Deep Learning for Multi-Modal Systems

DS3 Data Science Summer School

Dr Christian Arnold (Cardiff University)

August 4, 2022

Logistics

- The repo with all material is at `https://github.com/chrisguarnold/ds3_mmdl`
- Please do ask questions in the chat: Huy and Olga will handle them.

My Background

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 DL, e.g.
  - **Voting Fraud Detection:** Multimodal data (satellite and micro data) to detect voting irregularities in remote areas.
  - **Private Synthetic Data:** Generative Adversarial Networks for differentially private synthetic micro data.

Industry Background

- Data Scientist with KIANA and KPMG
- Scientific advisory committee for the GSS, Office for National Statistics
- Alan Turning Institute Networking Grant for AI in Public Policy

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. Foundation
3. Basic Multimodal Learning
4. Extensions

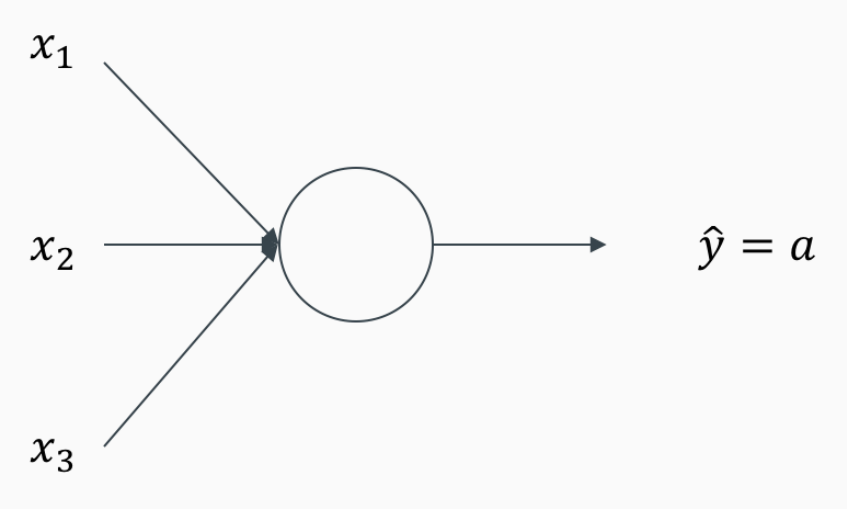
Today's Learnings

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

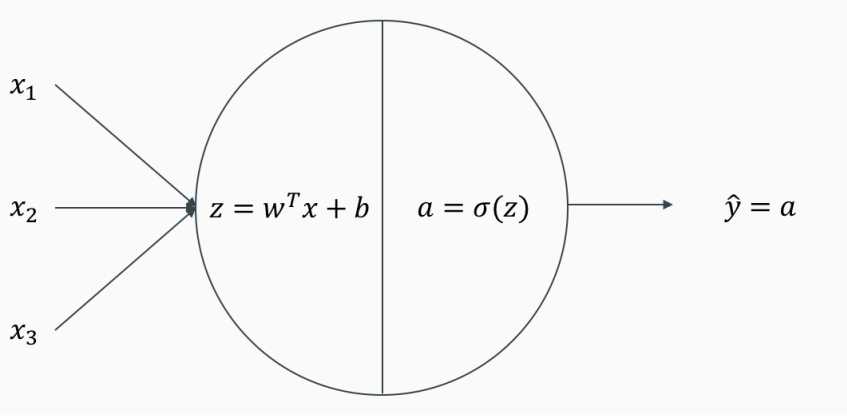
Foundation

Deep Learning

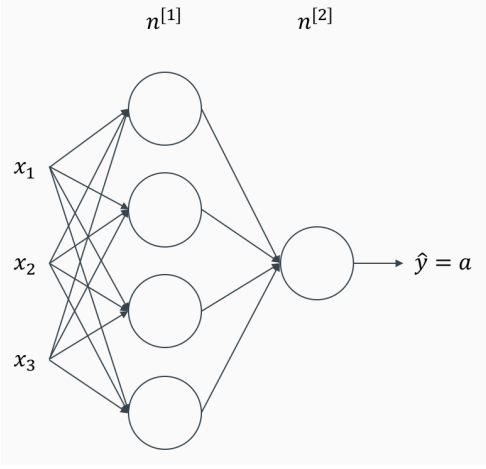
Logistic Regression



One Neuron



(Shallow) Neural Net



The Different Layers

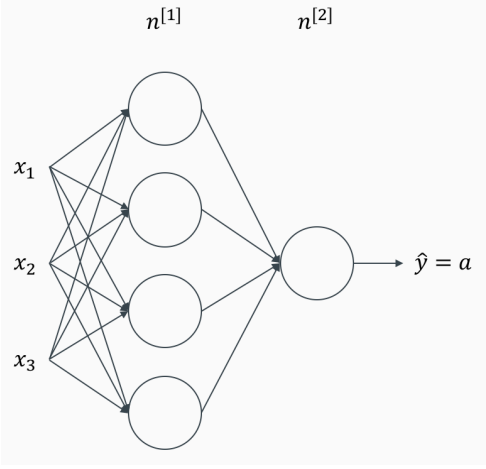
- Input layer
- Hidden layer
- Output layer

How to Train Neural Nets?

The Mantra

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

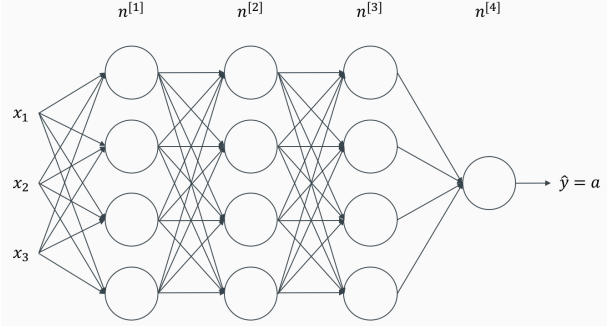
(Shallow) Neural Net



Four Steps to Repeat

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

Deep Neural Net



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

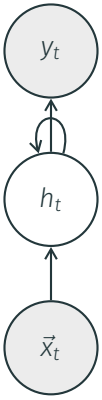


Figure 1: A simple RNN loops the output of h_t as an additional input to h_{t+1}.

