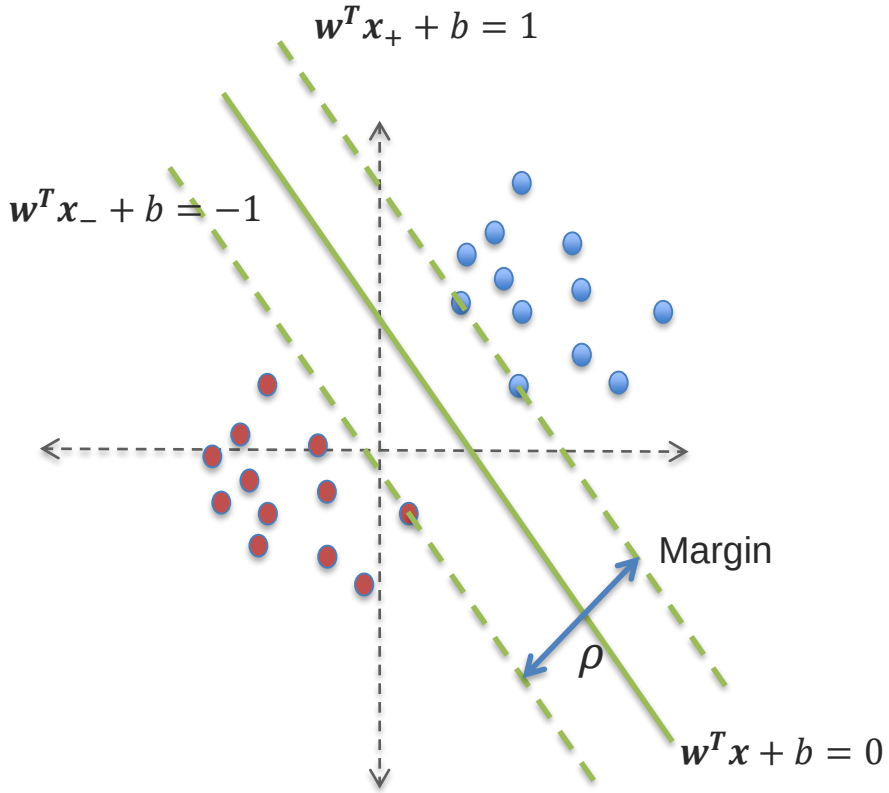
$$y_i \in \{-1,1\}$$

$$\boldsymbol{w} \in \mathbb{R}^D$$

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 \ (\forall i \in \{1, \ldots, n\})$$

SVM formulation

- Which separating line should we pick
- Intuitively pick max-margin classification
  - Empirical risk minimization guarantees



- How to compute the margin:
- $\rho = \frac{w^T}{\|w\|}(x_+ - x_-) = \frac{w^T x_+ - w^T x_-}{\|w\|} =$

$$= \frac{w^T x_+ - w^T x_- + b - b}{\|w\|} = \frac{2}{\|w\|}$$

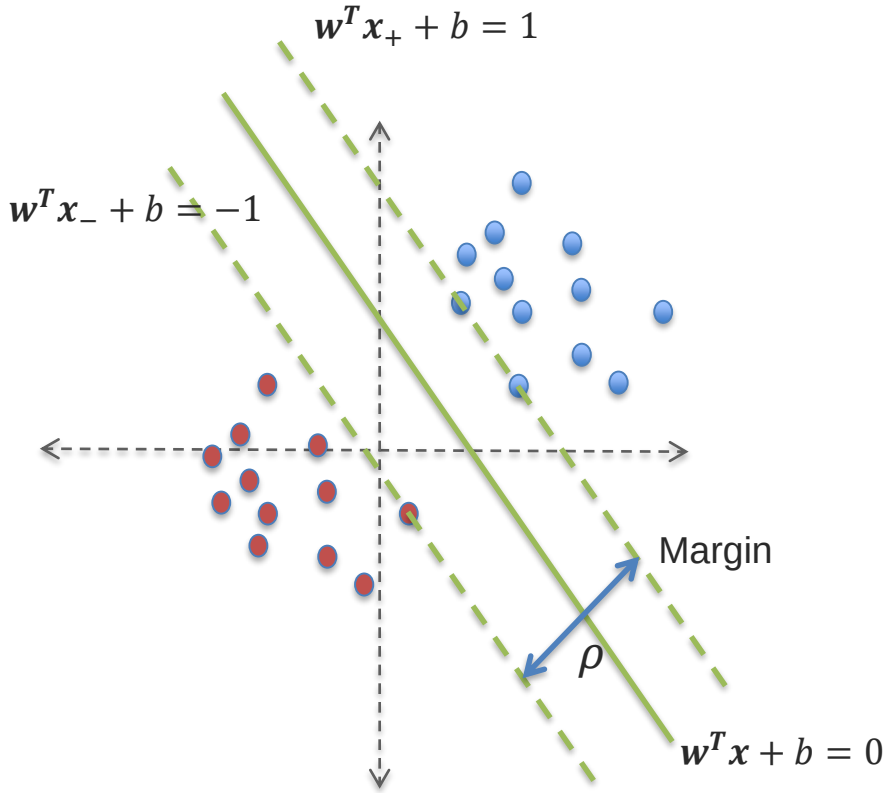
SVM formulation

- Margin size $\rho = \frac{2}{\|\boldsymbol{w}\|}$
- We want to maximize:

$$\frac{2}{\|\boldsymbol{w}\|}$$

- This is equivalent to minimizing:

$$\frac{1}{2}\|\boldsymbol{w}\|^2$$



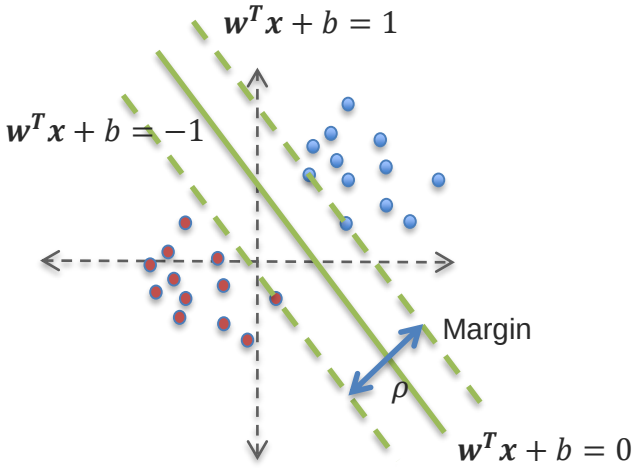
Hard margin SVM formulation

minimize: $\boldsymbol{w}$ $\frac{1}{2}\|\boldsymbol{w}\|^2$ ← Maximize the margin

subject to:

$$\boldsymbol{w}^T\boldsymbol{x}_i + b \geq 1 \quad \text{, for } y_i = 1$$
$$\boldsymbol{w}^T\boldsymbol{x}_i + b \leq -1 \quad \text{, for } y_i = -1$$

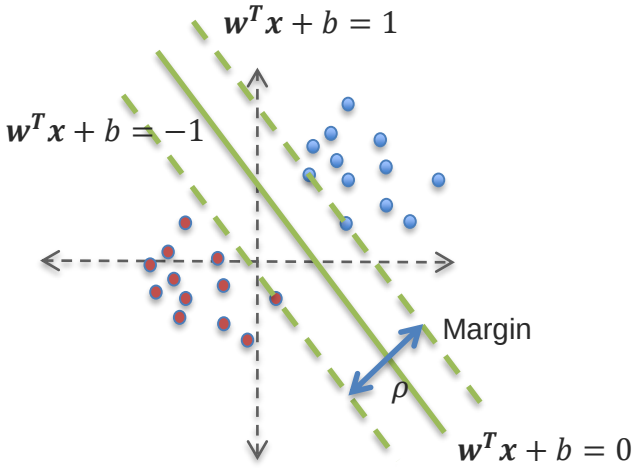
← Separate the data



Hard margin SVM formulation

minimize: $\boldsymbol{w}$ $\frac{1}{2}\|\boldsymbol{w}\|^2$ ← Maximize the margin

subject to: $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1$ ← Separate the data



The problem is convex (has a unique minimum) and we can actually find an optimal solution using Quadratic Programming

Non-separable data

Previous formulation assumes that data is linearly separable, what if it is not?

