

# Deep Learning for Multi-Modal Systems

DS3 Hertie School

---

Dr Christian Arnold (Cardiff University)

August 25, 2023

# Logistics

- The repo with all material is at  
[https://github.com/chrisguarnold/ds3\\_mmdl](https://github.com/chrisguarnold/ds3_mmdl)
- Please do ask questions in the chat: We will handle them as a team.

This is the most exciting time  
to be a social scientist.  
Ever.

# My Background

## Academic Background

- Senior Lecturer at Cardiff University
- PhD, University of Mannheim
- Research interests using MMDL, e.g.
  - **Voting Fraud Detection:** Multimodal data (satellite and micro data) to detect voting irregularities in remote areas.
  - **Elite Communication:** Making the most of MM data to study communication of political elites.

## Industry Background

- Data Scientist with KIANA and KPMG
- Scientific advisory committee for the GSS, Office for National Statistics
- Detecting voting fraud in the field, National Democratic Institute

# What I Assume About Your Background

## General Background

- Background mostly in Social Sciences
- At the moment: postgraduate academic qualification
- You know R or Python
- You know machine learning

## Experience with Deep Neural Networks

Basic exposure —[xxxxxxxxxx]——— Expert data scientist

# Today's Agenda

---

1. Introduction
2. Foundations
3. Basic Multimodal Learning
4. Extensions

# Today's Learnings

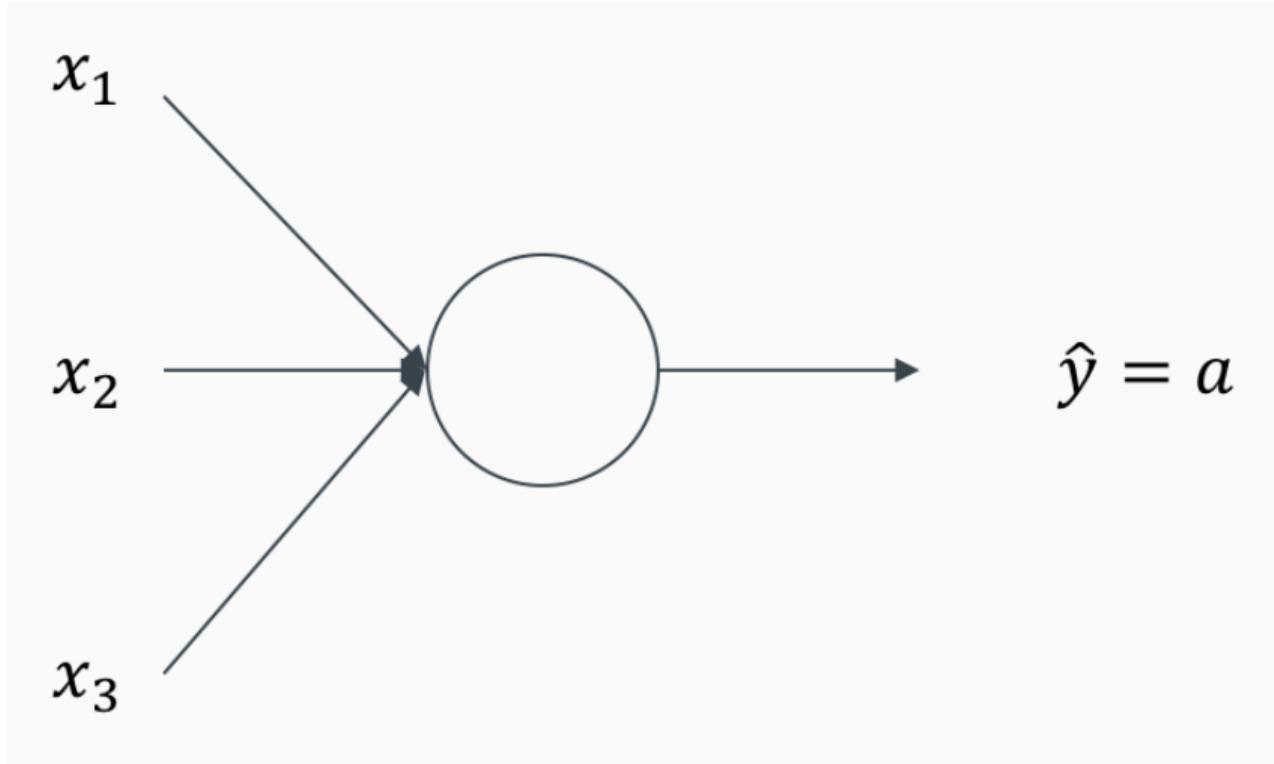
Topic	Knowledge	Experience	Skill
Introduction	X	X	
Foundations	X		
Basic MML	X	X	X
Extensions	X	X	

# Foundations

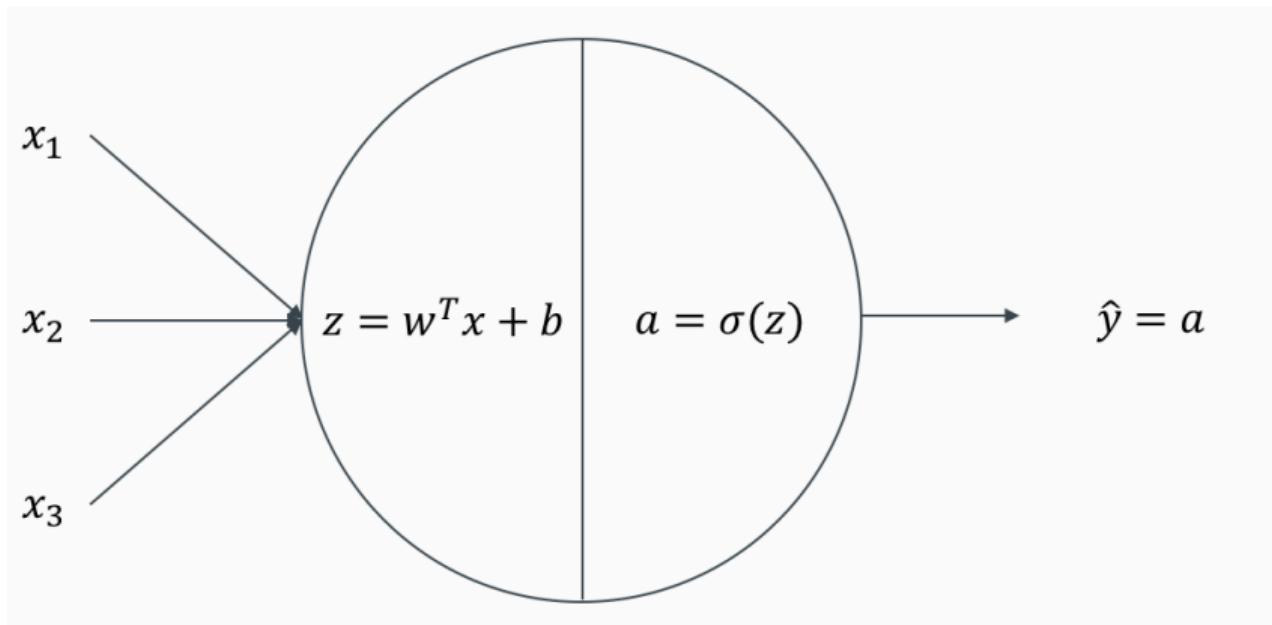
---

## Deep Learning

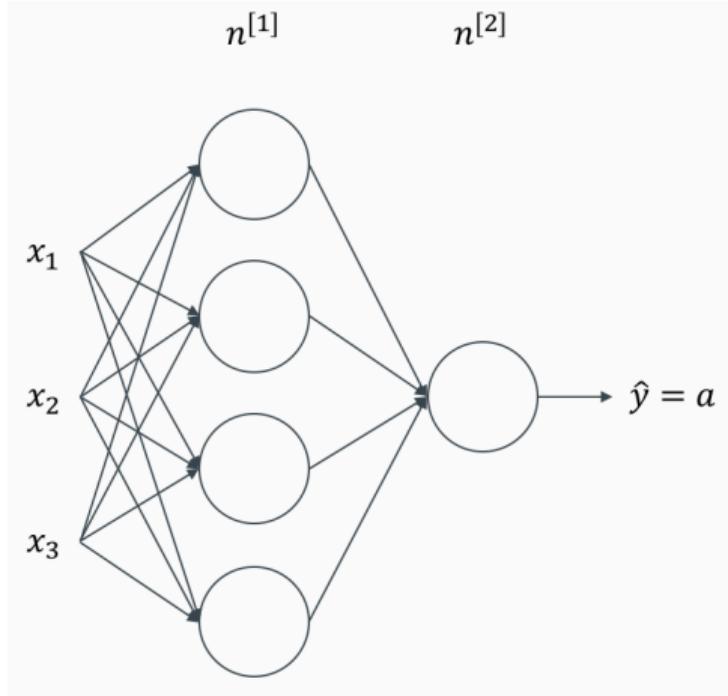
# Logistic Regression



# One Neuron



# (Shallow) Neural Net



## The Different Layers

- Input layer
- Hidden layer
- Output layer

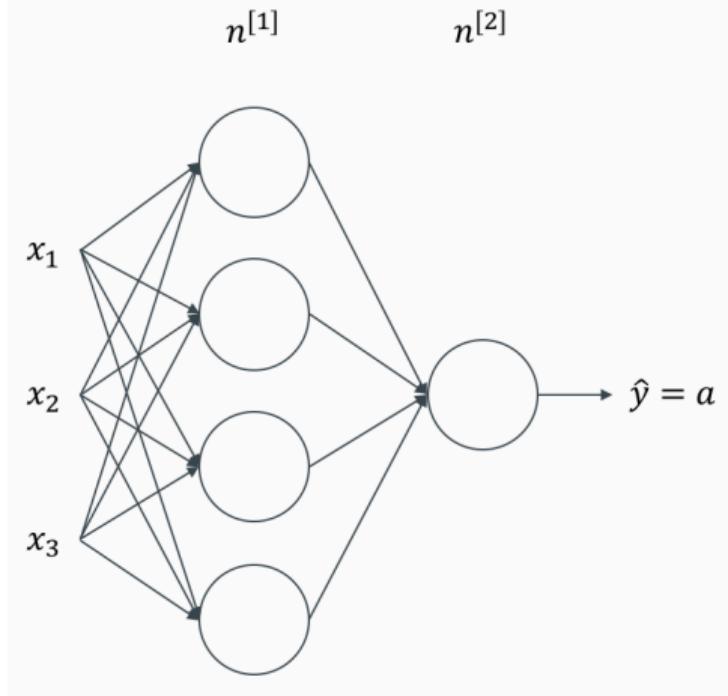
# How to Train Neural Nets?

---

## The Mantra

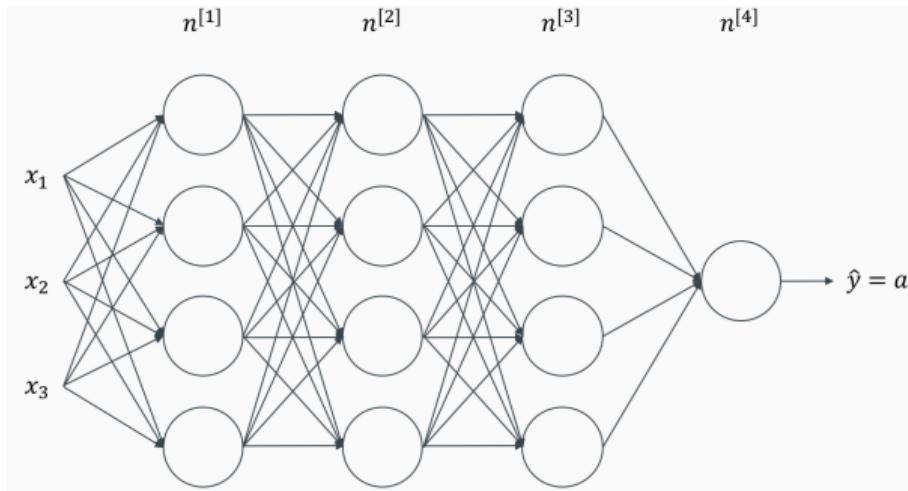
- Predict
- Calculate how wrong the prediction is
- Propagate the information back
- Update weights

# (Shallow) Neural Net



## Four Steps to Repeat

- Predict
- Calculate how wrong the prediction is
- Propagate the information back
- Update weights