Unrolling the RNN

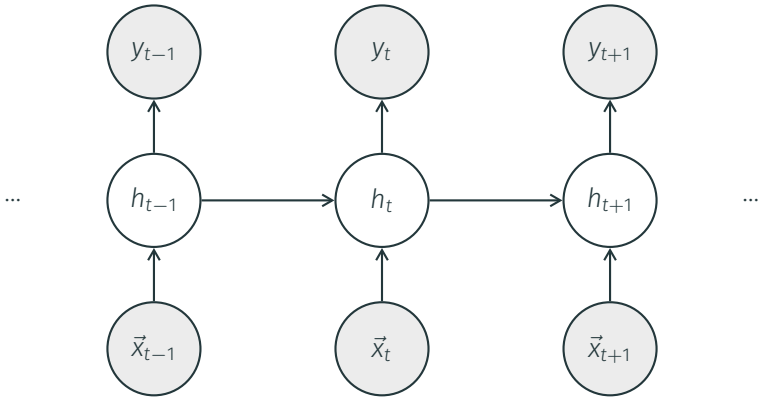


Figure 2: Unrolling the loop shows what happens in the RNN.

Visual Data Analysis with Convolutional Neural Nets (CNN)

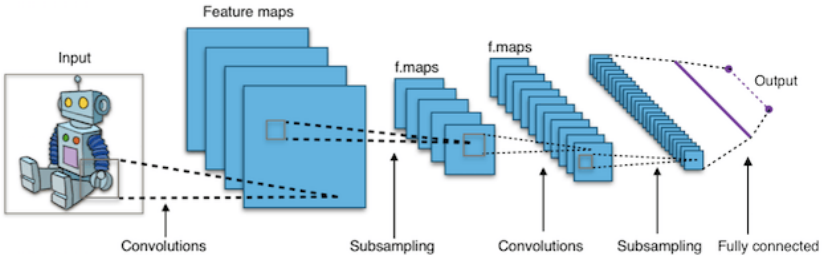


Figure 3: Typical CNN Architecture. Source: Keras Tutorial

In Sum: A Look at Different NN architectures

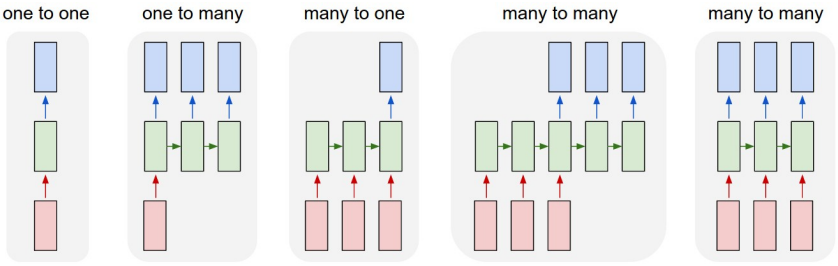


Figure 4: Different NN architectures. Source: Andrej Karpathy

Foundation

Multi Modal Learning

Multi Modal Learning & Social Sciences

Development of Data Sources

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

Multi Modal Learning & Social Sciences

Development of Data Sources

- Structured data $\rightarrow$ 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

Modalities

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

Multimodality

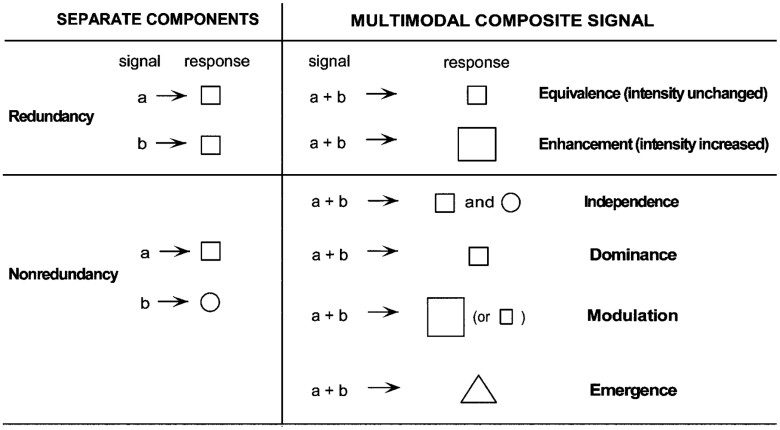
Multi Modal Learning

- Involves multiple modalities
- Data is heterogenous and interconnected

Spectrum: Homogenous and Heterogenous Modalities

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

Cross Modal Interactions



Source: Partan and Marler (2005)

Examples for Multi Modal Learning 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 Learning 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, Mondrak 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 Learning 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, Mondrak 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.

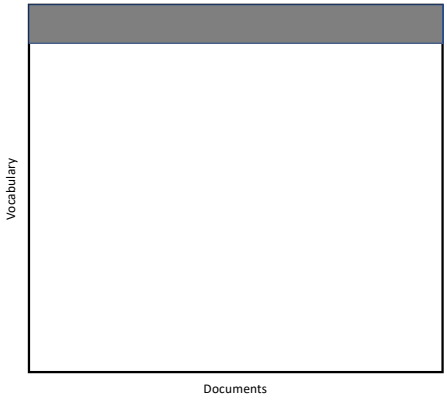
Basic Multimodal Learning

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, respondes 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, respondes well to Australia’s sunshine.
- I dined off bread and chees and this excellent bardiwac.

What is the Meaning of ‘Bardiwac’?

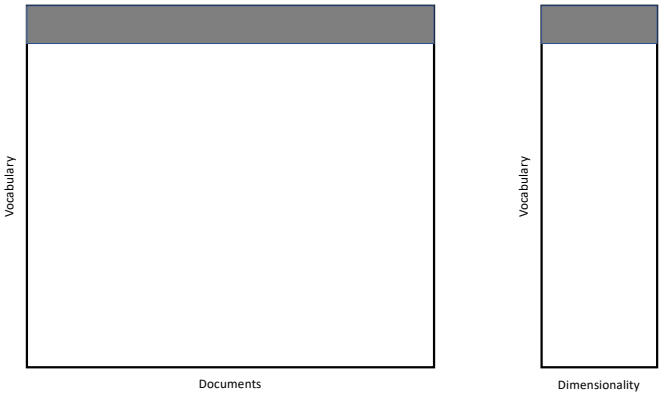
Distribution hypothesis: meaning of a word can be approximated by surrounding words.