The new formulation:

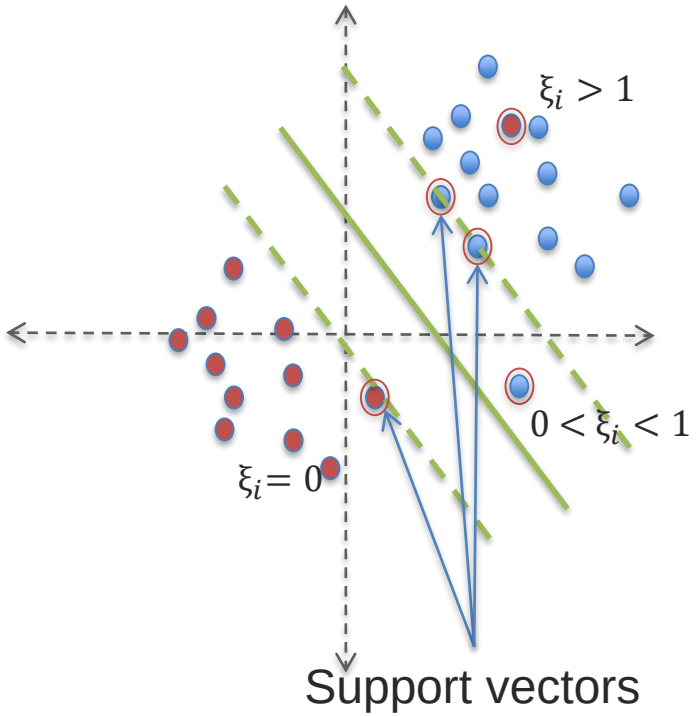
$$y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right) \geq 1 - \xi_i$$

Adding a "slack" variable:

$$\xi_i \geq 0$$

When:

- $\xi_i = 0$, *classification is correct*
- $0 < \xi_i < 1$, *violating the margin*
- $\xi_i > 1$, *missclassification*



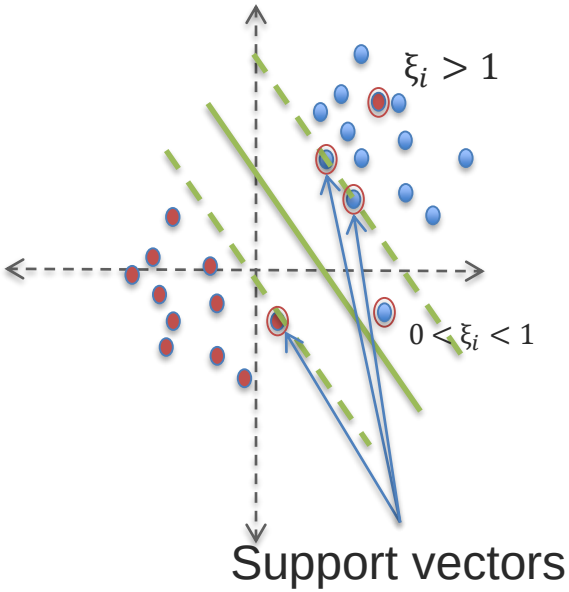
Soft margin SVM formulation

minimize:
$\boldsymbol{w}, \xi$

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{M}\xi_i$$

subject to:

$$y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$



Allows to misclassify a number of samples in order to get a wider margin

C – tradeoff between margin size and missclassification, a **hyperparameter**

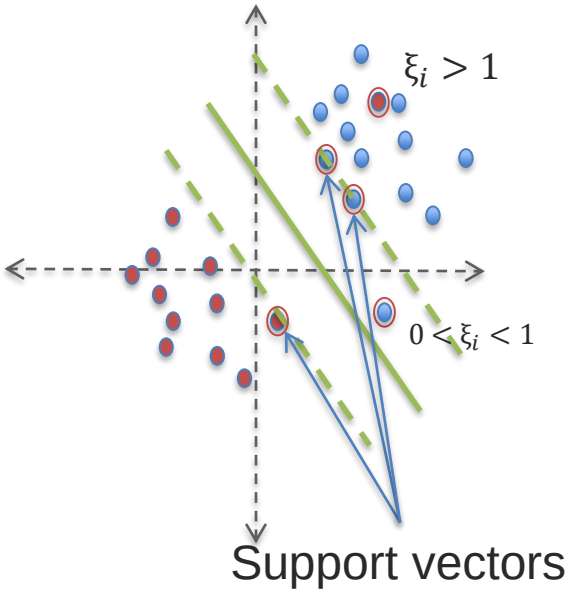
Soft margin SVM optimization

minimize:
$\boldsymbol{w}, \boldsymbol{\xi}$

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{M}\xi_i$$

subject to:

$$y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$



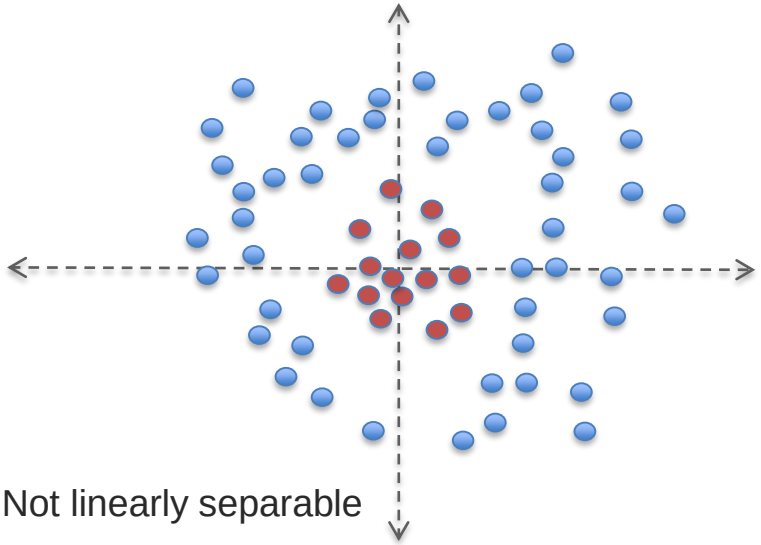
As before we can use quadratic programming to arrive at a solution, just now we need to minimize across $\boldsymbol{w}$ and ξ

What is a Kernel function

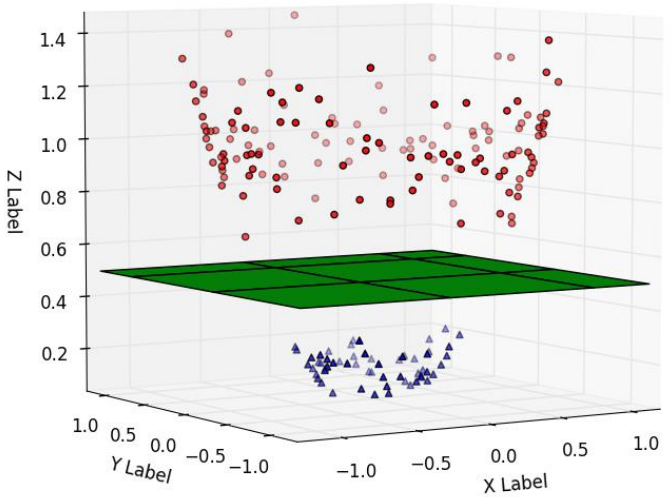
- What is a kernel function

  $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$, where $\phi: D \rightarrow Z$
- Kernel function performs an inner product in feature map space ϕ
- Inner product (a generalization of the dot product) is often denoted as $\langle ., . \rangle$ in SVM papers
- $\boldsymbol{x} \in \mathbb{R}^D$(but not necessarily), but $\phi(\boldsymbol{x})$ can be in any space – same, higher, lower or even in an infinite dimensional space
- Acts as a similarity metric between data points

Non-linearly separable data



Not linearly separable



Same data, but now linearly separable

- Want to map our data to a linearly separable space
- Instead of $\boldsymbol{x}$, want $\phi(\boldsymbol{x})$, in a separable space ($\phi(\boldsymbol{x})$ is a feature map)
- What if $\phi(\boldsymbol{x})$ is much higher dimensional? We do not want to learn more parameters and mapping could become very expensive

Radial Basis Function Kernel (RBF)

- Arguably the most popular SVM kernel
- $K(x_i, x_j) = \exp -\frac{1}{2\sigma^2}\|x_i - x_j\|^2$
- $\phi(\boldsymbol{x}) =?$ It is infinite dimensional and fairly involved, no easy way to actually perform the mapping to this space, but we know what an inner product looks like in it
- σ – a hyperparameter
- With a really low sigma the model becomes close to a KNN approach (potentially very expensive)