## The Different Layers

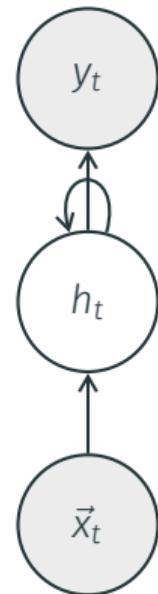
- Predict
- Calculate how wrong the prediction is
- Propagate the information back
- Update weights

# A Geometric Interpretation of a DNN

---

$$a = g(wx + b)$$

# A Simple RNN



**Figure 1:** A simple RNN loops the output of  $h_t$  as an additional input to  $h_{t+1}$ .

## Unrolling the RNN

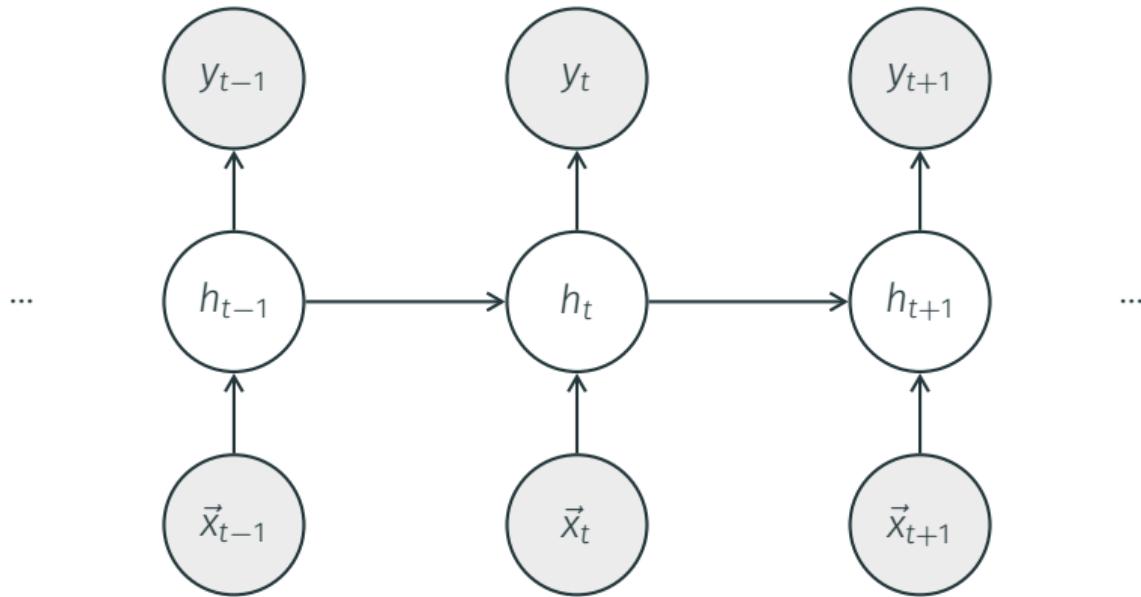


Figure 2: Unrolling the loop shows what happens in the RNN.

# Visual Data Analysis with Convolutional Neural Nets (CNN)

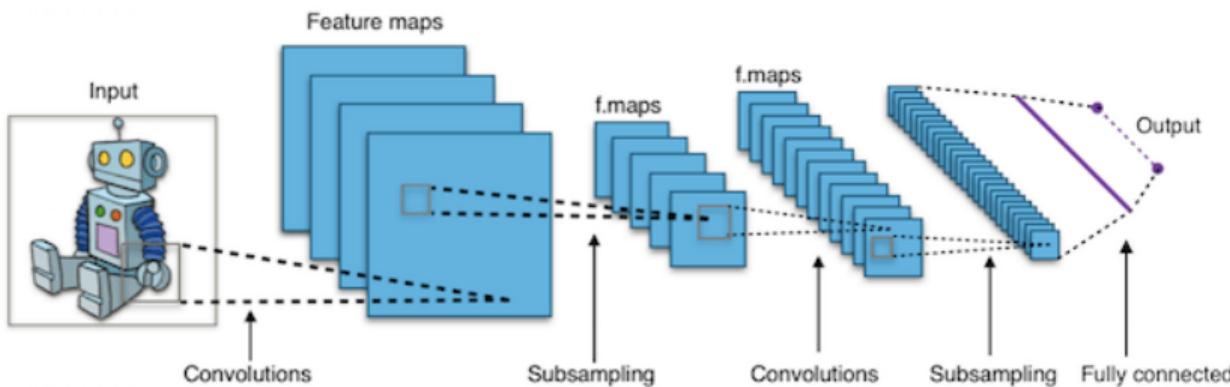
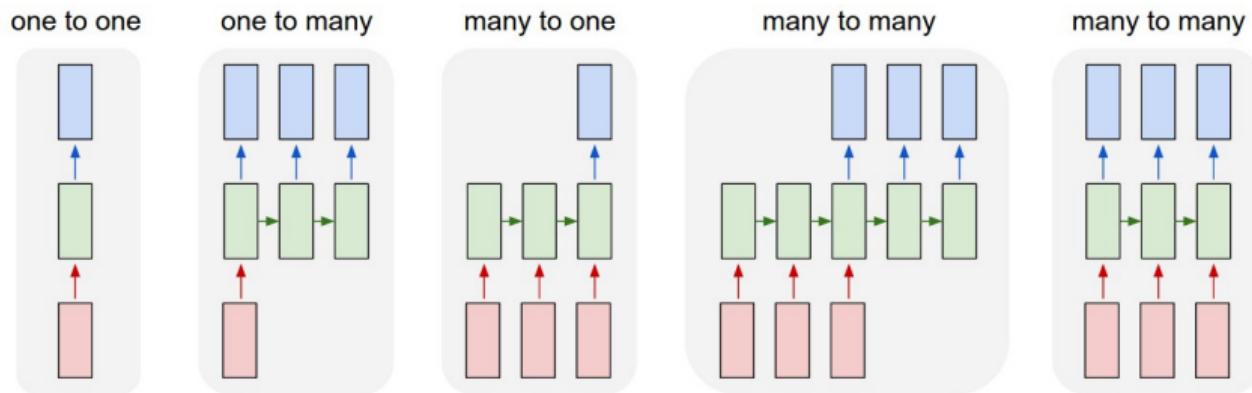


Figure 3: Typical CNN Architecture. Source: Keras Tutorial

# In Sum: A Look at Different NN architectures



**Figure 4:** Different NN architectures. Source: Andrej Karpathy

# Foundations

---

## Multi Modal Data



## Development of Data Sources

- Structured data → unstructured data
- Text data: documents, speeches
- But is a text really all?



## Development of Data Sources

- Structured data → unstructured data
- Text data: documents, speeches
- But is a text really all?

## Social Media Over Time

- Twitter/Facebook
- Instagram/Snapchat
- TikTok/Instagram/YouTube Shorts

## The McGurk Effect (1976)



<https://www.youtube.com/watch?v=PWGeUztTkRA>

# Examples for Multimodal Signals

## Humans

- Language: Words, syntax, speech acts
- Acoustic: Intonation, voice quality, vocal expressions
- Visual: gestures, body language, facial expressions
- Touch
- Any body signal: MRI, ECG,

## Society

- Surveys
- Governmental statistics

## Technology

- Mobile phones
- Internet
- GPS
- Remote sensing

## Modality

“A modality refers to the way in which something happens or is experienced.”  
(Baltrušaitis et al. 2017)

## Spectrum: Closeness to Sensors

- Audio, image or video data
- Language, entity and object detection
- Sentiment intensity, object categories

## Definition

“Multimodality is the joint experience of the same phenomenon through several modalities.” (Liang et al, 2023: 5)

## Spectrum: Homogenous and Heterogenous Modalities

- Same picture from different angles
- Text in two different languages
- Image and its description
- Image and socio economic data

# The Challenges of Multimodal Data

## Heterogeneity

- Information: Which modality to solve a task? How much noise or missing data?
- Data layout: compare image vs. sentence
- Data representation: different atomic units (words, audio, images)

## Connectedness

- Correlation of signals:  
e.g. a clap
- Semantically: e.g. a wink and its words

## Interacting

- In all sorts of ways...

# Cross Modal Interactions

SEPARATE COMPONENTS		MULTIMODAL COMPOSITE SIGNAL		
Redundancy	signal a → <input type="checkbox"/>	signal a + b → <input type="checkbox"/>	response <input type="checkbox"/>	Equivalence (intensity unchanged)
	response b → <input type="checkbox"/>	signal a + b → <input type="checkbox"/>	response <input type="checkbox"/>	Enhancement (intensity increased)
		a + b → <input type="checkbox"/> and <input type="circle"/>		Independence
Nonredundancy	signal a → <input type="checkbox"/>	signal a + b → <input type="checkbox"/>	response <input type="checkbox"/>	Dominance
	response b → <input type="circle"/>	signal a + b → <input type="checkbox"/> (or <input type="checkbox"/> )	response <input type="checkbox"/>	Modulation
		a + b → <input type="triangle"/>		Emergence

Source: Partan and Marler (2005)

# Examples for Multi Modal Data Analysis in Political Sciences

## Rittman, Ringwald and Nyhuis

- Are legislators more likely to deliver emphatic speeches if district preferences on a bill align with their vote choice?
- Video and Audio input to predict speech emphasis in 2-seconds segments

# Examples for Multi Modal Data Analysis in Political Sciences

## Rittman, Ringwald and Nyhuis

- Are legislators more likely to deliver emphatic speeches if district preferences on a bill align with their vote choice?
- Video and Audio input to predict speech emphasis in 2-seconds segments

## Dietrich, Mondak and Williams

- Video, audio and text to identify the emotional intensity associated with responses obtained from in-person, online and telephone interviews.

# Examples for Multi Modal Data Analysis in Political Sciences

## Rittman, Ringwald and Nyhuis

- Are legislators more likely to deliver emphatic speeches if district preferences on a bill align with their vote choice?
- Video and Audio input to predict speech emphasis in 2-seconds segments

## Dietrich, Mondak and Williams

- Video, audio and text to identify the emotional intensity associated with responses obtained from in-person, online and telephone interviews.

## Dietrich and Ko (2022)

- During COVID 19 TV coverage: Is Fauci on the programme and does his presence change the content of the news coverage?
- Image data and text transcripts.

# Foundations

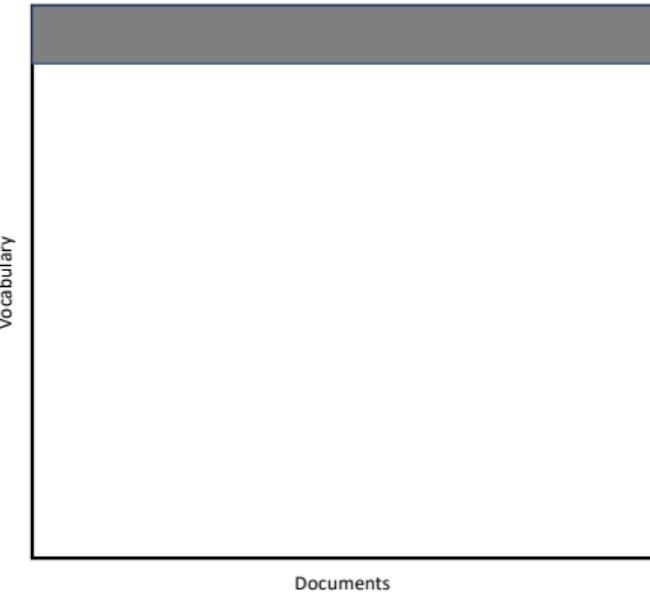
---

## Representation

# Representation in Language

## Term Document Matrix

- Corpus with documents and vocabulary
- Count each word in each document
- Loss of word order and grammar
- Static



# What is the Meaning of ‘Bardiwac’?

---

## What is the Meaning of ‘Bardiwac’?

---

Distribution hypothesis: meaning of a word can be approximated by surrounding words.

## What is the Meaning of ‘Bardiwac’?

---

Distribution hypothesis: meaning of a word can be approximated by surrounding words.

- He handed her a glass of bardiwac.

## What is the Meaning of ‘Bardiwac’?

---

Distribution hypothesis: meaning of a word can be approximated by surrounding words.

- He handed her a glass of bardiwac.
- Beef dishes are made to complement the bardiwacs.

## What is the Meaning of ‘Bardiwac’?

---

Distribution hypothesis: meaning of a word can be approximated by surrounding words.

- He handed her a glass of bardiwac.
- Beef dishes are made to complement the bardiwacs.
- Nigel staggered to his feed, face flushed from too much bardiwac.

## What is the Meaning of ‘Bardiwac’?

---

Distribution hypothesis: meaning of a word can be approximated by surrounding words.