- 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, respondes well to Australia’s sunshine.
- I dined off bread and chees and this excellent bardiwac.
- The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.

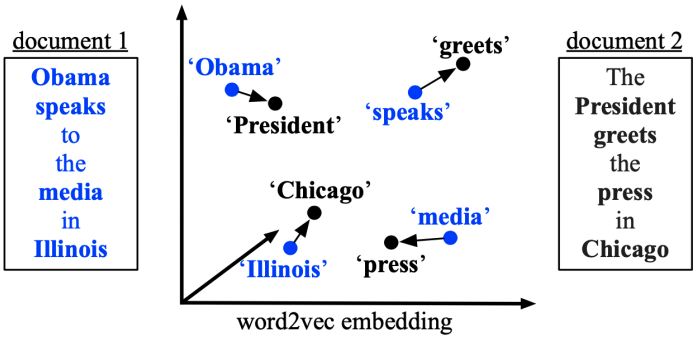
Word Embeddings

Evolution of Models

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

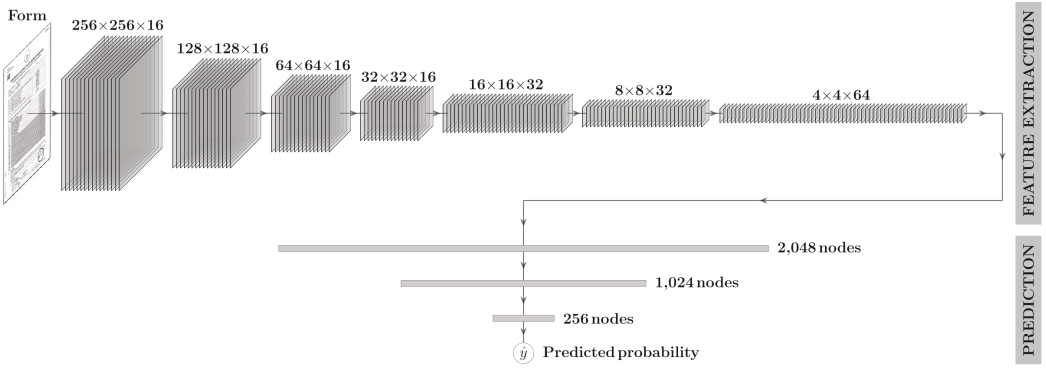


Doing Things with Wordembeddings

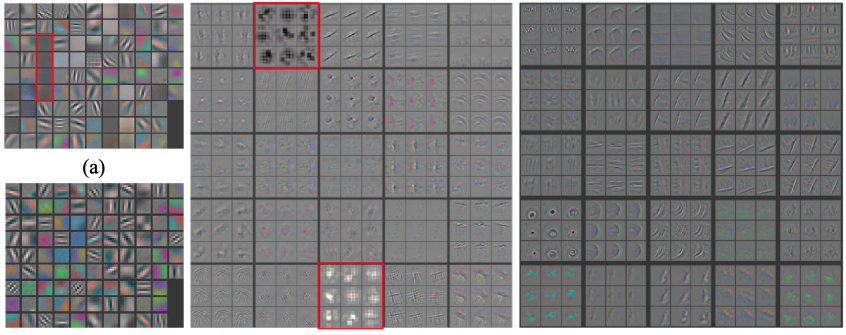


Source: Kusner et al. (2015)

Visual Embeddings

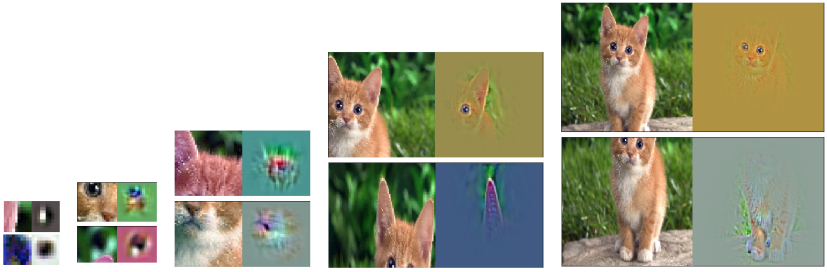


Filters



Source: Qin et al. (2018)

Features



Source: Qin et al. (2018)

Basic Multimodal Learning

Fusion

Fusion

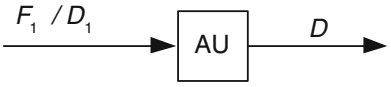
Multimodal Fusion

Process of joining information from modalities for prediction

Examples

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

Building Blocks: Analysis Unit

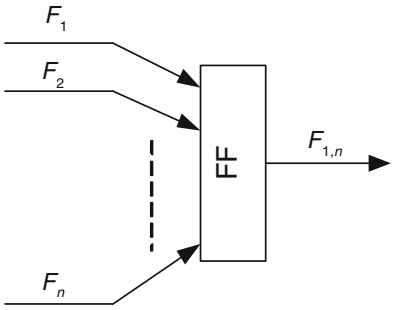


with

- F: Features
- D: Decisions

Source: Atrey et al. (2010)

Building Blocks: Feature Fusion Unit

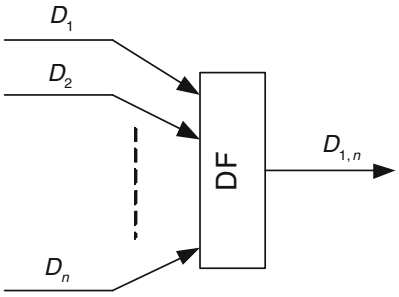


with

- F: Features

Source: Atrey et al. (2010)