Some other kernels

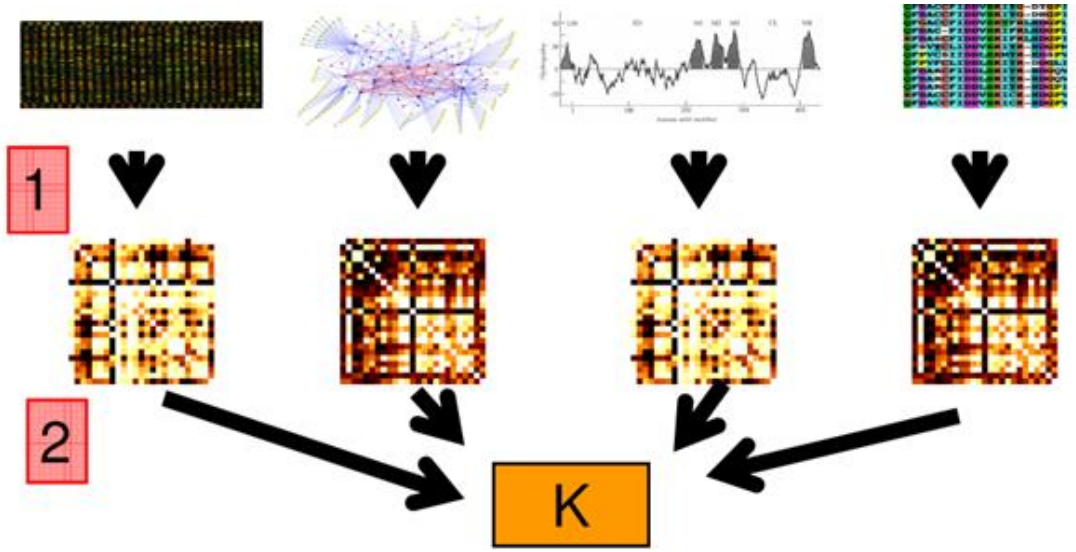
- Other kernels exist
  - Histogram Intersection Kernel – good for histogram features
  - String kernels – specifically for text and sentence features
  - Proximity distribution kernel
  - (Spatial) pyramid matching kernel
- The restriction is for Gram matrix to be positive semi definite, otherwise anything goes

Different properties of different signals

- How do we deal with heterogeneous or multimodal data?
- The data of interest is not in a joint space so appropriate kernels might be different
- Multiple Kernel Learning (MKL) is a way to address this
  - Was popular for image classification and retrieval before deep learning approaches came around (winner of 2010 VOC challenge, ImageClef 2011 challenge)
  - MKL - fell slightly out of favor when deep learning approaches became popular
  - Still useful when large datasets are not available

Multiple Kernel Learning

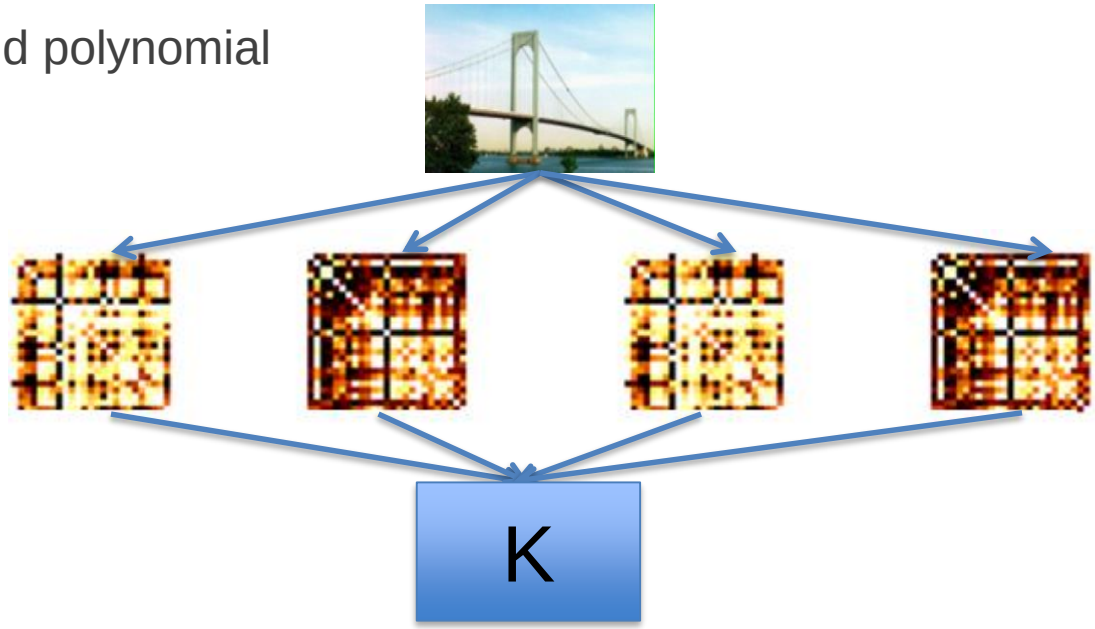
- Instead of providing a single kernel and validating which one works optimize in a family of kernels (or different families for different modalities)
- Works well for unimodal and multimodal data, very little adaptation is needed



[Lanckriet 2004]

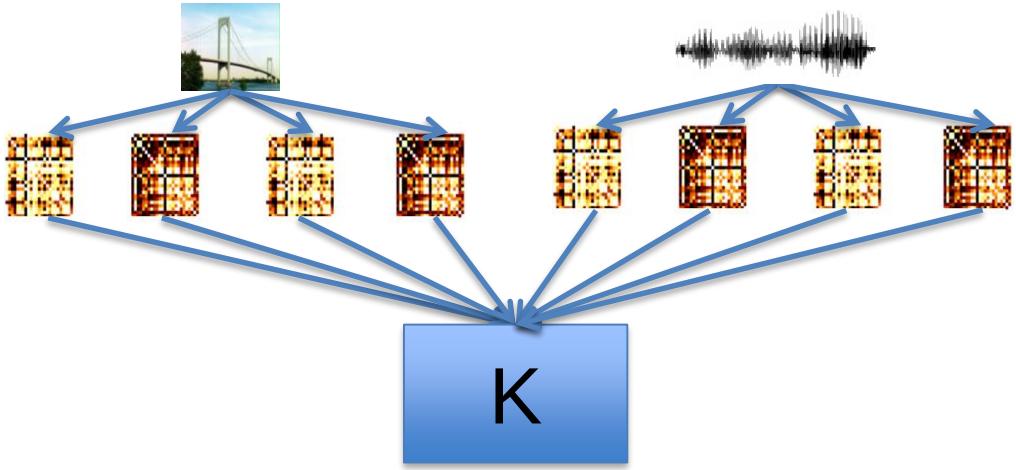
MKL in unimodal case

- Pick a family of kernels and learn which kernels are important for the classification case
- For example a set of RBF and polynomial kernels



MKL in multimodal/multiview case

- Pick a family of kernels for each modality and learn which kernels are important for the classification case
- Does not need to be different modalities, often we use different views of the same modality (HOG, SIFT, etc.)



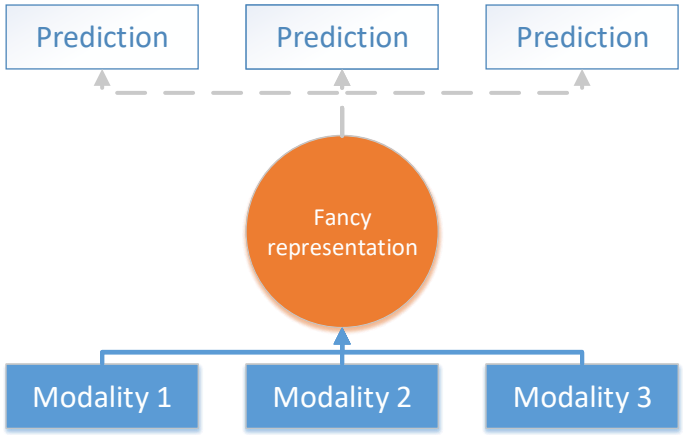
Multiple Kernel Learning

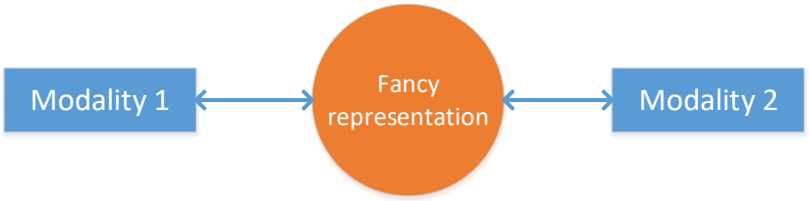
- Allows to reduce amount of cross validation, (e.g. instead of using σ as a hyperparameter in RBF learn which values are important)
- Instead of feature selection throw all of them at MKL and let the kernels learn which ones are important
- Dealing with different format and scale data (real, ordinal, nominal)
- A technically sound way of combining features
- Feature combination and classifier training is done simultaneously

Multimodal machine learning recap

Challenge 1 - Multimodal representation

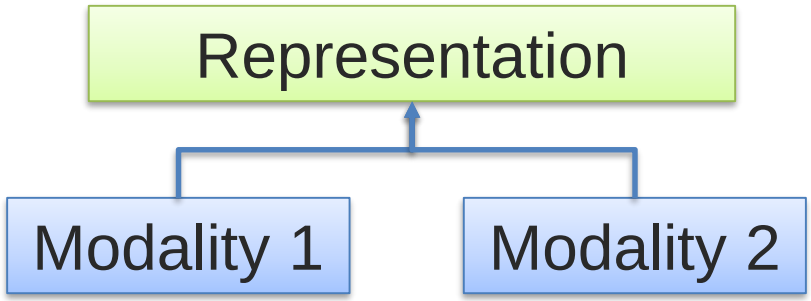
- "Computer interpretable description of the multimodal data (e.g., vector, tensor)"
  - Missing modalities
  - Heterogeneous data (symbols vs signals)
  - Static vs. sequential data
  - Different levels of noise
- Focus throughout the course, particularly weeks 3-7