- He handed her a glass of bardiwac.
- Beef dishes are made to complement the bardiwacs.
- Nigel staggered to his feed, face flushed from too much bardiwac.
- Malbec, one of the lesser-known bardiwac grapes, responds well to Australia's sunshine.

## What is the Meaning of ‘Bardiwac’?

---

Distribution hypothesis: meaning of a word can be approximated by surrounding words.

- He handed her a glass of bardiwac.
- Beef dishes are made to complement the bardiwacs.
- Nigel staggered to his feed, face flushed from too much bardiwac.
- Malbec, one of the lesser-known bardiwac grapes, responds well to Australia's sunshine.
- I dined off bread and cheese and this excellent bardiwac.

## What is the Meaning of ‘Bardiwac’?

---

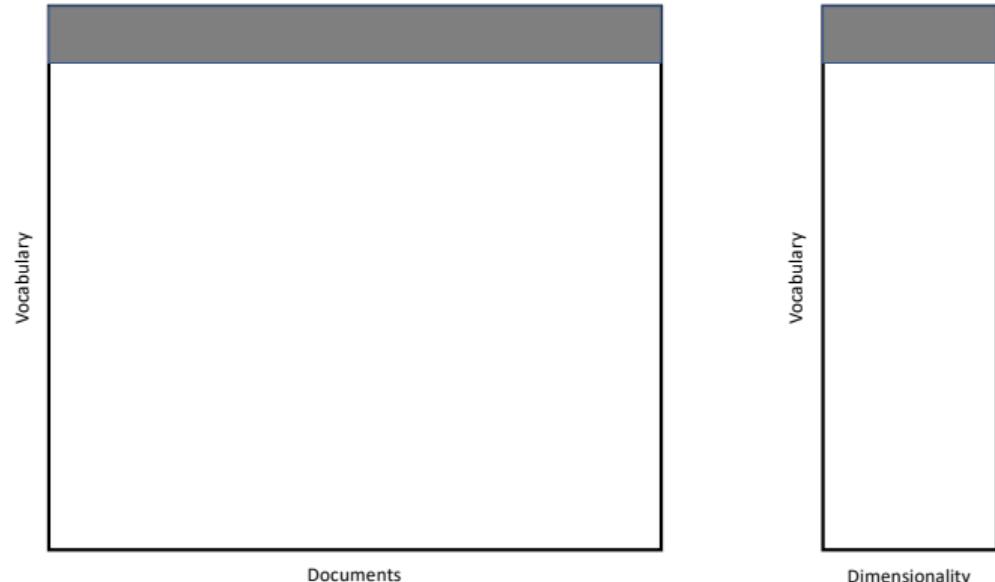
Distribution hypothesis: meaning of a word can be approximated by surrounding words.

- He handed her a glass of bardiwac.
- Beef dishes are made to complement the bardiwacs.
- Nigel staggered to his feed, face flushed from too much bardiwac.
- Malbec, one of the lesser-known bardiwac grapes, responds well to Australia's sunshine.
- I dined off bread and cheeses and this excellent bardiwac.
- The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.

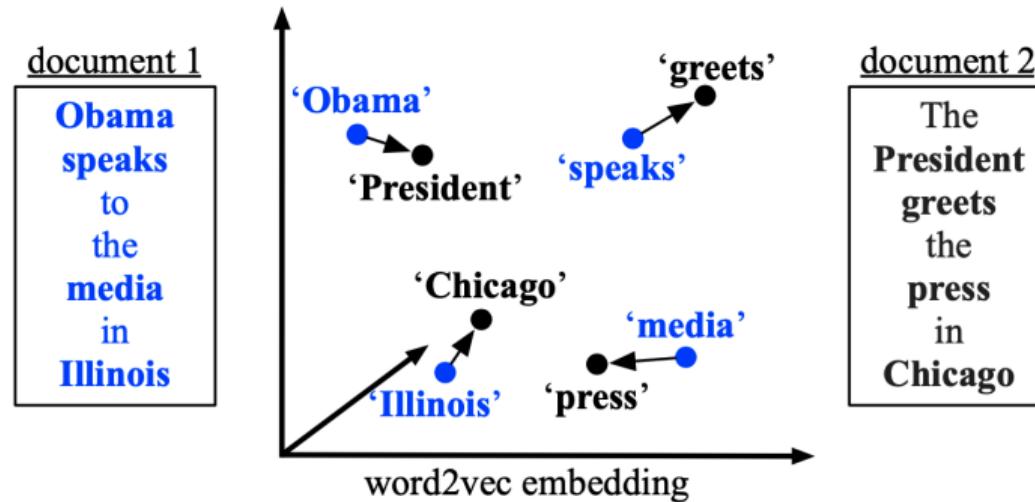
# Word Embeddings

## Evolution of Models

- Word2Vec
- GloVe
- BERT
- RoBERTa
- and of course BART,  
CamemBERT,  
RetriBERT, HerBERT,  
HUBERT ...

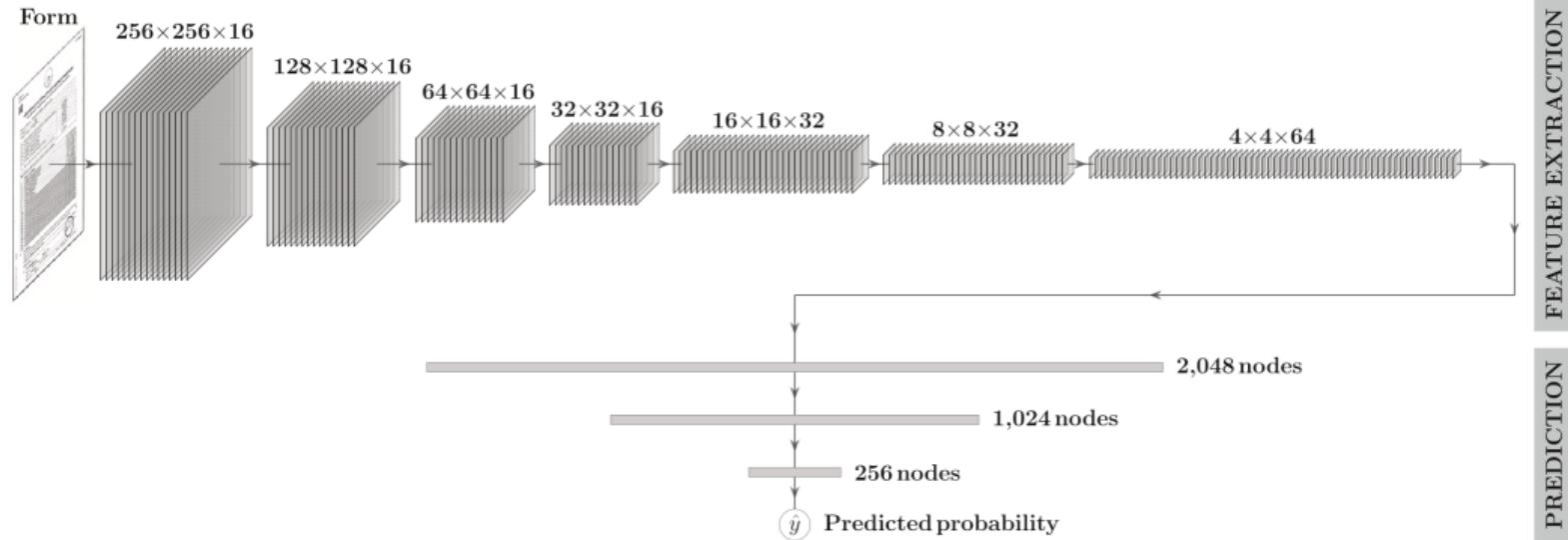


# Doing Things with Wordembeddings

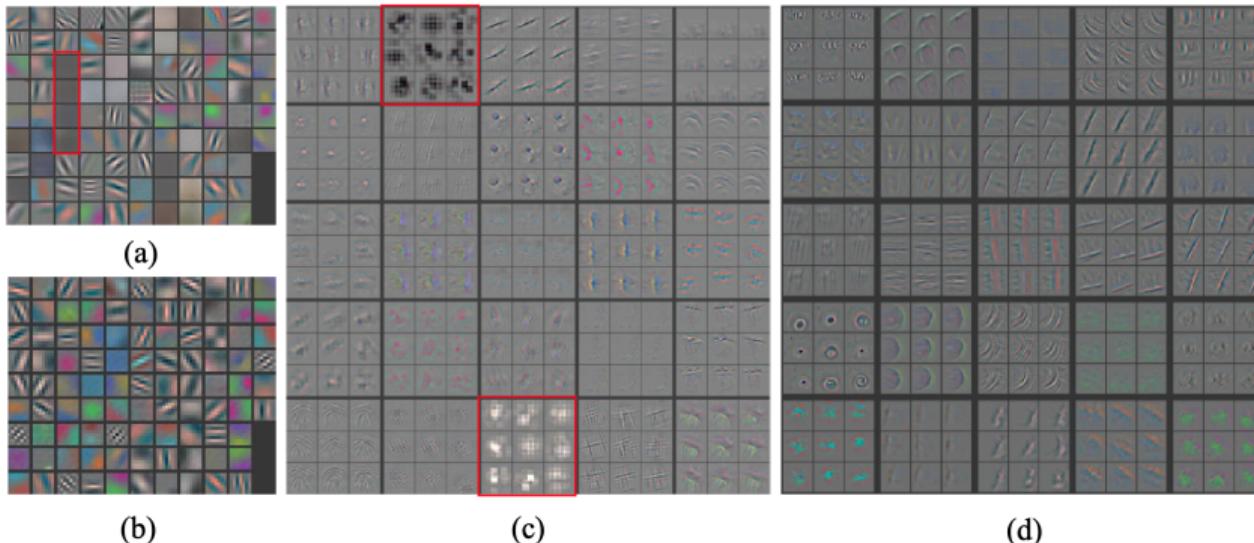


Source: Kusner et al. (2015)

# Visual Embeddings

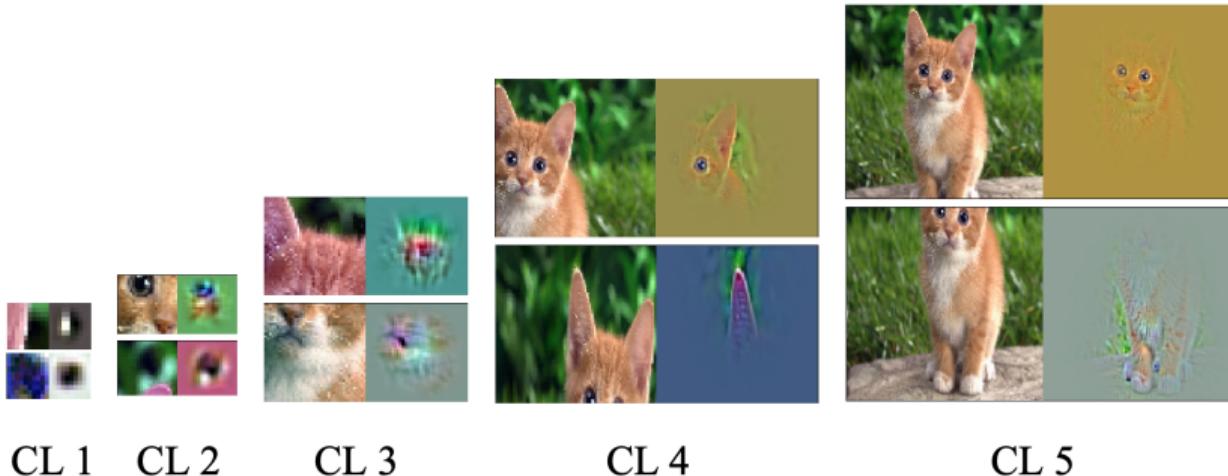


# Filters



Source: Qin et al. (2018)

# Features



Source: Qin et al. (2018)

# Basic Multimodal Learning

---

## Fusion

## Multimodal Fusion

Process of joining information from modalities for prediction

### Examples

- Audio-visual speech recognition
- Audio-visual emotion recognition
- Speaker identification
- ...