Building Blocks: Decision Fusion Unit

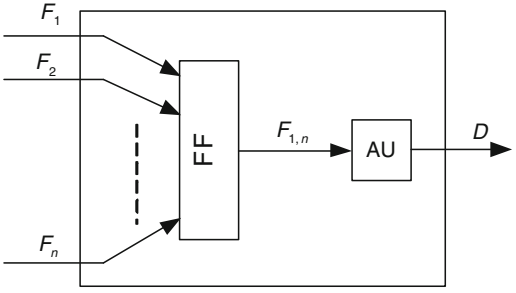


with

- D: Decisions

Source: Atrey et al. (2010)

Early Fusion/Feature Level MML

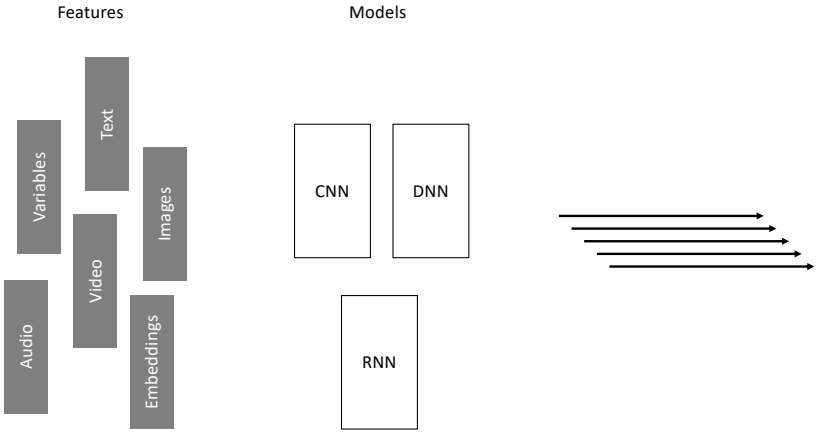


Source: Atrey et al. (2010)

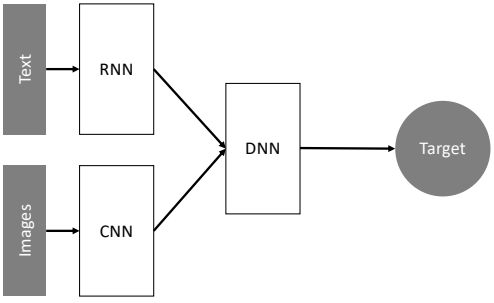
with

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

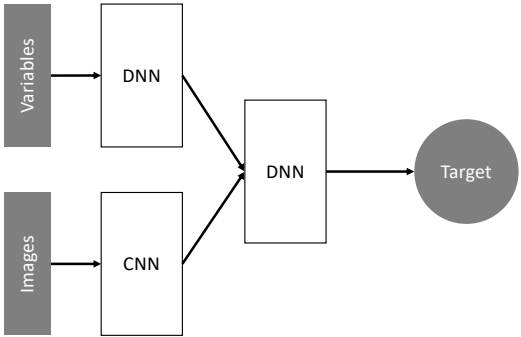
Quick Recap: What Do We Have to Play With?



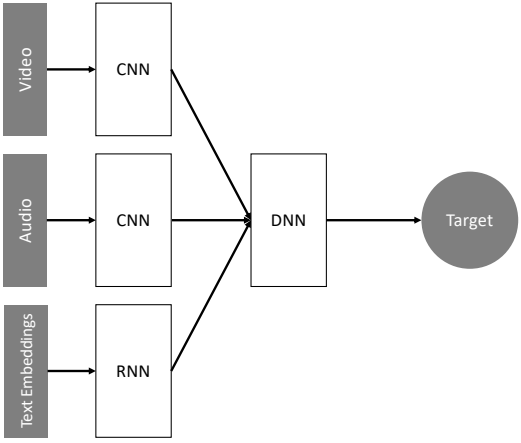
Early Fusion with Neural Networks



Early Fusion with Neural Networks



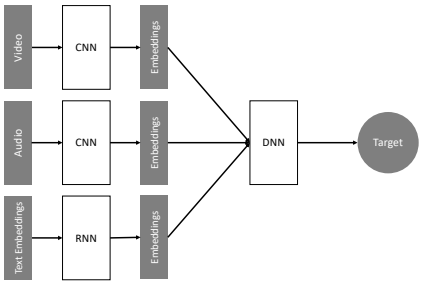
Early Fusion with Neural Networks



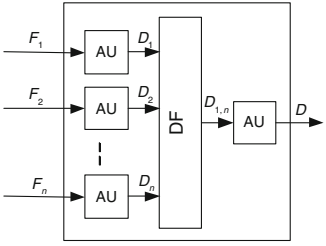
Early Fusion with Neural Networks

How does the data look like?

- Raw data?
- Features?
- Embeddings?



Late Fusion/Decision Level MML

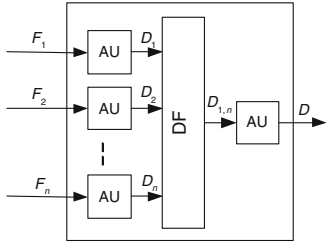


with

- F: Features
- D: Decisions

Source: Atrey et al. (2010)

Late Fusion/Decision Level MML



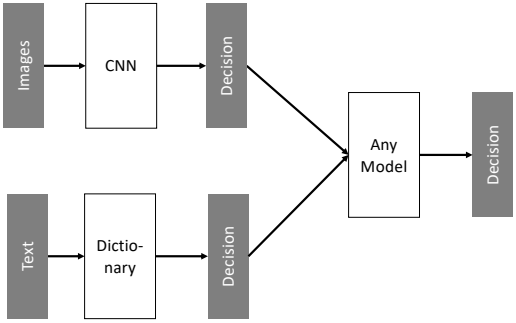
with

- F: Features
- D: Decisions

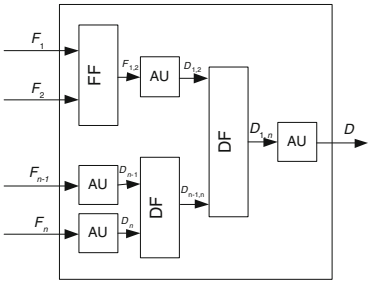
Source: Atrey et al. (2010)

- 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 syncronise
- No low level interactions between modalities
- Is this ensemble learning?

