Machine Learning for NLP

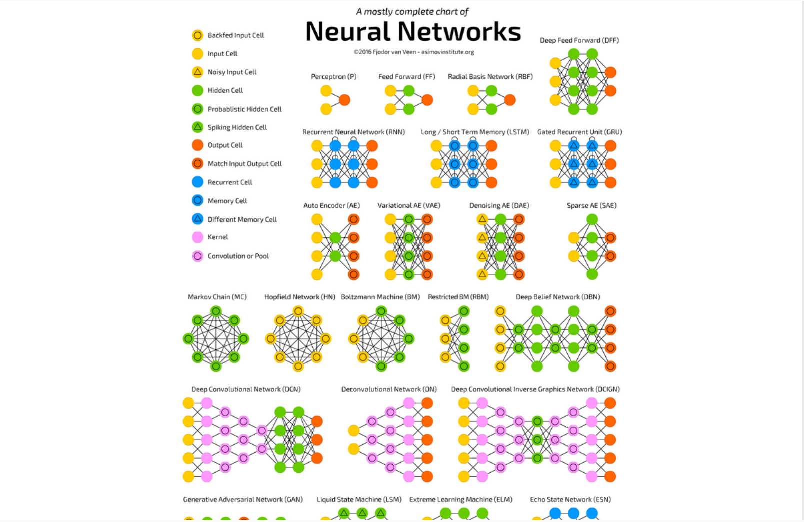
The Neural Network Zoo

Aurélie Herbelot

2019

Centre for Mind/Brain Sciences
University of Trento

The Neural Net Zoo



http://www.asimovinstitute.org/neural-network-zoo/

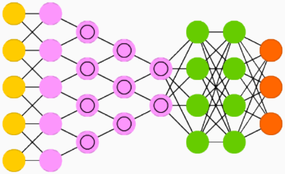
How to keep track of new architectures?

- The ACL anthology: 48,000 papers, hosted at https://aclweb.org/anthology/.
- arXiv on Language and Computation: https://arxiv.org/list/cs.CL/recent.
- Twitter...

Today:
a wild race through a few architectures

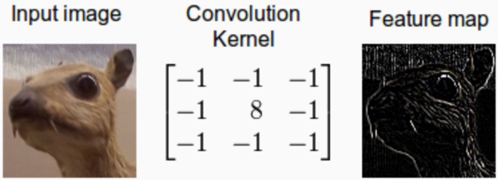
CNNs

- **Convolutional Neural Networks:** NNs in which the neuronal connectivity is inspired by the organization of the animal visual cortex.
- Primarily for vision but now also used for linguistic problems.
- The last layer of the network (usually of fairly small dimensionality) can be taken out to form a reduced representation of the image.



Convolutional deep learning

- Convolution is an operation that tells us how to *mix* two pieces of information.
- In vision, it usually involves passing a filter (kernel) over an image to identify certain features.



CNNs: what for?

- Identifying latent patterns in a sentence: syntax?
- CNNs can be used to induce a graph similar to a syntactic tree.

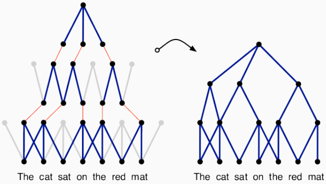
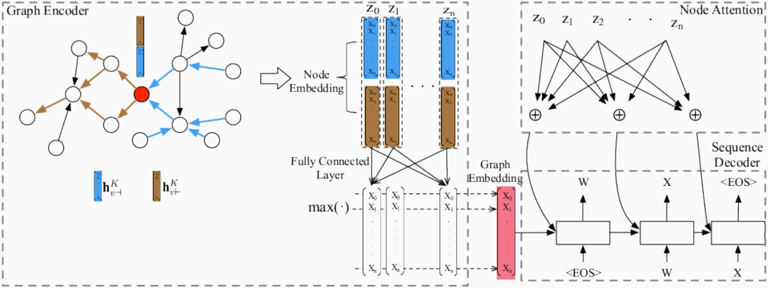


Figure 1: Subgraph of a feature graph induced over an input sentence in a Dynamic Convolutional Neural Network. The full induced graph has multiple subgraphs of this kind with a distinct set of edges; subgraphs may merge at different layers. The left diagram emphasises the pooled nodes. The width of the convolutional filters is 3 and 2 respectively. With dynamic pooling, a filter with small width at the higher layers can relate phrases far apart in the input sentence.

Kalchbrenner et al, 2014:
https://arxiv.org/pdf/1404.2188.pdf

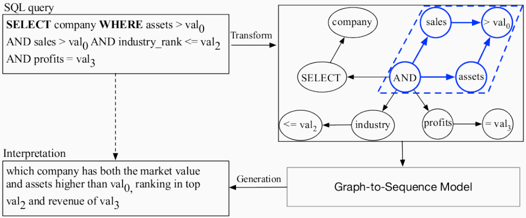
Graph2Seq architectures

- **Graph2Seq:** take a graph as input and convert it into a sequence.
- To embed a graph, we record the neighbours of a particular node and direction of connections.