## Building Blocks: Analysis Unit

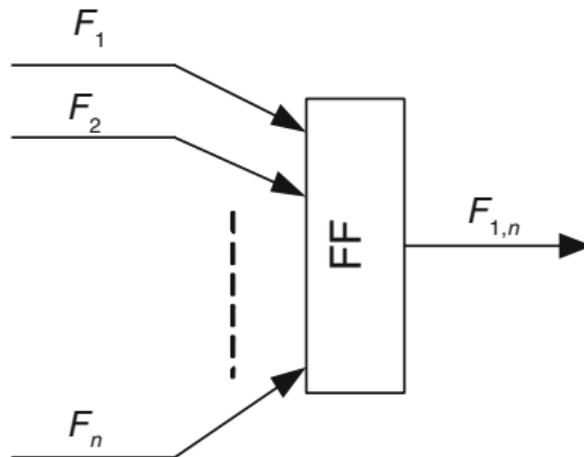


with

- F: Features
- D: Decisions

Source: Atrey et al. (2010)

## Building Blocks: Feature Fusion Unit

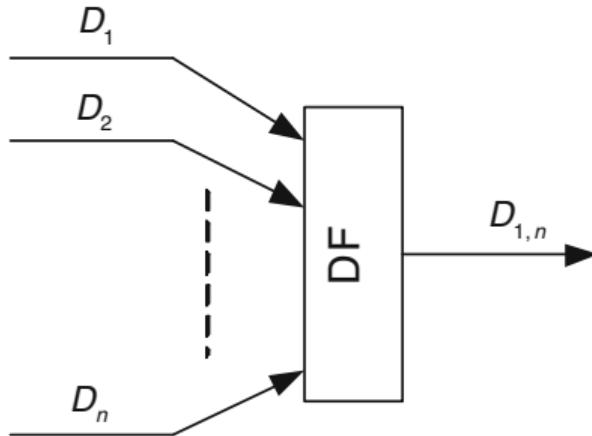


with

- F: Features

Source: Atrey et al. (2010)

## Building Blocks: Decision Fusion Unit

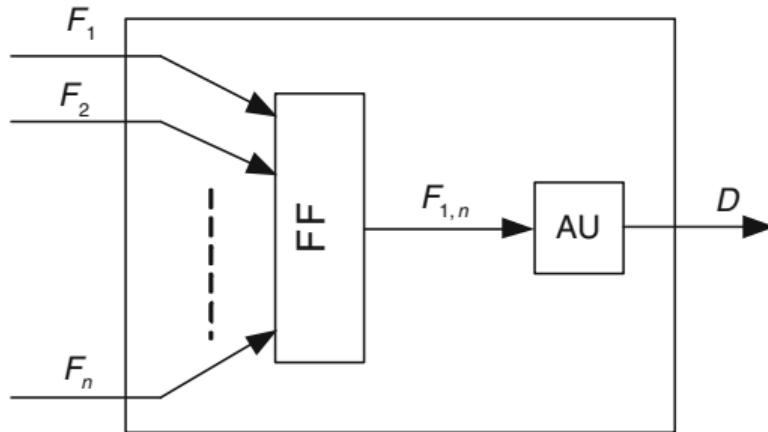


with

- D: Decisions

Source: Atrey et al. (2010)

# Early Fusion/Feature Level MML

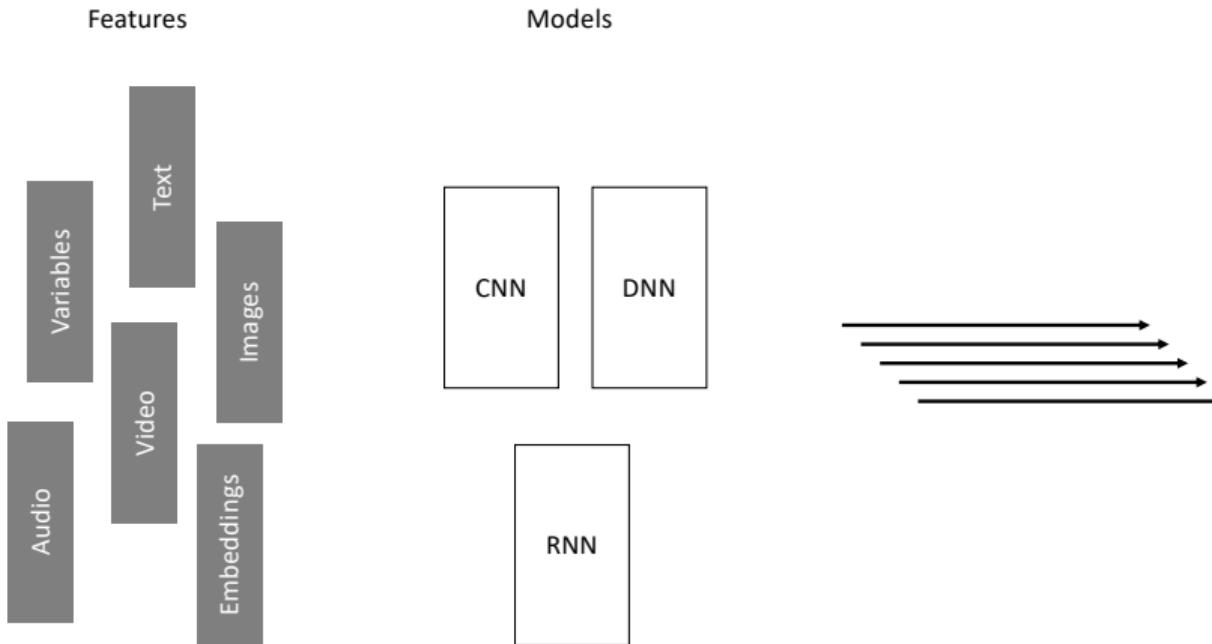


Source: Atrey et al. (2010)

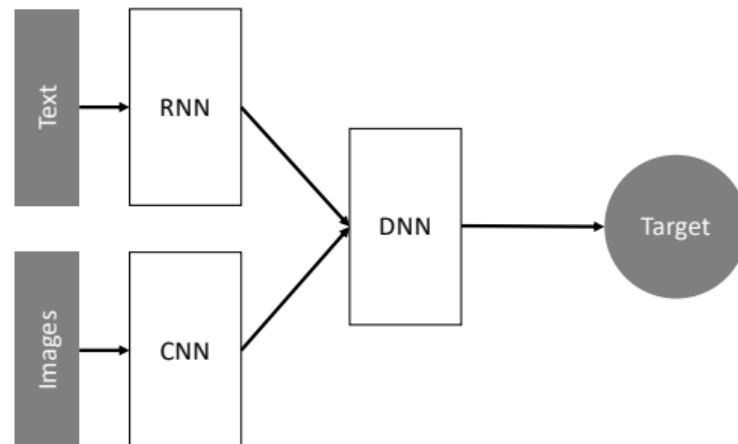
with

- F: Features
- D: Decisions
- Concatenate features
- Uses Correlations between features
- One learning phase
- May be high dimensional
- But: What if different granularities between modalities?

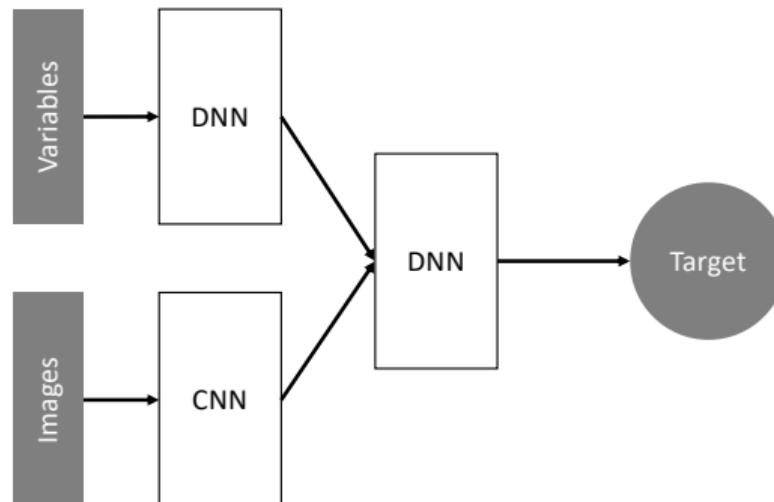
# Quick Recap: What Do We Have to Play With?



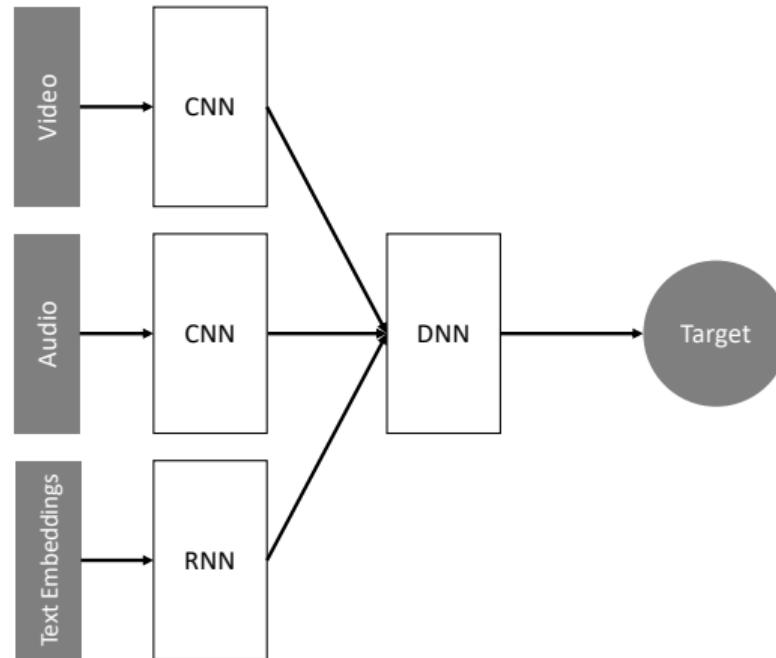
# Early Fusion with Neural Networks



# Early Fusion with Neural Networks



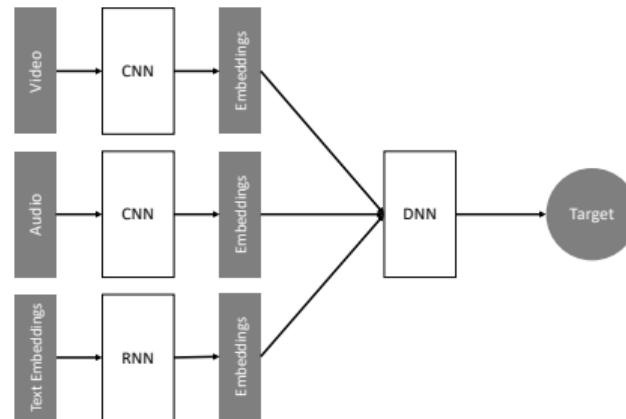
# Early Fusion with Neural Networks



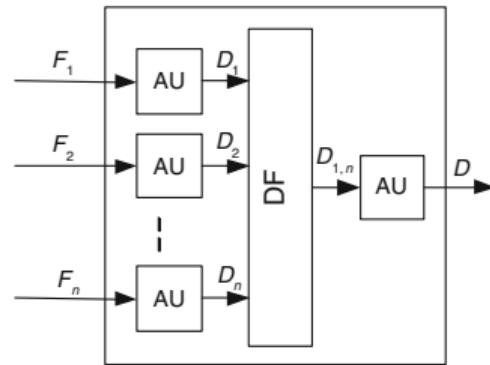
# Early Fusion with Neural Networks

How does the data look like?

- Raw data?
- Features?
- Embeddings?