Late Fusion with Neural Networks



Hybrid Fusion/Hybrid MML

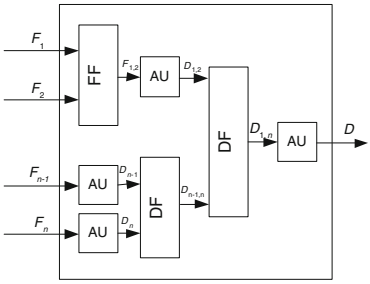


with

- F: Features
- D: Decisions

Source: Atrey et al. (2010)

Hybrid Fusion/Hybrid MML



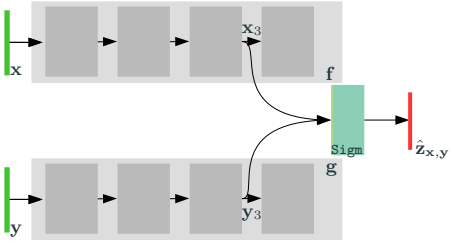
with

- F: Features
- D: Decisions

Source: Atrey et al. (2010)

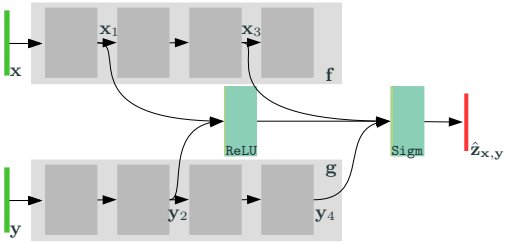
- Can have advantages of early and late fusion, e.g. interactions at lower and higher levels
- But may compound disadvantages, e.g. re training and complexity

Hybrid Fusion with Neural Networks



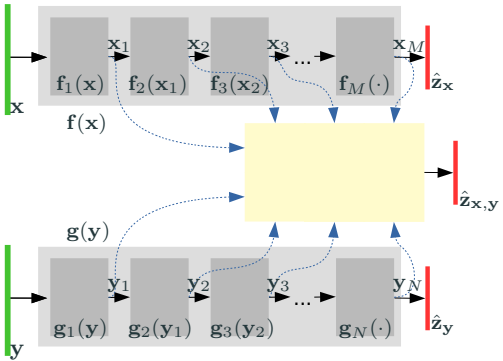
Source: Pérez-Rúa et al. (2019)

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)

Code

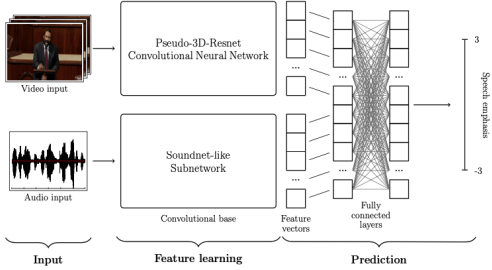
How to Build a Multi Modal Model with Keras

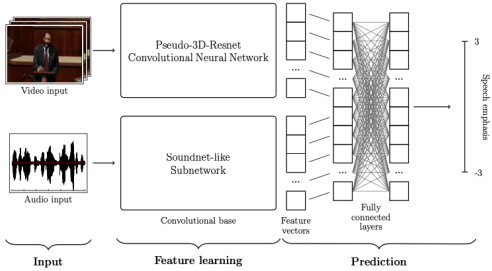
- Sequential and Functional API
- Model Fusion
- Google Collab: `https://colab.research.google.com/drive/1ZKqndxBsSZ2Yuu-kgafJnAgvnYDfHDjm?usp=sharing`

Basic Multimodal Learning

Examples from Political Science Revisited

Rittman, Ringwald and Nyhuis: *Public Opinion and Emphatic Legislative Speech*

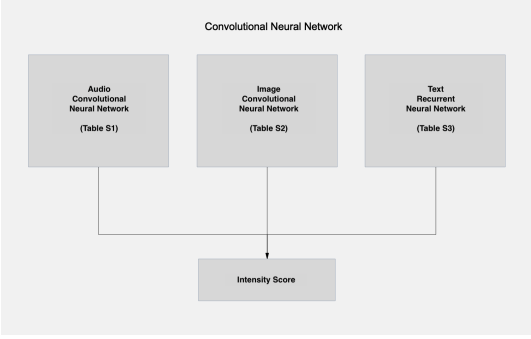


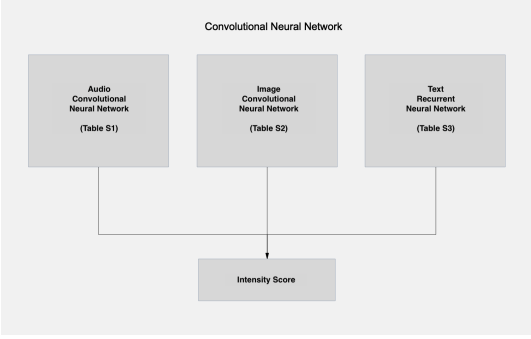


Modelling

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

Dietrich, Mondak and Williams: *Assessing Affective Polarization*

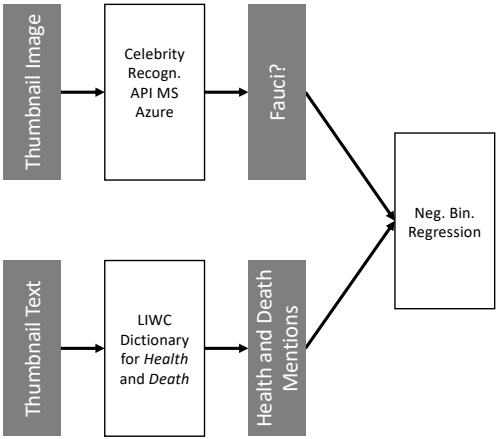




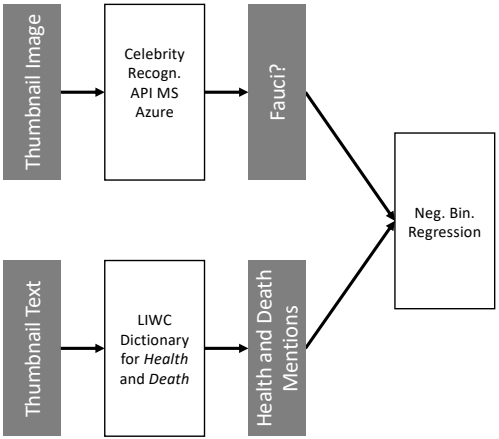
Modelling

- 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*



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

Code

Predicting House Prizes

- Predict the prize of a house on the basis of its images and characteristics
- Regression
- Google Collab: `https://colab.research.google.com/drive/1HVb5yBkN1TSiCo_1Nf4RW6RQ6aZgJO8w?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