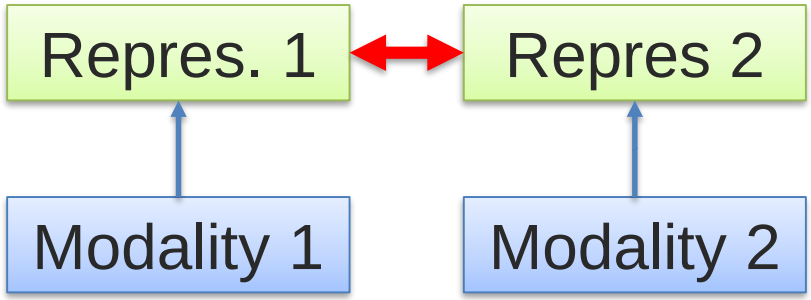
Challenge 1 - Multimodal representation types

A Joint representations:



- Simplest version: modality concatenation (early fusion)
- Can be learned supervised or unsupervised
- Multimodal factor analysis

B Coordinated representations:



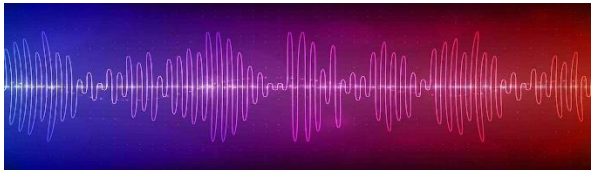
- Similarity-based methods (e.g., cosine distance)
- Structure constraints (e.g., orthogonality, sparseness)

Challenge 2 - Multimodal Translation / Mapping

- **Visual animations**

- **Image captioning**

- **Speech synthesis**



Translation / mapping:

"Process of changing data from one modality to another"

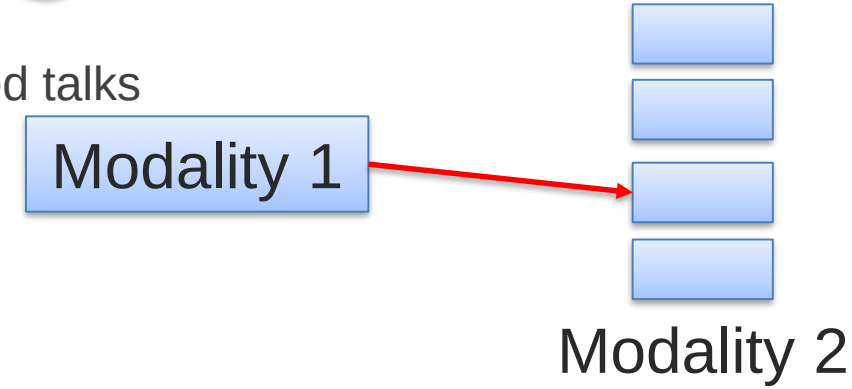
Challenges:

I. Different representations
II. Multiple source modalities
III. Open ended translations
IV. Subjective evaluation
V. Repetitive processes

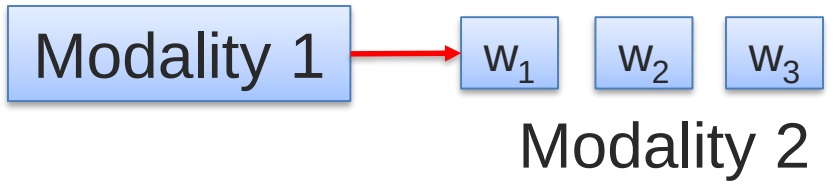
Challenge 2 - Multimodal Translation / Mapping

- Two major types
  - Example based
  - Generative
- Focus of some of the invited talks and Week 4, 5, 9

A **Bounded (example based) translations:**

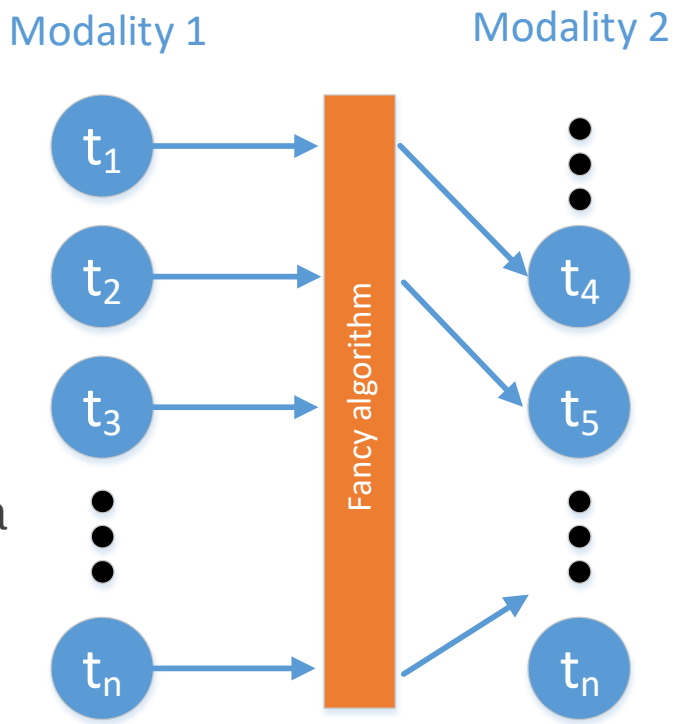


B **Open-ended (generative) translations:**



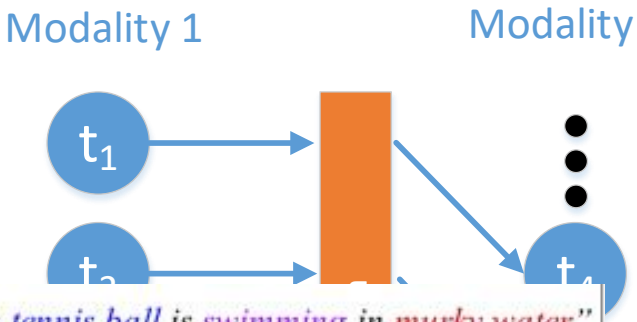
Challenge 3 - Alignment

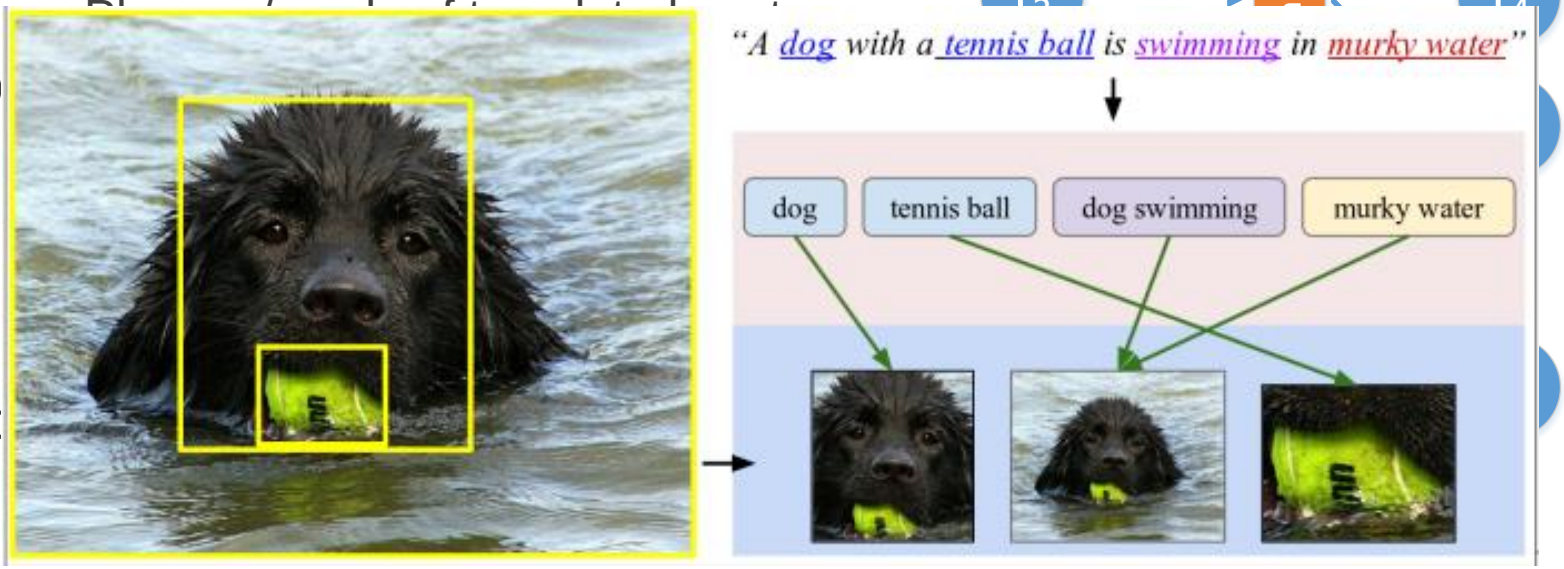
- Alignment of (sub)components of multimodal signals
- Examples
  - Images with captions
  - Recipe steps with a how-to video
  - Phrases/words of translated sentences
- Two types
  - Explicit – alignment is the task in itself
  - Latent – alignment helps when solving a different task (for example “Attention” models)
- Focus of week 9