# Late Fusion/Decision Level MML

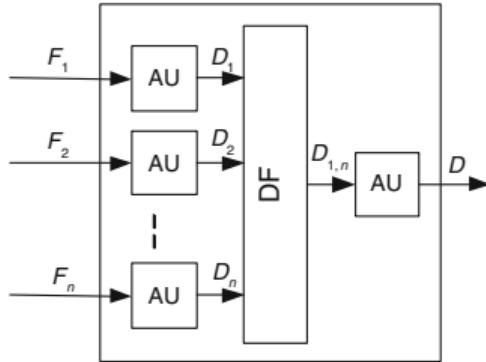


with

- F: Features
- D: Decisions

Source: Atrey et al. (2010)

# Late Fusion/Decision Level MML



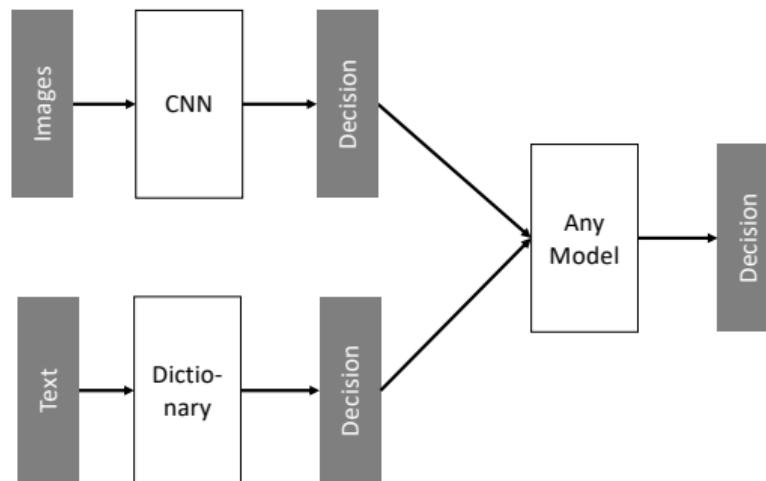
- Multiple training stages: customize each prediction model with domain knowledge
- Fusion mechanisms: voting, weighted sums, or machine learning
- Granularity of features does not need to be the same: decisions synchronize
- No low level interactions between modalities
- Is this ensemble learning?

with

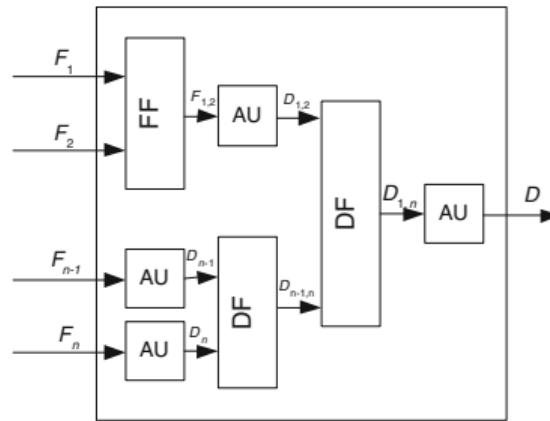
- F: Features
- D: Decisions

Source: Atrey et al. (2010)

# Late Fusion with Neural Networks



# Hybrid Fusion/Hybrid MML

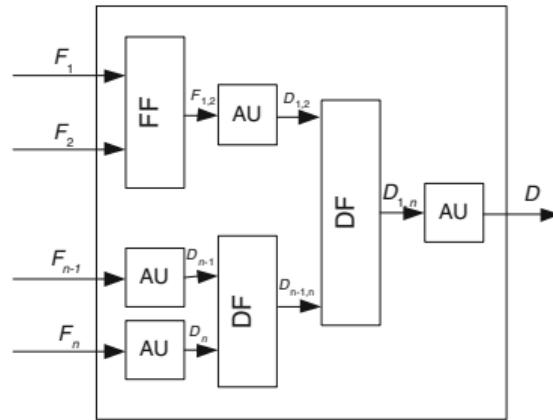


with

- F: Features
- D: Decisions

Source: Atrey et al. (2010)

# Hybrid Fusion/Hybrid MML



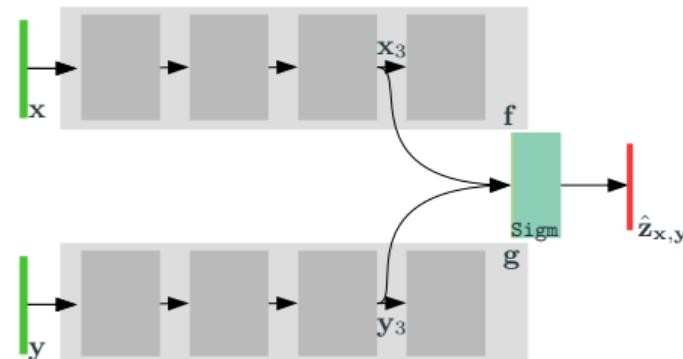
- Can have advantages of early and late fusion, e.g. interactions at lower and higher levels
- But may compound disadvantages, e.g. re training and complexity

with

- F: Features
- D: Decisions

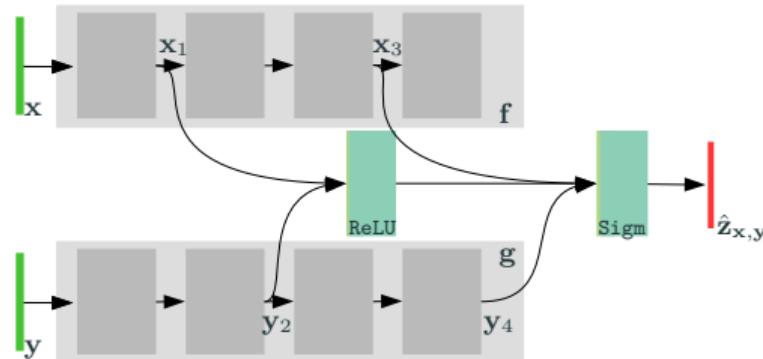
Source: Atrey et al. (2010)

# Hybrid Fusion with Neural Networks



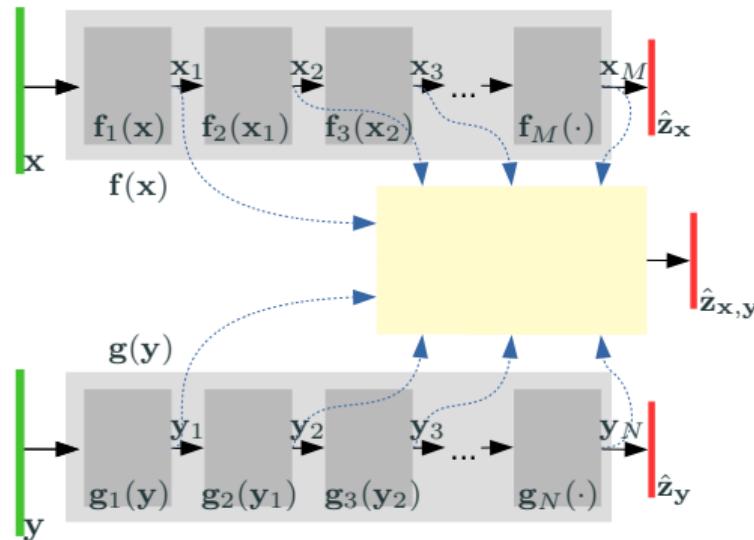
Source: Pérez-Rúa et al. (2019)

# Hybrid Fusion with Neural Networks



Source: Pérez-Rúa et al. (2019)

# Hybrid Fusion with Neural Networks: Fusion Layer Unit



Source: Pérez-Rúa et al. (2019)

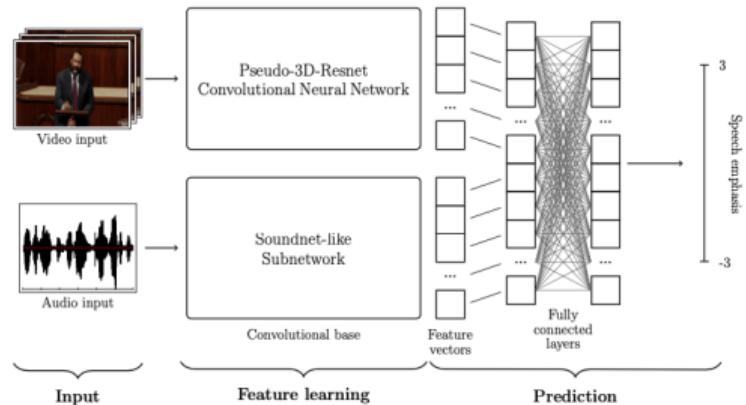
## How to Build a Multi Modal Model with Keras

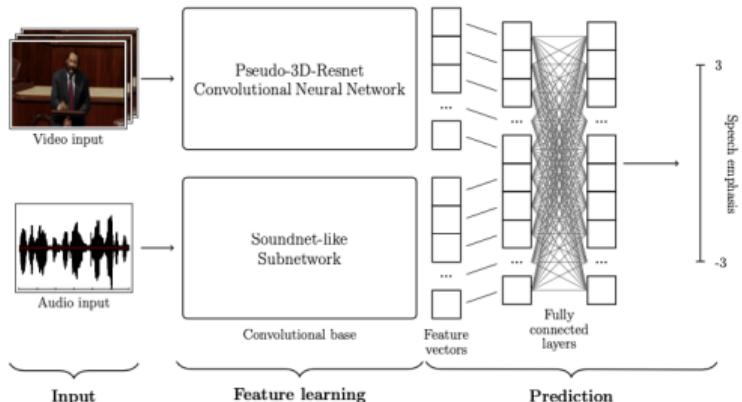
- Sequential and Functional API
- Model Fusion
- Code Nr. 1 Google Collab: <https://colab.research.google.com/drive/1ZKqndxBsSZ2Yuu-kgafJnAgvnYDfHDjm?usp=sharing>

# Basic Multimodal Learning

---

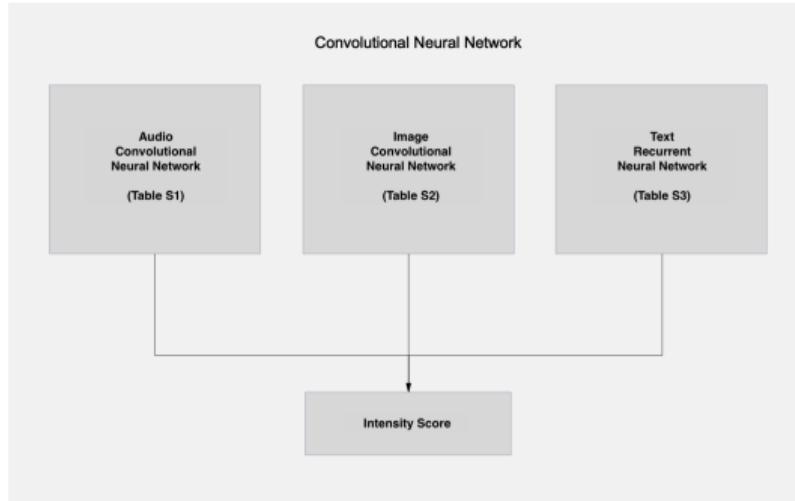
Examples from Political Science  
Revisited

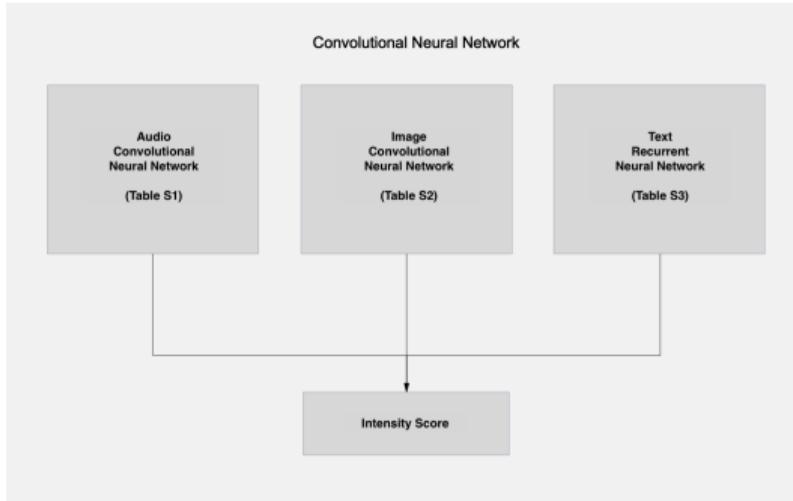




## Modelling

- Syncronised video input and audio input
- Early fusion
- Fusion mechanism: concatenation
- Training: Adam with low learning rate