Xu et al, 2018: https://arxiv.org/pdf/1804.00823

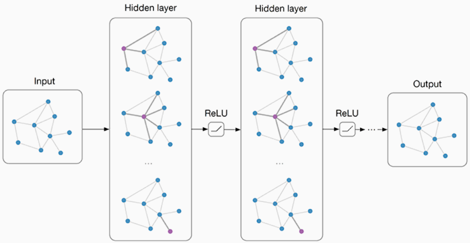
Graph2Seq: what for?



Language generation: the model has structured information from a database and needs to generate sentences describing operations over the structure.

GCNs

- **Graph Convolutional Networks:** CNNs that operate on graphs.
- Input, hidden layers and output all encapsulate graph structures.



GCNs: what for?

- Abusive language detection.
- Represent an online community as a graph and learn the language of each node (speaker). Flag abusive speakers.

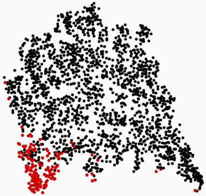
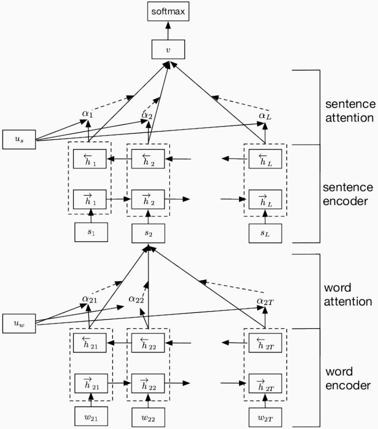


Figure 2: Visualization of the author profiles extracted from our GCN. Red dots represent the authors who are deemed abusive (racist or sexist) by the GCN.

Mishra et al, 2019: https://arxiv.org/pdf/1904.04073

Hierarchical Neural Networks

- **Hierarchical Neural Networks:** we have seen networks that take a graph as input. HNNs are shaped as acyclic graphs.
- Each node in the graph is a network.



Yang et al, 2016:
https://www.aclweb.org/anthology/N16-1174

Hierarchical Networks: what for?

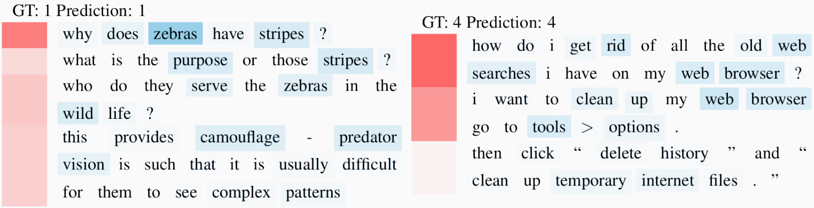
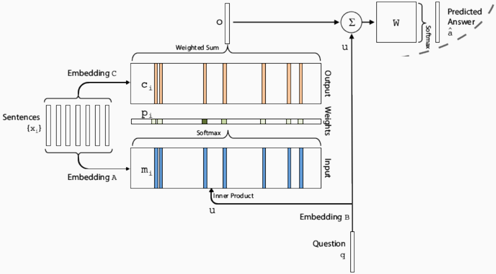


Figure 6: Documents from Yahoo Answers. Label 1 denotes Science and Mathematics and label 4 denotes Computers and Internet.

Document classification: the model attends to words in the document that it thinks are relevant to classify it into one or another class.

Memory Networks

- **Memory Networks:** NNs with a *store* of memories.
- When presented with new input, the MN computes the similarity of each memory to the input.
- The model performs attention over memory cells.



Sukhbaatar et al, 2015:
https://papers.nips.cc/paper/5846-end-to-end-memory-networks.pdf

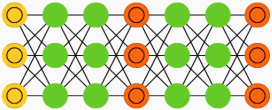
Memory Networks: what for?

Sam walks into the kitchen. Sam picks up an apple. Sam walks into the bedroom. Sam drops the apple. Q: Where is the apple? A. Bedroom	Brian is a lion. Julius is a lion. Julius is white. Bernhard is green. Q: What color is Brian? A. White	Mary journeyed to the den. Mary went back to the kitchen. John journeyed to the bedroom. Mary discarded the milk. Q: Where was the milk before the den? A. Hallway

Textual question answering: embed sentences as single memories. When presented with a question about the text, retrieve the relevant sentences.

GANs

- **Generative Adversarial Networks:** two networks working in collaboration.
- A *generative* network and a *discriminating* network.
- The discriminator works towards distinguishing real data from generated data while the generator learns to fool the discriminator.