Challenge 3 - Alignment

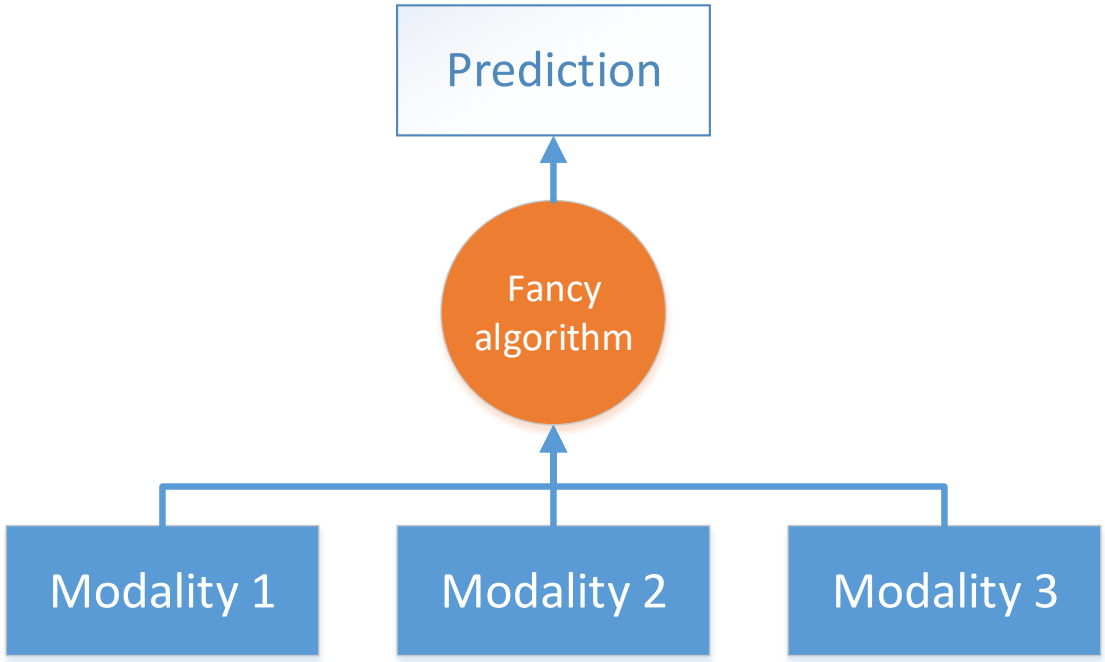
- Alignment of (sub)components of multimodal signals
- Examples
  - Images with captions
  - Recipe steps with a how-to video
- Tw
- Fc





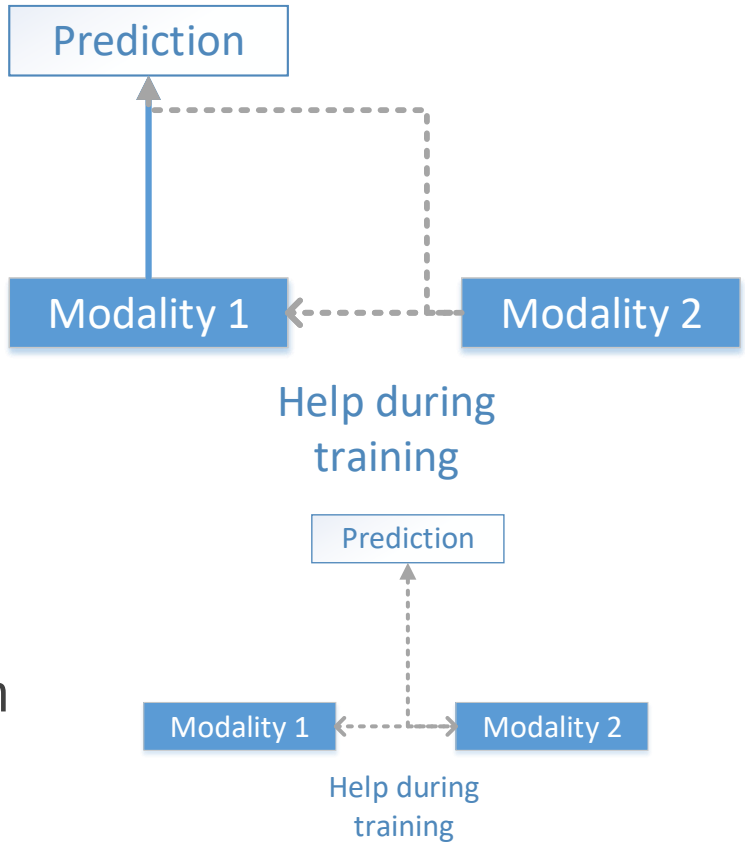
Challenge 4 - Multimodal Fusion

- Process of joining information from two or more modalities to perform a prediction
- Two major types
- Model Free
  - Early, late, hybrid
- Model Based
  - Kernel Methods
  - Graphical models
  - Neural networks
  - Attention Based
  - Memory Based
- Focus of Week 12



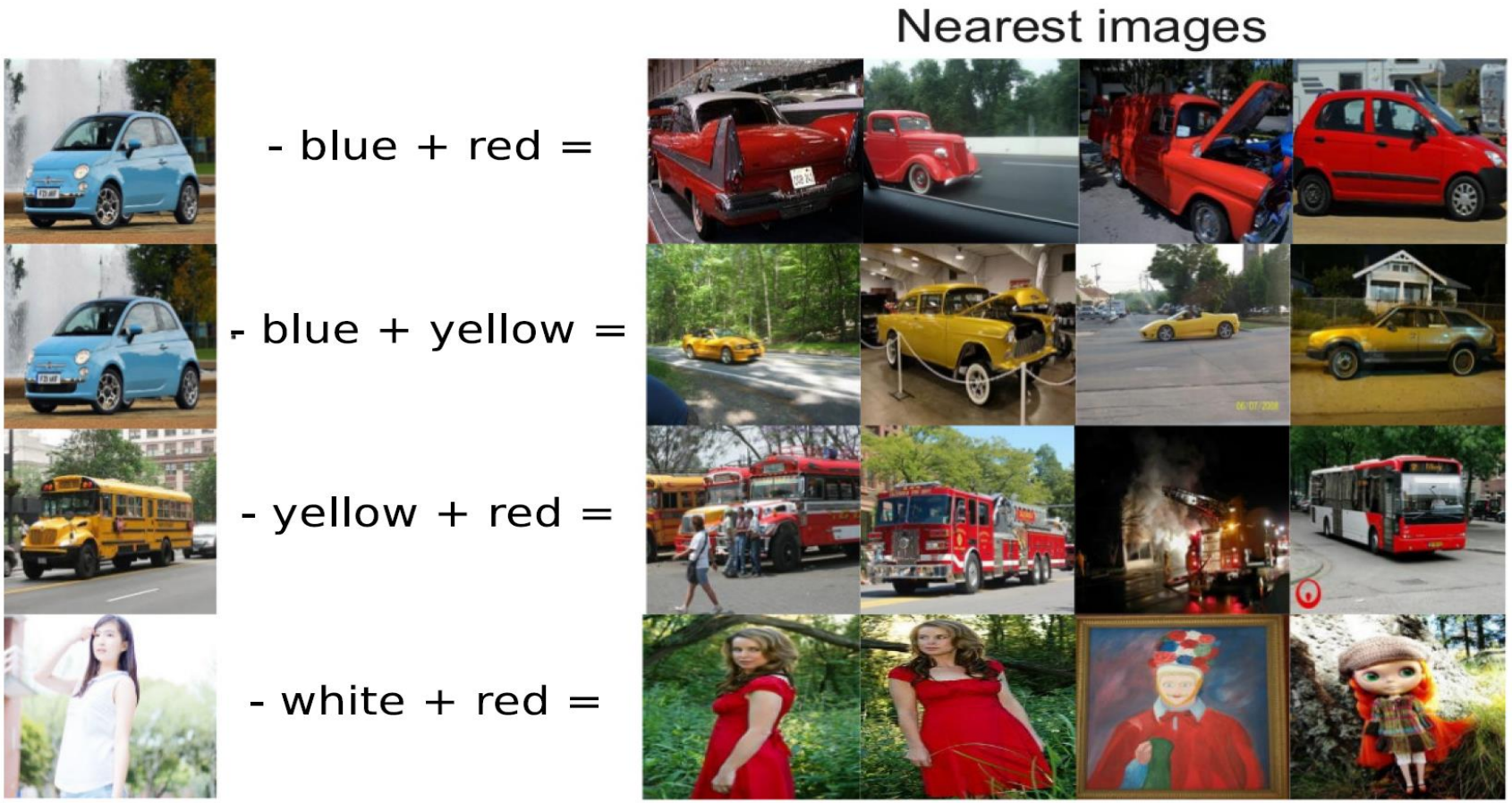
Challenge 5 – Co-learning

- How can one modality help learning in another modality?
  - One modality may have more resources
  - Bootstrapping or domain adaptation
  - Zero-shot learning
- How to alternate between modalities during learning?
  - Co-training (term introduced by Avrim Blum and Tom Mitchell from CMU)
- Transfer learning