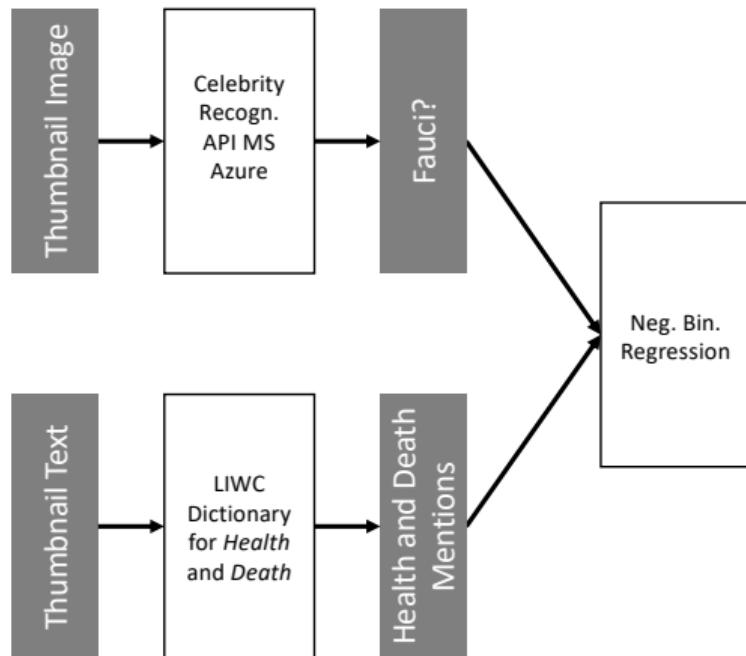


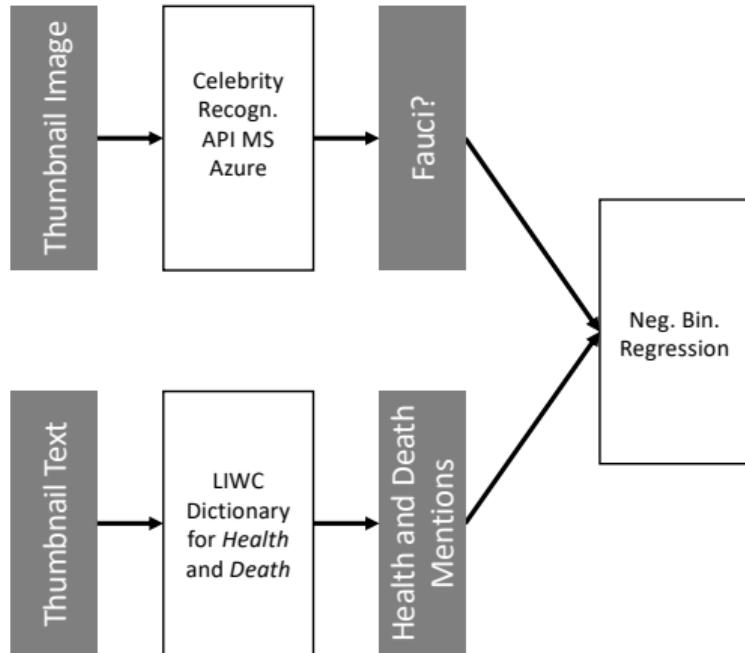


## Modelling

- Syncronised video, audio and text input at word level with KALDI (Ochshorn and Hawkins 2016)
- Early fusion
- Fusion mechanism: concatenation

# Dietrich and Ko (2022): *Finding Fauci*





## Modelling

- Internet archive: 2020 programming from CNN, Fox News, and MSNBC with search terms Covid and coronavirus
- One minute segments with closed-captioning and thumbnails
- Late fusion
- Low level models: proprietary “AI” and dictionaries
- High level model: Negative Binomial regression
- Fusion mechanism: decision level
- Syncronisation: decisions on one minute segments

## Predicting House Prizes

- Predict the prize of a house on the basis of its images and characteristics
- Regression
- Code Nr. 2 Google Collab: [https://colab.research.google.com/drive/1HVb5yBkN1TSiCo\\_1Nf4RW6RQ6aZgJ08w?usp=sharing](https://colab.research.google.com/drive/1HVb5yBkN1TSiCo_1Nf4RW6RQ6aZgJ08w?usp=sharing)

# Basic Multimodal Learning

---

## Learning and Optimization

# Learning and Optimization



## Challenge

- Data is from different sources.
- Different subgraphs

## Example: Sentiment with Images and Captions

- CNN require high decaying learning rate
- RNN like adaptive methods and SGD
- DNN like adaptive methods

## Solution: Pre-training

- Train each modality separately
- Assemble and fine-tune
- Adaptive methods (e.g. Adam) are not good for fine tuning because of their high initial momentum

## Example: Sentiment with Images and Captions

1. Train CNN & RNN separately on sentiment
2. Generate visual and verbal representations
3. Train DNN with representations on sentiment
4. Fine tune all on sentiment

## Solution: Transfer Learning

- VGG for CNN
- Language models for RNN
- Train DNN layer-by-layer

# Extensions

---

## Representation Fusion

## Definition

Learn a joint representation that models cross-modal interactions between individual elements of different modalities.

## Fusing Modalities is a Continuum

- Homogenous modalities, e.g. two different camera angles
- Heterogenous modalities, e.g. a picture and text

# A Closer Look at Fusion

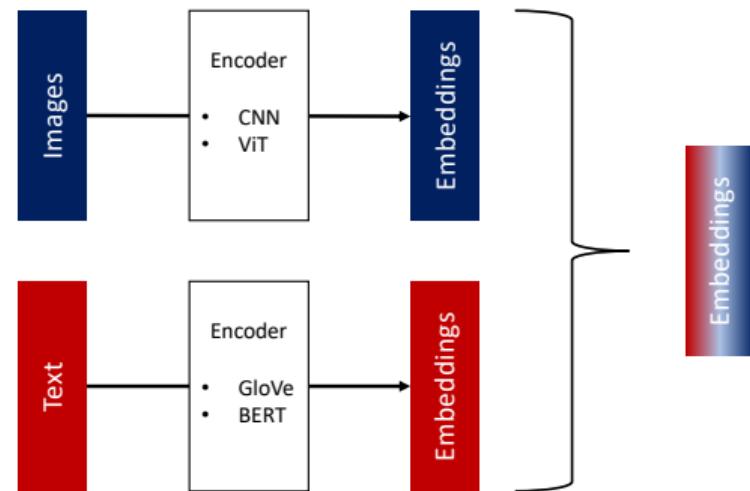
## Definition

Learn a joint representation that models cross-modal interactions between individual elements of different modalities.

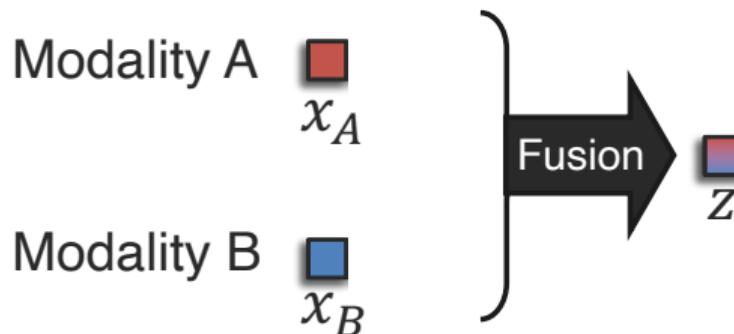
## Fusing Modalities is a Continuum

- Homogenous modalities, e.g. two different camera angles
- Heterogenous modalities, e.g. a picture and text

## Challenge: How to Fuse Heterogenous Modalities?



# Unimodal Univariate Case



Additive Interaction

$$z = w_1x_1 + w_2x_2 + \epsilon$$

Multiplicative Interaction

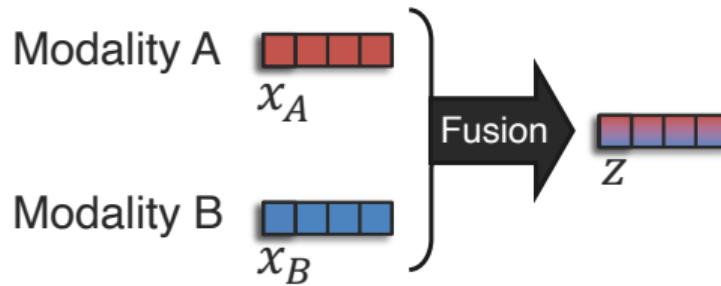
$$z = w_3(x_1x_2) + \epsilon$$

Additive and Multiplicative Interaction

$$z = w_1x_1 + w_2x_2 + w_3(x_1x_2) + \epsilon$$

Source: Morency et al. (2022)

# Unimodal Multivariate Case



Additive Interaction

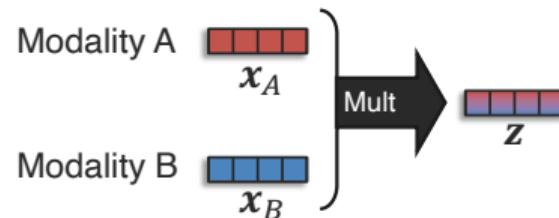
$$z = w_1x_1 + w_2x_2 = W \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Source: Morency et al. (2022)

# Unimodal Multivariate Case

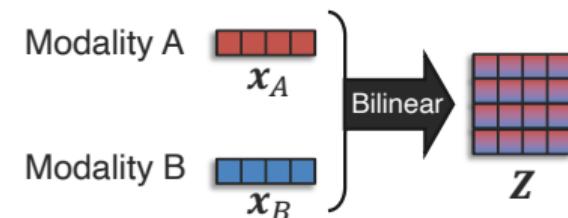
## Multiplicative Fusion

$$z = w(x_1 \times x_2)$$



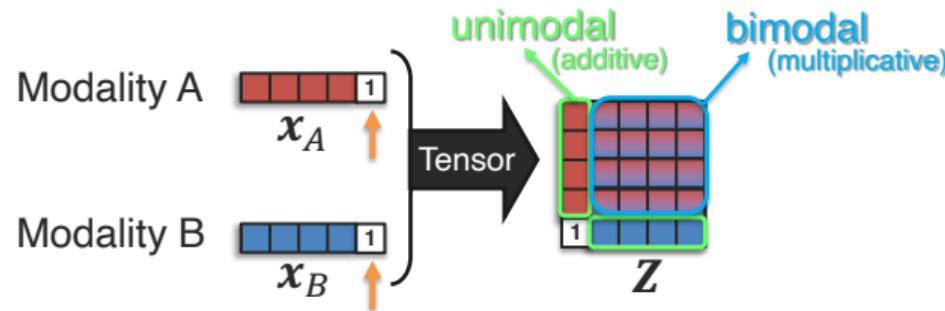
## Billinear Fusion

$$Z = w(x_1^T \times x_2)$$



Source: Morency et al. (2022)

# Unimodal Multivariate Case



## Tensor Fusion

$$Z = w([x_1]^\top \times [x_2])$$

Source: Morency et al. (2022), Zadeh et al. (2017)

# Extensions

---

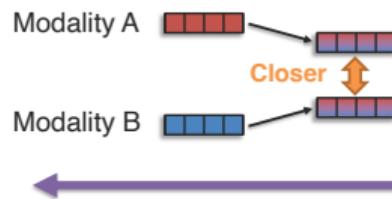
## Representation Coordination

# Representation Coordination

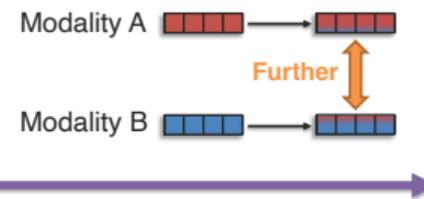
## Definition

Learn multimodally-contextualized representations that are coordinated through their cross-modal interactions.

### Strong Coordination:

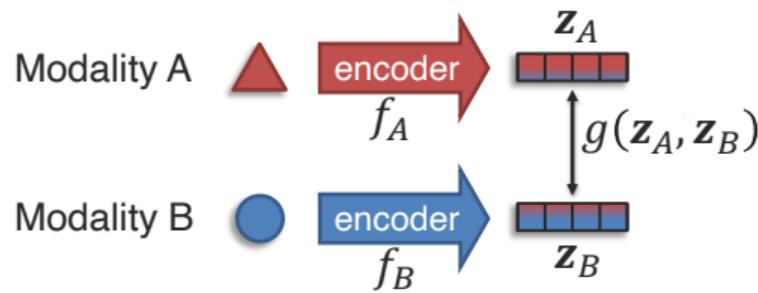


### Partial Coordination:

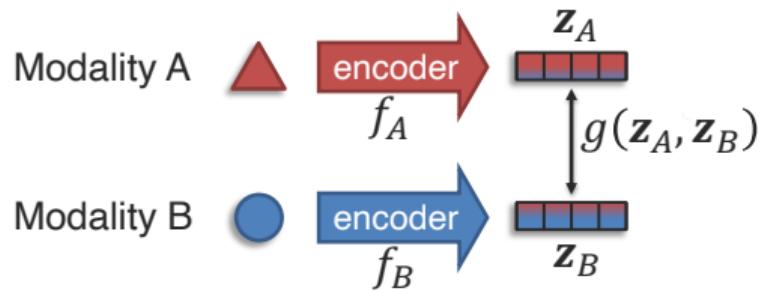


Source: Morency et al. (2022)

## Examples for Coordination Functions $g()$ :



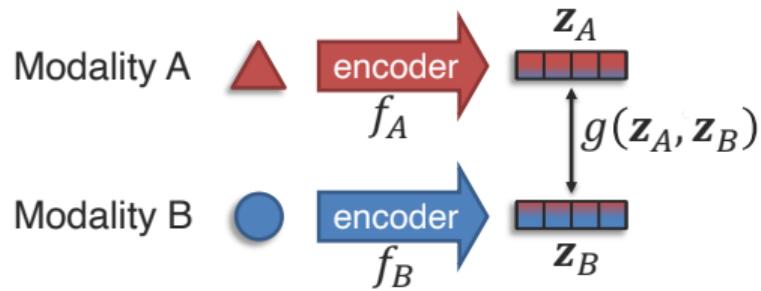
## Examples for Coordination Functions $g()$ :