Learning and Optimization

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

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

A Closer Look at Fusion

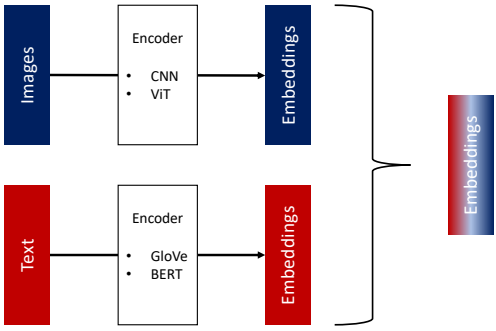
Definition

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

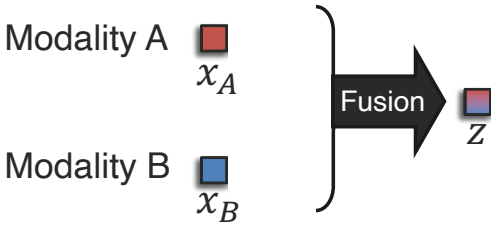
Fusing Modalities is a Continuum

- 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

$$z = w_1 x_1 + w_2 x_2 + \epsilon$$

Multiplicative Interaction

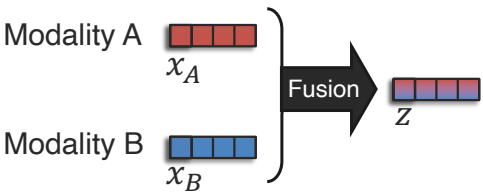
$$z = w_3 (x_1 x_2) + \epsilon$$

Additive and Multiplicative Interaction

$$z = w_1 x_1 + w_2 x_2 + w_3 (x_1 x_2) + \epsilon$$

Source: Morency et al. (2022)

Unimodal Multivariate Case



Additive Interaction

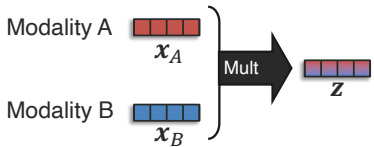
$$z = w_1 x_1 + w_2 x_2 = W \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Source: Morency et al. (2022)

Unimodal Multivariate Case

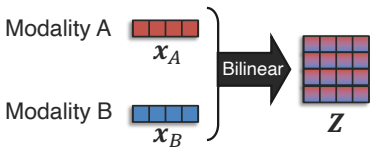
Multiplicative Fusion

$$z = w(x_1 \times w_2)$$



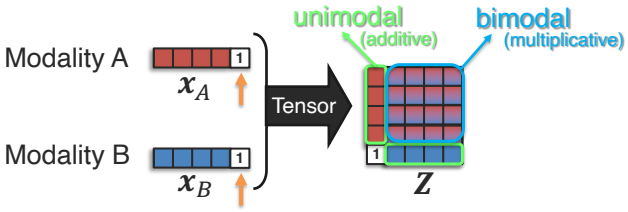
Billinear Fusion

$$Z = w(x_1^T \times w_2)$$



Source: Morency et al. (2022)

Unimodal Multivariate Case



Tensor Fusion

$$Z = w([x_1 1]^T \times [x_2 1])$$

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

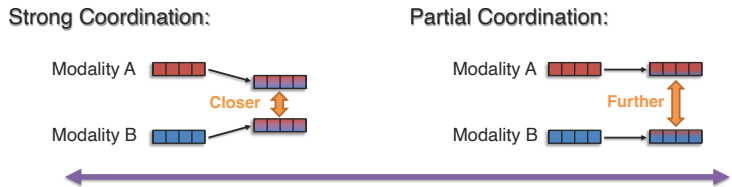
Extensions

Representation Coordination

Representation Coordination

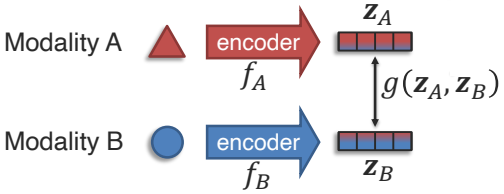
Definition

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

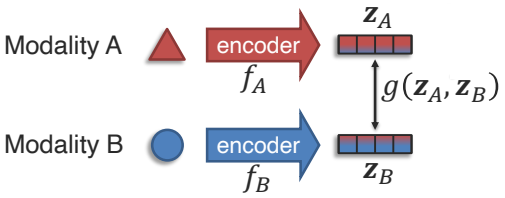


Source: Morency et al. (2022)

Examples for Coordination Functions $g()$:



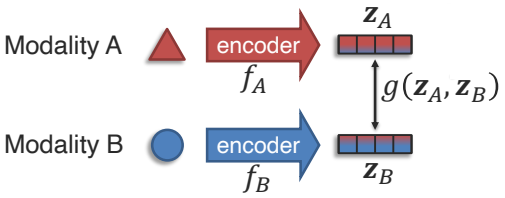
Examples for Coordination Functions $g()$:



Cosine Similarity

$$g(z_A, z_B) = \frac{k(z_A, z_B)}{\|z_A\| \|z_B\|}$$

Examples for Coordination Functions $g()$:



Source: Morency et al. (2022)