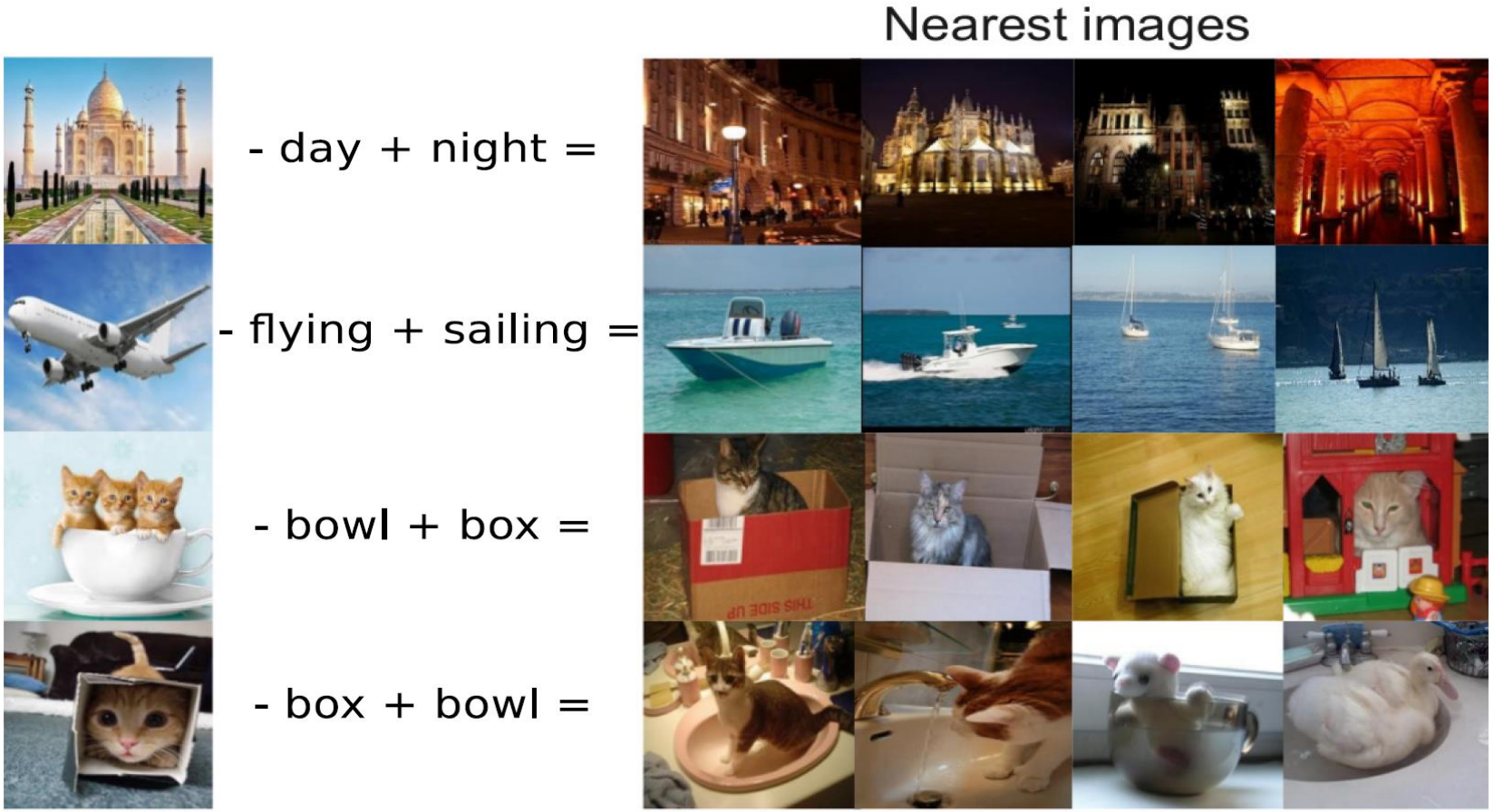
New(ish) and exciting directions

Multimodal representation 1



[Kiros et al., Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, 2014]

Multimodal representation 1

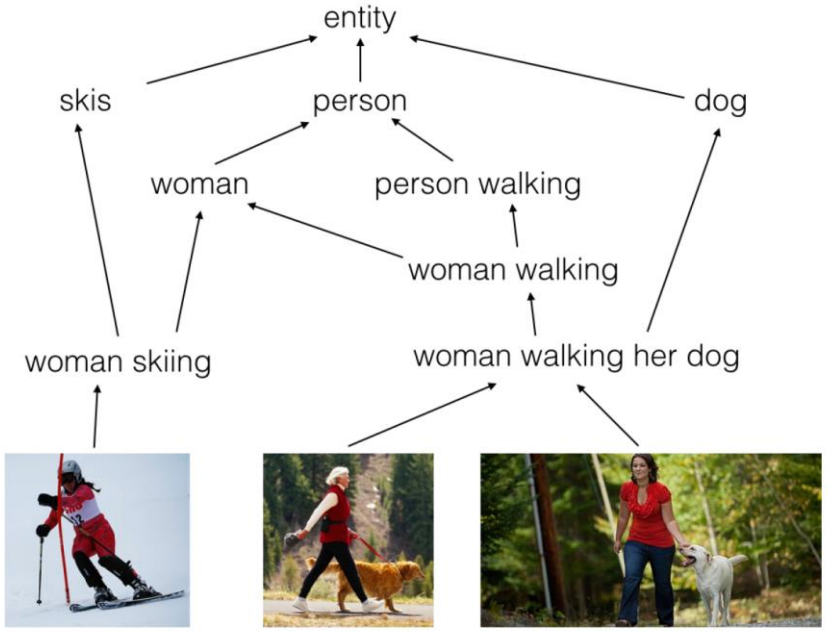


[Kiros et al., Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, 2014]

Multimodal representation 2

- We talked about coordinated representations, but mostly enforced “simple” coordination
  - Can we take it further?
- Replaces symmetric similarity

$$x \preceq y \text{ if and only if } \bigwedge_{i=1}^{N} x_i \geq y_i$$

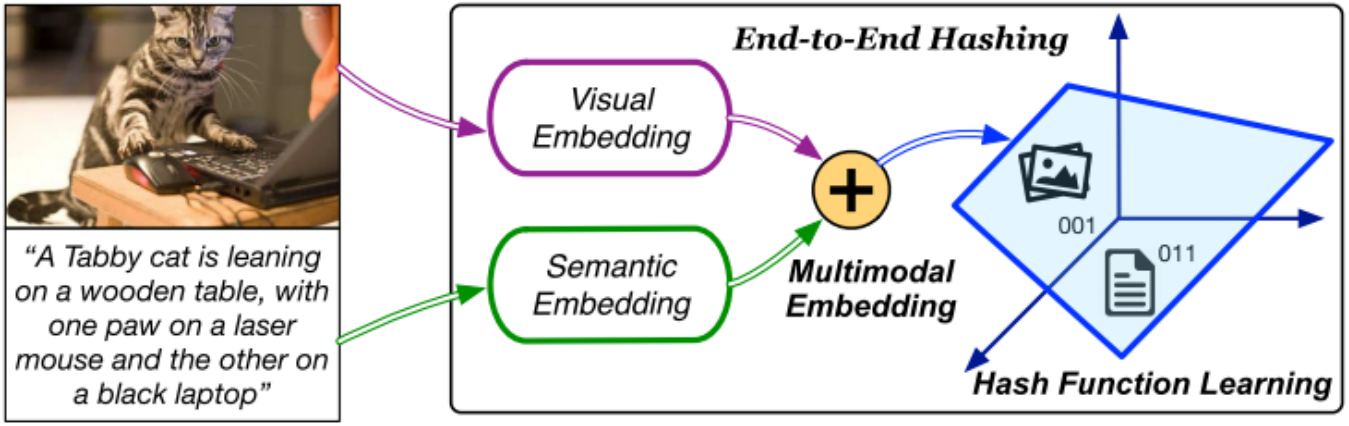


- Enforce approximate structure when training the embedding

[Vendrov et al. Order-embeddings of images and language, ICLR 2016]