Cosine Similarity

$$g(\mathbf{z}_A, \mathbf{z}_B) = \frac{k(\mathbf{z}_A, \mathbf{z}_B)}{\|\mathbf{z}_A\| \|\mathbf{z}_B\|}$$

## Examples for Coordination Functions $g()$ :



Cosine Similarity

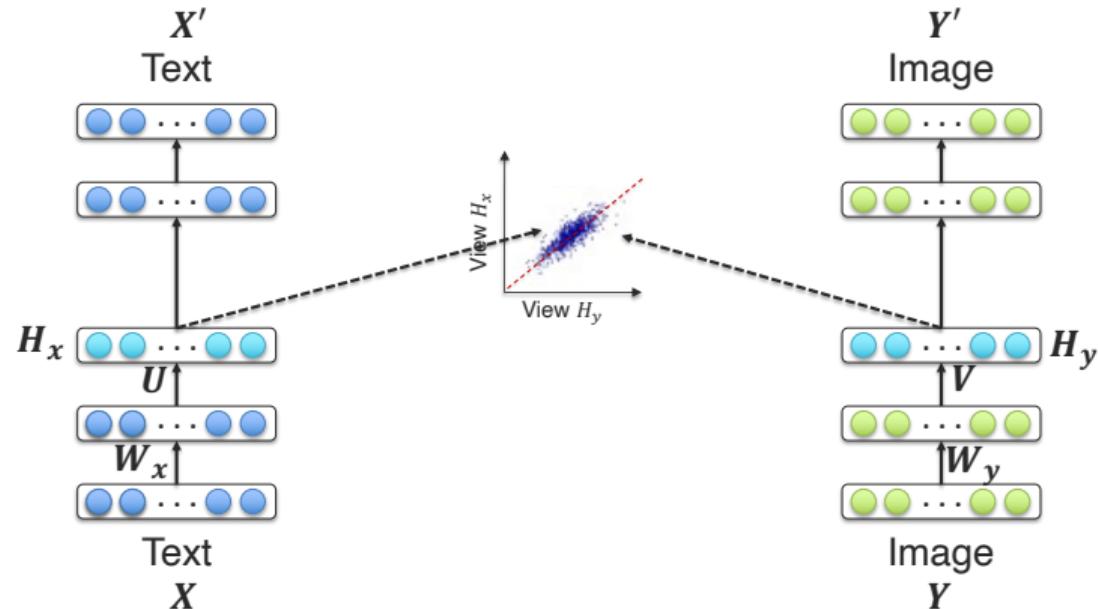
$$g(\mathbf{z}_A, \mathbf{z}_B) = \frac{k(\mathbf{z}_A, \mathbf{z}_B)}{\|\mathbf{z}_A\| \|\mathbf{z}_B\|}$$

Canonical Correlation Analysis (CCA)

$$\underset{V, U, f_A, f_B}{\operatorname{argmax}} \operatorname{corr}(z_A, z_B)$$

Source: Morency et al. (2022)

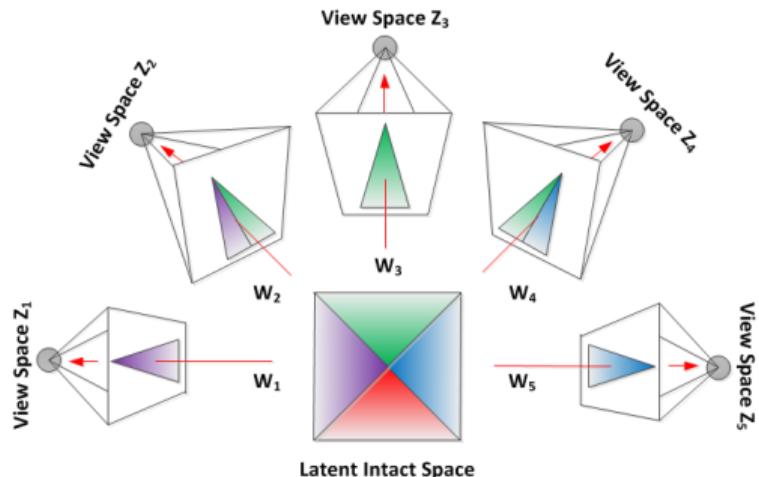
# Deep Canonically Correlated Autoencoders (DCCAE)



- Cost function?
- How to train?

Source: Morency et al. (2022), Wang et al. (2015)

# Intuition: Why is a Coordinated Embedding Space Useful?



## Multiple Views from the Same Object

- Intact representation that is complete and not damaged
- Individual views  $z_i$  are partial representations of the intact representation

Source: Xu and Tao (2015)

# Intuition: Why is a Coordinated Embedding Space Useful?



- day + night =
- flying + sailing =
- bowl + box =
- box + bowl =



Nearest images



- blue + red =
- blue + yellow =
- yellow + red =
- white + red =



Nearest images

Source: Kiros et al. (2014)

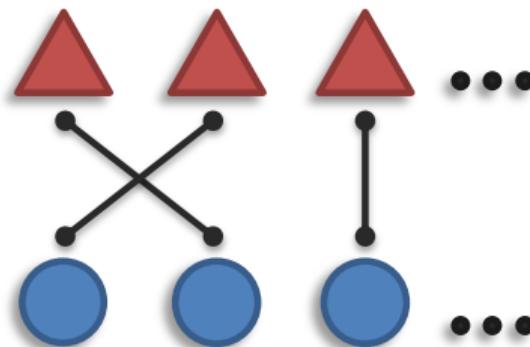
# Extensions

---

## Alignment

## Definition

“Finding relationships and correspondences between sub-components of instances from two or more modalities.” (Baltrušaitis et al., 2018: 432)



## Connection Types

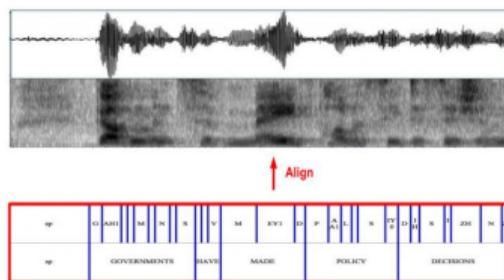
- Equivalence: No extra information
- Correspondence: Enhanced information
- Dependencies: Complementary information

Source: Morrency et al. (2022)

# Explicit Multimodal Alignment

## In a Nutshell

- Alignment is the task itself.
- → Alignment is the loss function.



## Examples

- Images and captions
- Speech signal to transcript
- Co-referring expressions



Source: Morency (2020)

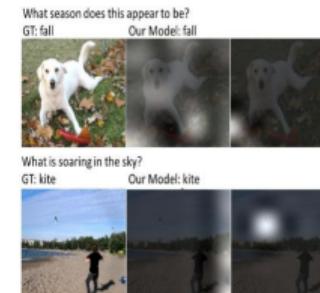
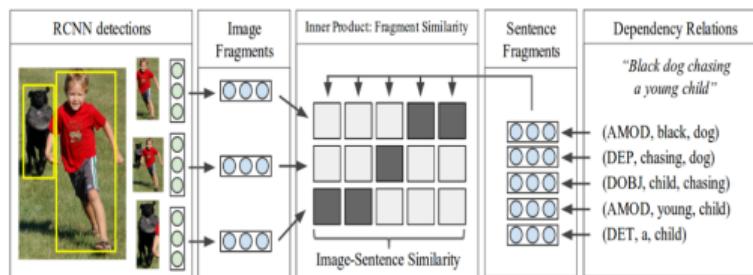
# Implicit Alignment

## In a Nutshell

- Alignment is an intermediate step that is built into the model
- Goal of the task is something else.
- If the model aligns well, the model performance will be higher.

## Examples

- Machine translation
- Visual question answering
- Cross-modal retrieval



# How Do Political Scientists Align?

---

## Global Level

- Dietrich, Enos, and Sen, 2019: Verbal contribution level
- Dietrich, Hayes, and O'Brien, 2019: Speech level
- Rittmann, forthcoming: Speech level
- Tarr, Hwang, and Imai, 2022: TV campaign ads

## Time Based

- Dietrich and Ko, 2022: 1 minute interval
- Rittmann et al., 2020: 2 seconds interval
- Boussalis et al., 2021: 1 second interval
- Knox and Lucas, 2021: 25 milliseconds interval

## Alignment Annotates Data Across Modalities

- Labels
- Model

## Alignment Annotates Data Across Modalities

- Labels
- Model

## Multimodal Data is Heterogeneous

- Data quality in individual modalities
- Overall performance improvement

## Alignment Annotates Data Across Modalities

- Labels
- Model

Multimodal Data is Heterogeneous

- Data quality in individual modalities
- Overall performance improvement

Multimodal Data is Connected

- New features
- Crossmodal queries

## Alignment Annotates Data Across Modalities

- Labels
- Model

### Multimodal Data is Heterogeneous

- Data quality in individual modalities
- Overall performance improvement

### Multimodal Data is Connected

- New features
- Crossmodal queries

### Multimodal Data is Interacting

- Model interpretability
- Human understanding of model interpretability

## Data: Political Ads in the 2020 US Presidential Race

- Source: Wesleyan Media Project
- Target/DV: How positive or negative is the ad? (tone)

## Alignment with Cross Modal Attention

- Cross Modal Attention Architecture (Tsai et al. 2020)
- Features
  - Text: GloVe Embeddings
  - Video: 41 facial action units from every 4th frame (Luo et al. 22)
  - Audio: 34 audio from pyAudioAnalysis (Giannakopoulos 2015)

## Alignment to Improve Predictions (Arnold and Küpfer 2023)

---

Model	F1-Score
Audio Only	82.22
Vision Only	81.92
Text Only	92.79
A+V+T no crossalignment	91.67
A+V+T with crossalignment	95.07

## Data: Speeches in the 19th German Parliament

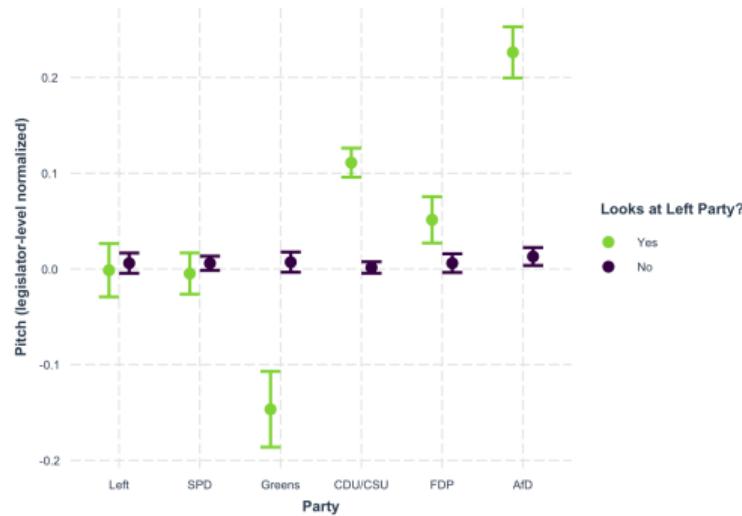
- Speeches from the *Bundestag* homepage
- Identify videos with speaker on the basis of the lettering of the podium



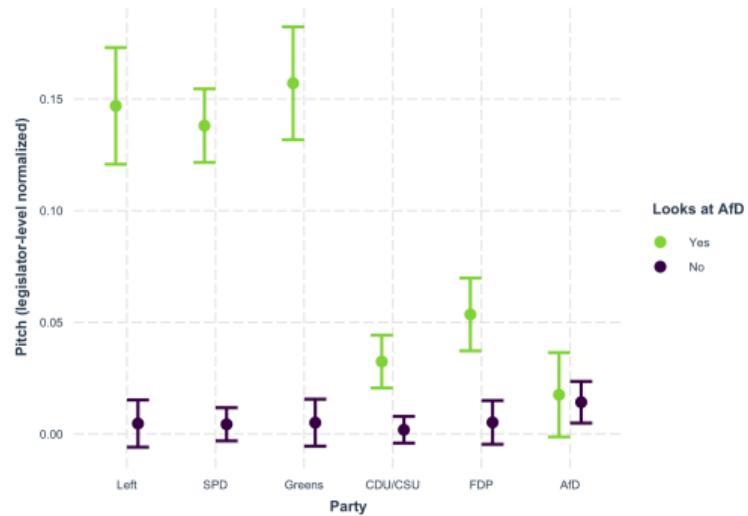
## Explicit Alignment

- Transcription at word level: WhisperX (Bain et al., 2023)
- Speaker gaze at extreme factions: SixDRepNet (Hempel et al., 2022)

The Pitch When Looking at *Die Linke*



The Pitch When Looking at AFD



## Multimodal Entailment: Does a Social Media Post Contradict or Imply Each Other?

- Recognizing Multimodal Entailment tutorial was held virtually at ACL-IJCNLP 2021 on August 1st.
- Event: <https://multimodal-entailment.github.io/>
- Baseline model:  
[https://keras.io/examples/nlp/multimodal\\_entailment](https://keras.io/examples/nlp/multimodal_entailment)

## Readings

- Barua, Arnab, Mobyen Uddin Ahmed, and Shahina Begum. "A Systematic Literature Review on Multimodal Machine Learning: Applications, Challenges, Gaps and Future Directions." *IEEE Access* (2023).
- Liang, Paul Pu, Amir Zadeh, and Louis-Philippe Morency. "Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions." *arXiv preprint arXiv:2209.03430* (2022).
- Parcalabescu, Letitia, Nils Trost, and Anette Frank. "What is multimodality?" *arXiv preprint arXiv:2103.06304* (2021).
- Xu, Peng, Xiatian Zhu, and David A. Clifton. "Multimodal learning with transformers: A survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
- Goodfellow, Ian and Yoshua Bengio and Aaron Courville. 2016. *Deep Learning*. MIT Press.