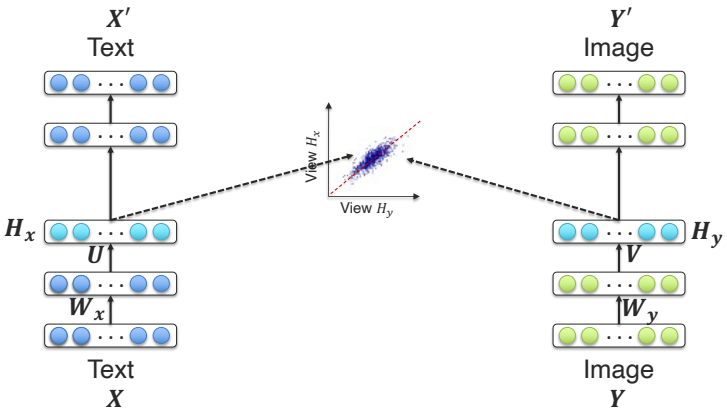
Cosine Similarity

$$g(z_A, z_B) = \frac{k(z_A, z_B)}{\|z_A\| \|z_B\|}$$

Canonical Correlation Analysis (CCA)

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

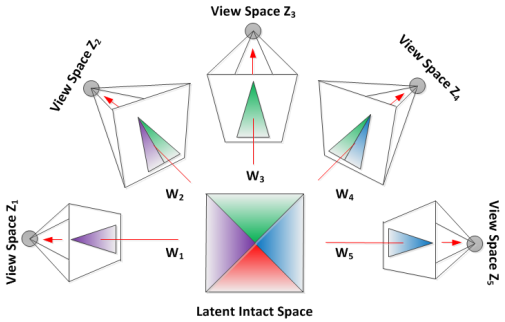
Deep Canonically Correlated Autoencoders (DCCAE)



- Cost function?
- How to train?

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

Intuition: Why is a Coordinated Embedding Space Useful?

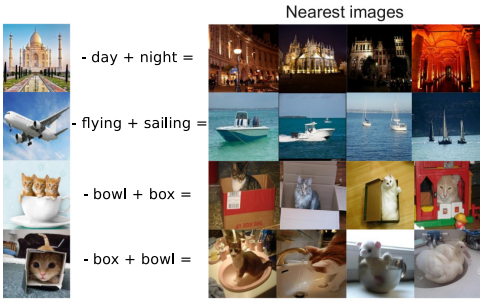


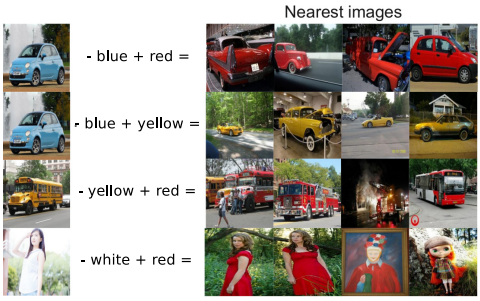
Source: Xu and Tao (2015)

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

Intuition: Why is a Coordinated Embedding Space Useful?





Source: Kiros et al. (2014)

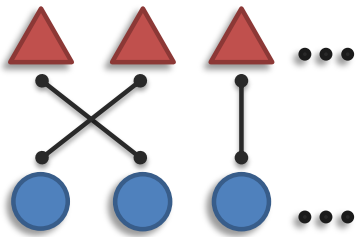
Extensions

Alignment

Multimodal Alignment

Definition

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



Source: Morrency et al. (2022)

Connection Types

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

Multimodal Alignment

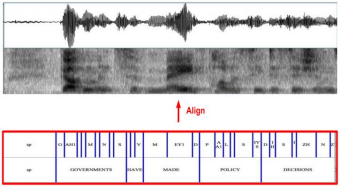
Two Ways of Alignment

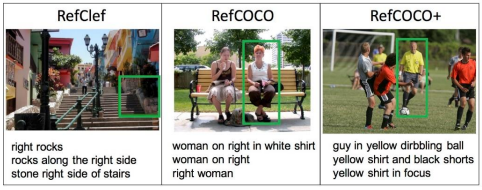
- Explicit aligment:
  - Alignment is the task itself.
  - $\rightarrow$ Alignment is the loss function.
- Implicit alignment:
  - Alignment is an intermediate step that is built into the model
  - Goal of the task is something else.
  - If the model aligns well, the model performance will be higher.

Explicit Multimodal Alignment

Examples

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





Source: Morency (2020)

Dynamic Time Warping: Unimodal

Two unaligned unimodal temporal signals

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

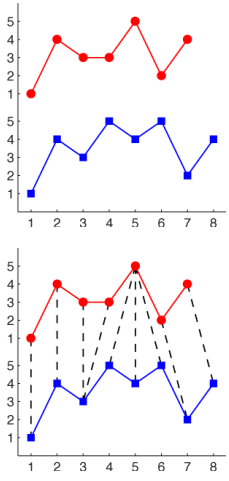
$$Y = [y_1, y_2, \ldots, y_{n_y}]$$

Find set of indices to minimize alignment difference

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

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

Source: Morency (2020)

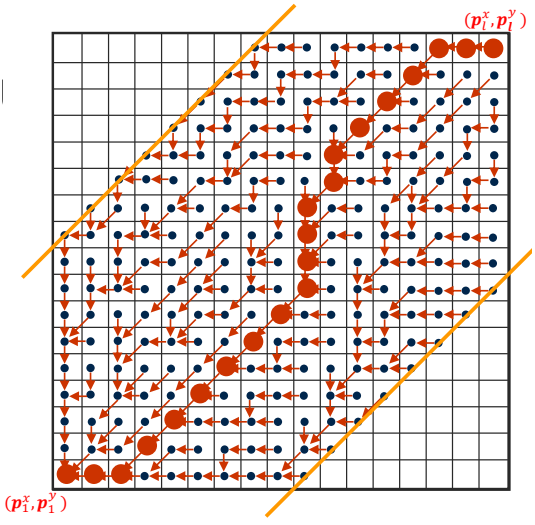


Dynamic Time Warping

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

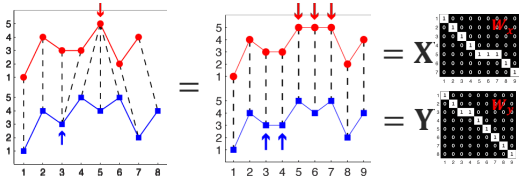
Source: Morency (2020)