Multimodal representation 3

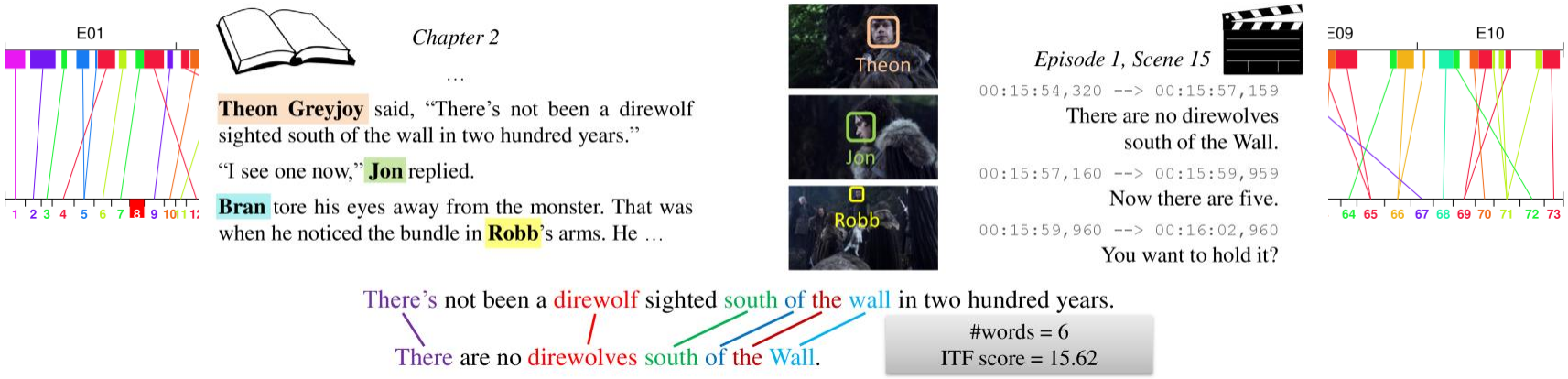
- We talked about coordinated representations, but mostly enforced “simple” coordination
- We can make embeddings more suitable for retrieval
  - Enforce a Hamming space (binary n-bit space)



[Cao et al. Deep visual-semantic hashing for cross-modal retrieval, KDD 2016]

Multimodal alignment 1

- Aligning very different modalities
- Books to scripts/movies

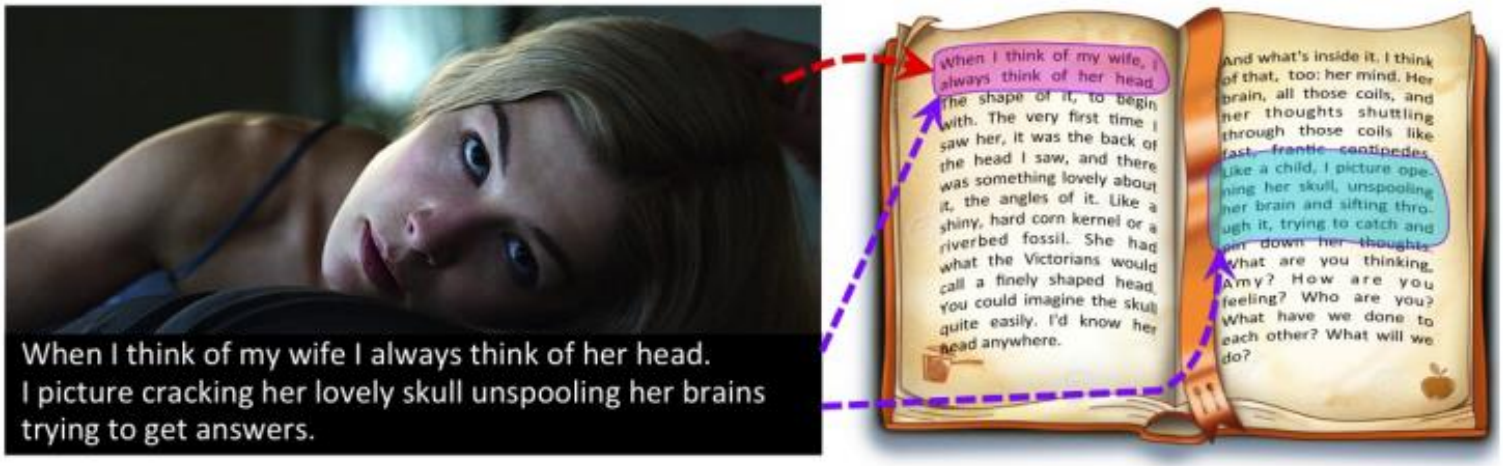


- Hand-crafted similarity based approach

[Tapaswi et al. Book2Movie: Aligning Video scenes with Book chapters, CVPR 2015]

Multimodal alignment 2

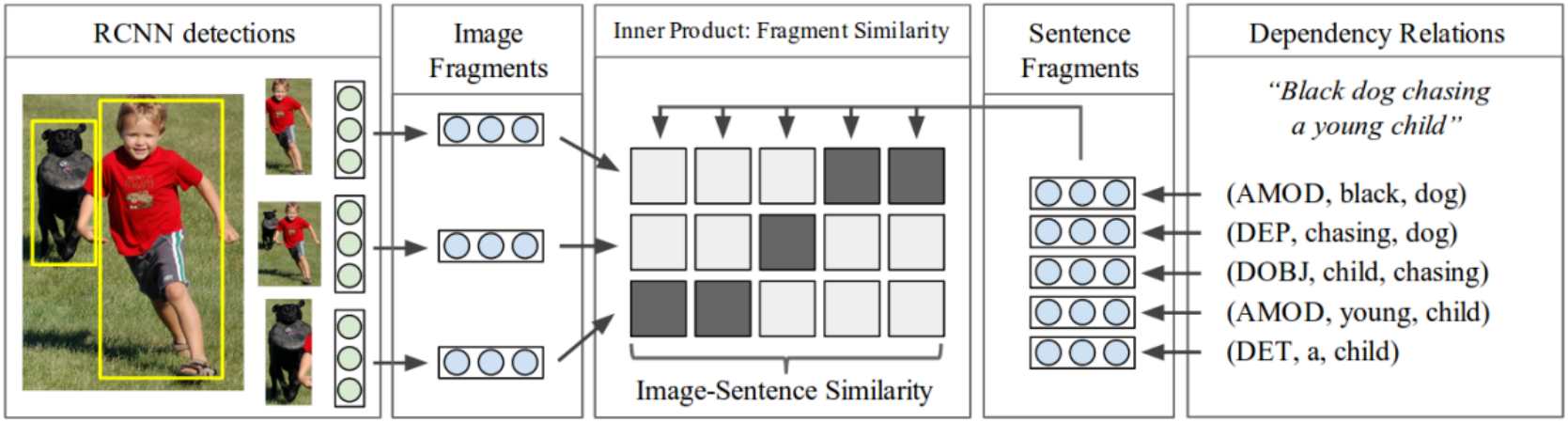
- Aligning very different modalities
- Books to scripts/movies



- Supervision based approach

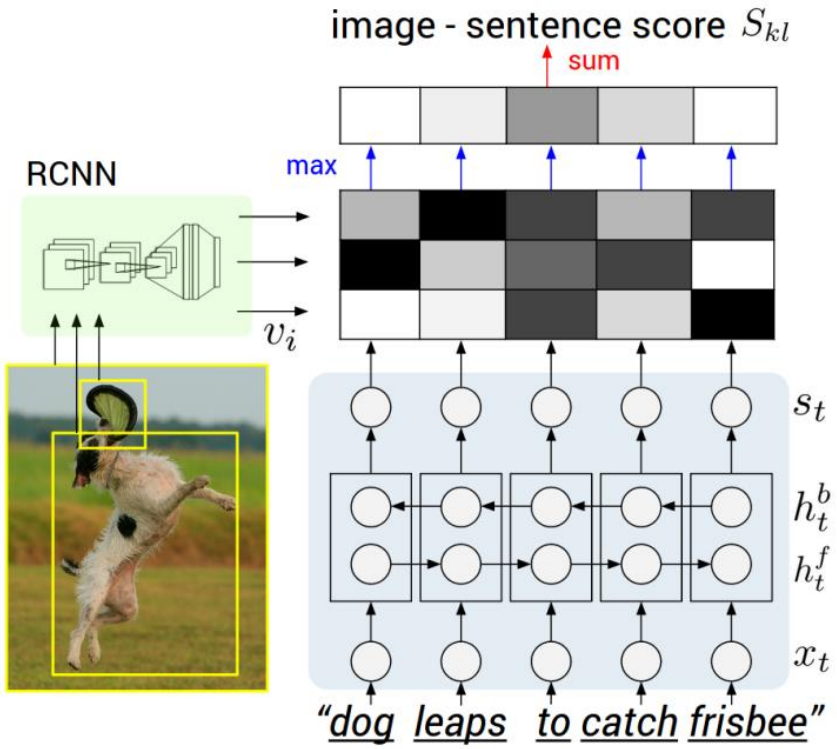
[Zhu et al. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books, ICCV 2015]