## Internet Resources

- Louis-Philippe Morency. *Multimodal Machine Learning*.  
<https://cmu-multicomp-lab.github.io/mmmml-course/fall2020/>.
- Morency, Louis-Philippe, Paul Liang and Amir Zadeh. 2022. *Tutorial on MultiModal Machine Learning*. NAACL 2022.  
<https://cmu-multicomp-lab.github.io/mmmml-tutorial/schedule/>.
- Morency, Louis-Philippe, Paul Liang and Amir Zadeh. 2022. *Tutorial on MultiModal Machine Learning*. CVPR 2022.  
<https://cmu-multicomp-lab.github.io/mmmml-tutorial/cvpr2022/>.

Working with Deep Learning? **Get in touch!**

*Dr Christian Arnold*

Cardiff University

@chrisguarnold

# Dynamic Time Warping: Unimodal

Two unaligned unimodal temporal signals

$$X = [x_1, x_2, \dots, x_{n_x}]$$

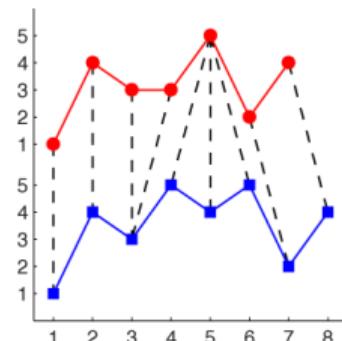
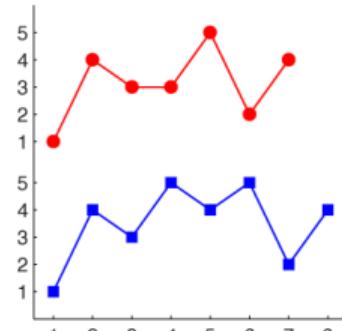
$$Y = [y_1, y_2, \dots, y_{n_y}]$$

Find set of indices to minimize alignment difference

$$L(p^x, p^y) = \sum_{t=1}^l \|x_{p_t^x} - y_{p_t^y}\|_2^2$$

→ Dynamic Time Warping is designed to find the vectors  $p^x$  and  $p^y$

Source: Morency (2020)

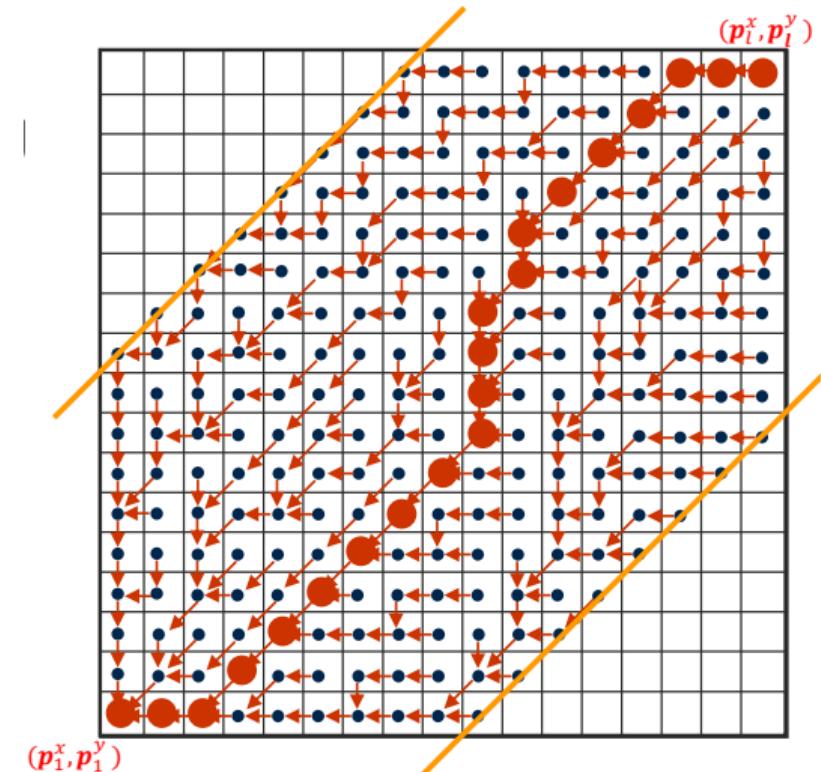


# Dynamic Time Warping

Solve: Lowest Cost Path in Cost Matrix

- Monotonicity: no going back
- Continuity: no gaps
- Boundary conditions: start and end at same points
- Warping window: don't get too far from diagonal
- Slope constraint: Do not insert or skip too much

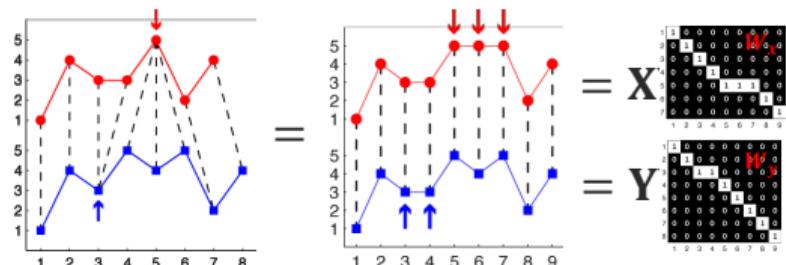
Source: Morency (2020)



# Dynamic Time Warping: Alternative Formulation

Objective So Far

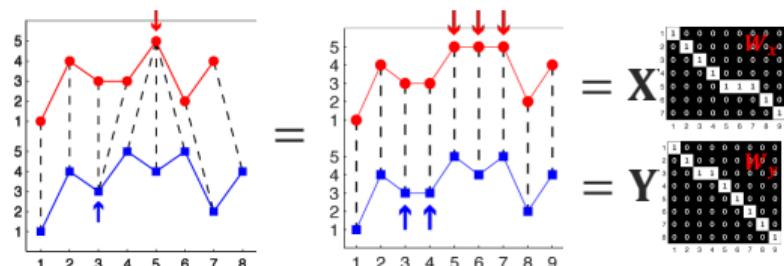
$$L(p^x, p^y) = \sum_{t=1}^l \|x_{p_t^x} - y_{p_t^y}\|_2^2$$



# Dynamic Time Warping: Alternative Formulation

Objective So Far

$$L(p^x, p^y) = \sum_{t=1}^l \|x_{p_t^x} - y_{p_t^y}\|_2^2$$



Alternative Objective

$$L(W_x, W_y) = \|XW_x - YW_y\|_F^2$$

with

- $X, Y$  original signals
- $W_x, W_y$  alignment matrices to find

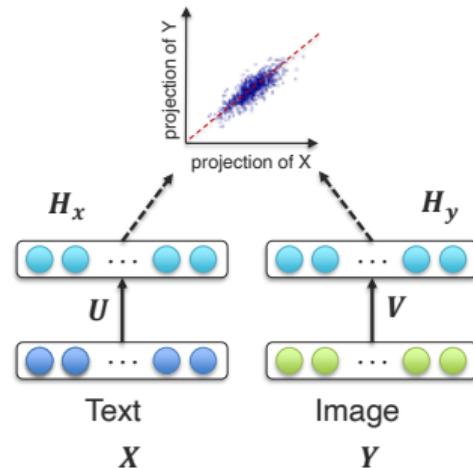
But: Computationally complex, sensitive to outliers, unimodal

Source: Morency (2020)

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

## Canonical Correlation Analysis

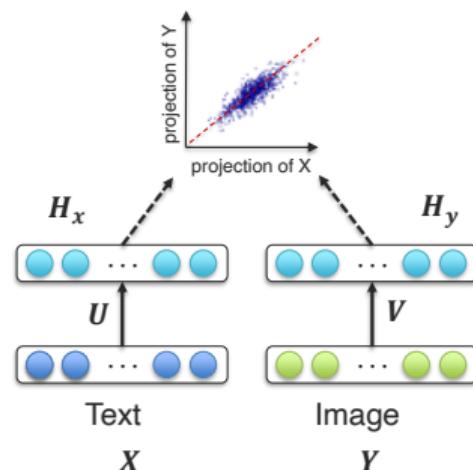
$$L(U, V) = \|U^T X - V^T Y\|_F^2$$



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

## Canonical Correlation Analysis

$$L(U, V) = \|U^T X - V^T Y\|_F^2$$



## Canonical Time Warping

$$L(U, V, W_x, W_y) = \|U^T X W_x - V^T Y W_y\|_F^2$$

- Allows to align multi modal and multi view
- $W_x, W_y$  for temporal alignment
- $U, V$  for cross-modal alignment

Source: Morrency (2020), Zhou and De la Torre (2009)

# Canonical Time Warping + NN = Deep Canonical Time Warping

## Canonical Time Warping

$$L(U, V, W_x, W_y) = \|U^T X W_x - V^T Y W_y\|_F^2$$

- $W_x, W_y$  for temporal alignment
- $U, V$  for cross-modal alignment

## Deep Canonical Time Warping

$$L(\theta_1, \theta_2, W_x, W_y) = \|f_{\theta_1(X)} X W_x - f_{\theta_2(Y)} Y W_y\|_F^2$$

- Addresses non-linearities in alignment

Source: Zhou and De la Torre (2009), Trigeorgis et al. (2016)

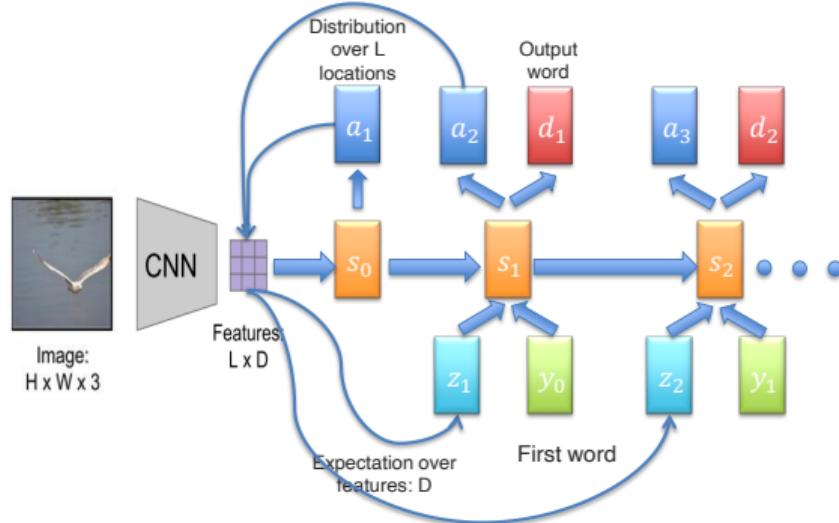
# Visual Captioning with Soft Attention



Source: Xu et al. (2015)

Multi Modal DL | DS3 Hertie School | August 25, 2023

# Soft Attention to Look at More Fine Grained Features



Source: Morency (2020)

## Attention

- Output from response maps to the convolution (e.g.  $4 \times 4 \times 64$ )
- Where to focus on the image
- For each location in the image ( $1 \times 64$ )

## Im Sum

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