Dynamic Time Warping: Alternative Formulation

Objective So Far

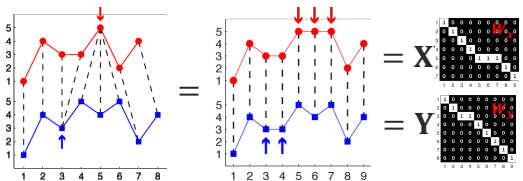
$$L(p^x, p^y) = \sum_{t=1}^{l} \left\| x_{p_t^x} - y_{p_t^y} \right\|_2^2$$



Dynamic Time Warping: Alternative Formulation

Objective So Far

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



Source: Morency (2020)

Alternative Objective

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

with

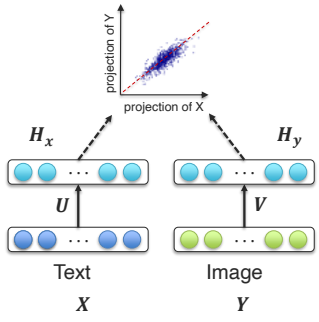
- X, Y original signals
- W_x, W_y alignment matrices to find

But: Computationally complex, sensitive to outliers, unimodal

Dynamic Time Warping + Canonical Correlation Analysis = Canonical Time Warping

Canonical Correlation Analysis

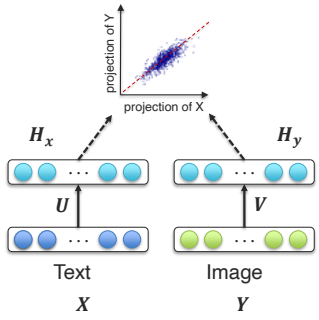
$$L(U, V) = \left\| U^T X - V^T Y \right\|_F^2$$



Dynamic Time Warping + Canonical Correlation Analysis = Canonical Time Warping

Canonical Correlation Analysis

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



Canonical Time Warping

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

- 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

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

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

Deep Canonical Time Warping

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

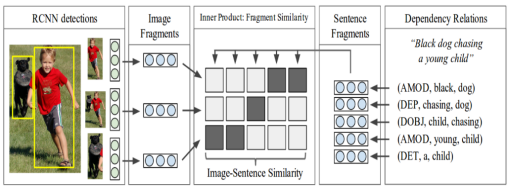
- Addresses non-linearities in alignment

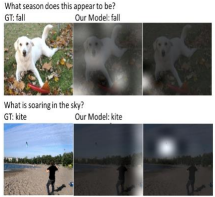
Source: Zhou and De la Torre (2009), Trigeorgis et al. (2016)

Implicit Alignment

Examples

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



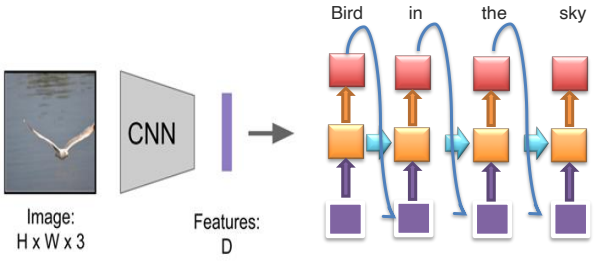


Source: Morency (2020)

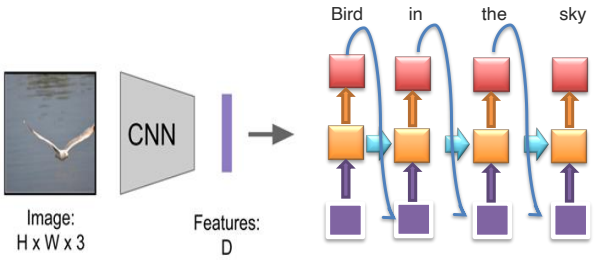
Visual Captioning with Soft Attention

Source: Xu et al. (2015)

Recap: Initialise RNN for Generating Captions



Recap: Initialise RNN for Generating Captions

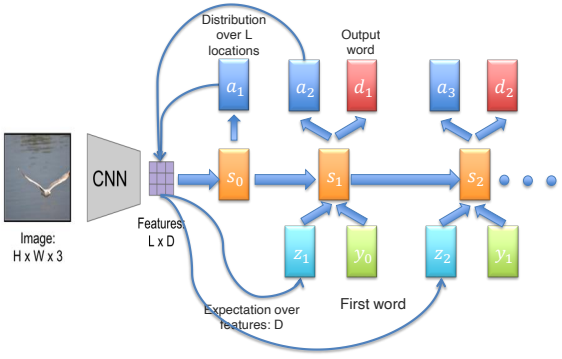


Why is it not a good idea to use the final layer of the CNN?

- Will the information of the image reach the end of the sentence?

Source: Morency (2020)

Soft Attention to Look at More Fine Grained Features



Source: Morency (2020)

Attention

- Output from response maps to the convolution (e.g. 4x4x64)
- Where to focus on the image
- For each location in the image (1x64)

Im Sum

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

Code

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`

Sources

Readings

- Atrey, Pradeep K., M. Anwar Hossain, Abdulmotaleb El Saddik and Mohan S. Kankanhalli. “Multimodal Fusion for Multimedia Analysis: A Survey.” *Multimedia Systems* (2010) 16: 345–379.
- Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. “Multimodal Machine Learning: A Survey and Taxonomy.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (2018): 423-443.”
- Bengio, Yoshua, Aaron Courville, and Pascal Vincent. “Representation learning: A Review and New Perspectives.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013): 1798-1828.
- Goodfellow, Ian and Yoshua Bengio and Aaron Courville. 2016. *Deep Learning*. MIT Press.

Sources

Internet Resources

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

Working with Deep Learning? Get in touch!

Dr Christian Arnold
Cardiff University
@chrisguarnold