Multimodal alignment and translation 1



[Karpathy et al. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping, 2014]

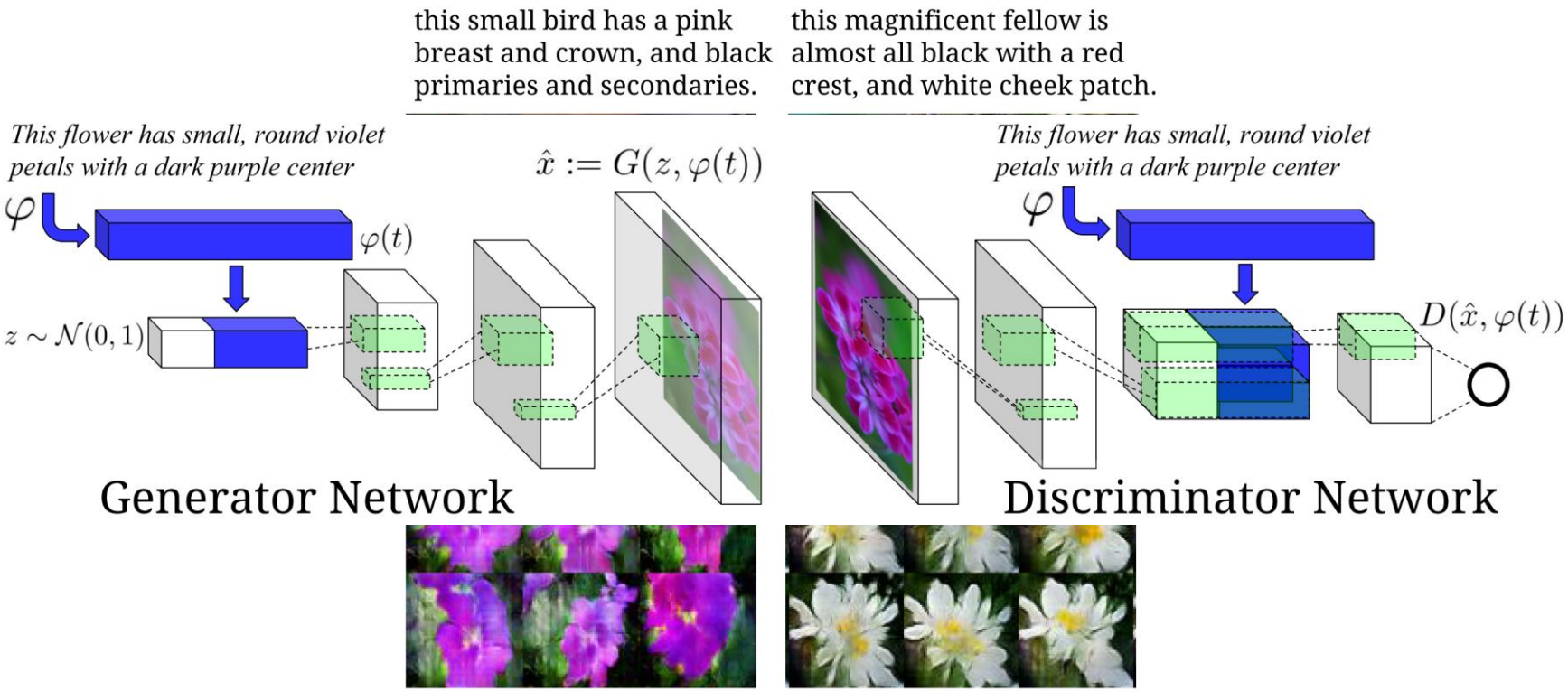
Multimodal alignment and translation 1



[Karpathy et al. Deep Visual-Semantic Alignments for Generating Image Descriptions, CVPR, 2015]

Multimodal translation 2

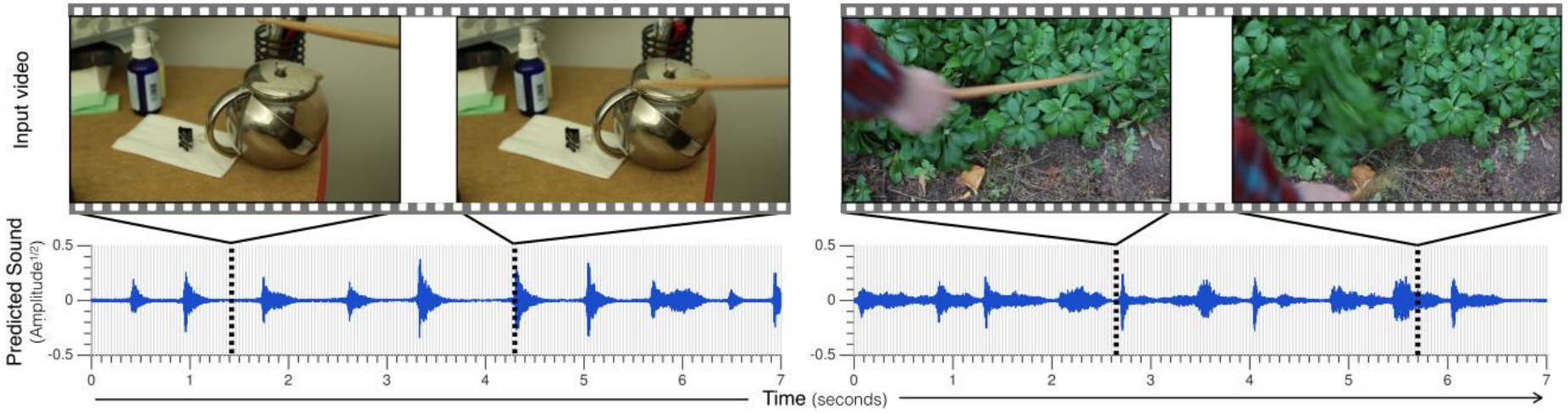
- Generative Adversarial Networks



[Reed et al. Generative Adversarial Text to Image Synthesis, ICML, 2016]

Multimodal translation 3

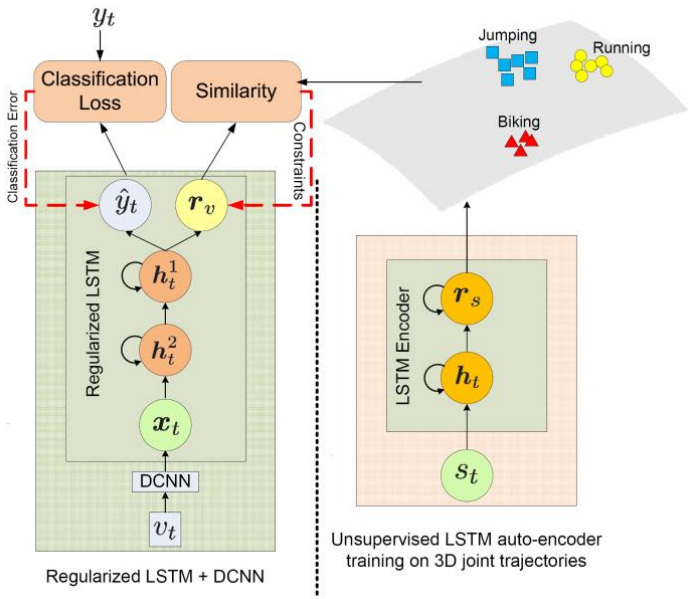
- Sound generation!

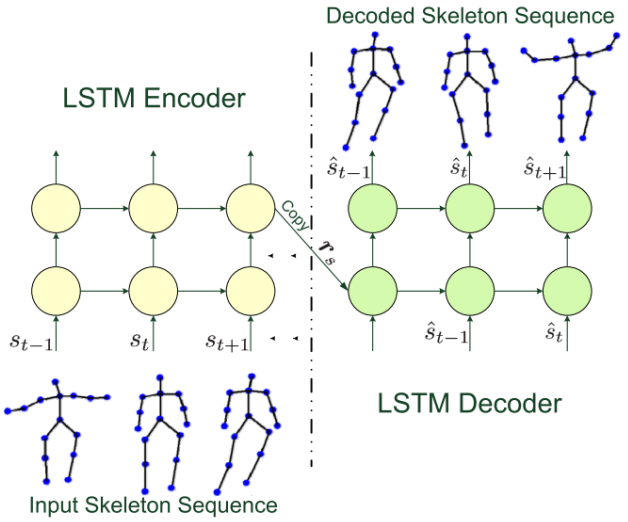


[Owens et al. Visually indicated sounds, CVPR, 2016]

Co-learning 1

- Better unimodal representation by regularizing using a different modality





Non parallel data!

[B. Mahasseni and S. Todorovic, "Regularizing Long Short Term Memory with 3D Human-Skeleton Sequences for Action Recognition," in CVPR, 2016]