GANs: what for?

- Generating images from text captions.
- Two-player game: the discriminator tries to tell generated from real images apart. The generator tries to produce more and more realistic images.

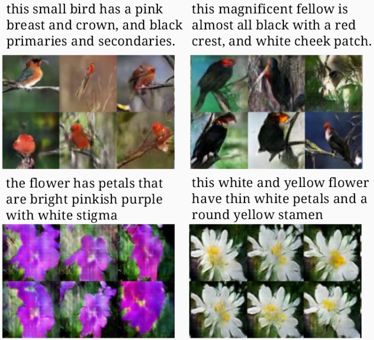
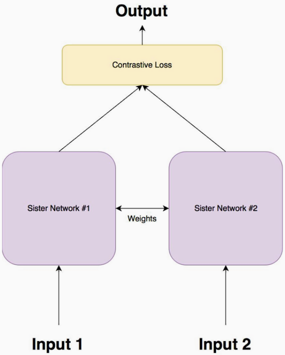


Figure 1. Examples of generated images from text descriptions. Left: captions are from zero-shot (held out) categories, unseen text. Right: captions are from the training set.

Reed et al, 2016:
http://jmlr.csail.mit.edu/proceedings/papers/v48/reed16.pdf

Siamese Networks

- **Siamese Networks:** learn to *differentiate* between two inputs.
- Use the same weights for two different input vectors and compute loss as a measure of *contrast* between the outputs.
- By getting a measure of contrast, we also get a measure of similarity.



https://hackernoon.com/one-shot-learning-with-siamese-networks-in-pytorch-8ddaab10340e

Siamese Networks: what for?

- Sentence similarity.
- By sharing the weights of two LSTMs, and combining their output via a contrastive function, we force them to concentrate on features that help assessing (dis)similarity in meaning.

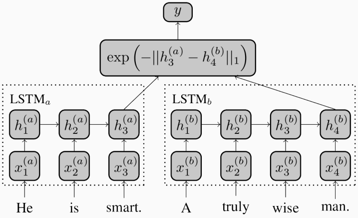
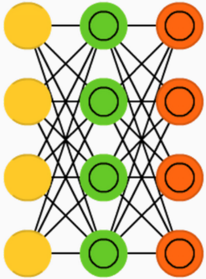


Figure 1: Our model uses an LSTM to read in word-vectors representing each input sentence and employs its final hidden state as a vector representation for each sentence. Subsequently, the similarity between these representations is used as a predictor of semantic similarity.

https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/viewPDFInterstitial/12195/12023

VAEs

- **AutoEncoders:** derived from FFNNs. They compress information into a (usually smaller) hidden layer (encoding) and reconstruct it from the hidden layer (decoding).
- **Variational Auto-Encoders:** an architecture that learns an approximated probability distribution of the input samples. Bayesian from the point of view of probabilistic inference and independence.



VAEs: what for?

- Model a smooth sentence space with syntactic and semantic transitions.
- Used for language modelling, sentence classification, etc.

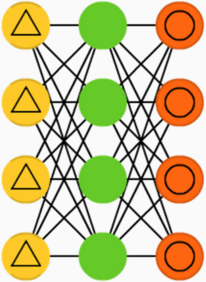
“ i want to talk to you . ”
“i want to be with you . ”
“i do n’t want to be with you . ”
i do n’t want to be with you .
she did n’t want to be with him .
he was silent for a long moment .
he was silent for a moment .
it was quiet for a moment .
it was dark and cold .
there was a pause .
it was my turn .
there is no one else in the world .
there is no one else in sight .
they were the only ones who mattered .
they were the only ones left .
he had to be with me .
she had to be with him .
i had to do this .
i wanted to kill him .
i started to cry .
i turned to him .
no .
he said .
“ no , ” he said .
“ no , ” i said .
“ i know , ” she said .
“ thank you , ” she said .
“ come with me , ” she said .
“ talk to me , ” she said .
“ do n’t worry about it , ” she said .

Table 8: Paths between pairs of random points in VAE space: Note that intermediate sentences are grammatical, and that topic and syntactic structure are usually locally consistent.

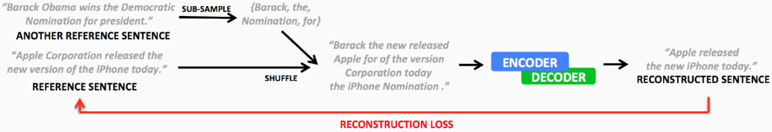
Bowman et al, 2016: https://www.aclweb.org/anthology/K16-1002

DAEs

- **Denoising AutoEncoders:** classic autoencoders, but the input is noisy.
- The goal is to force the network to look for the 'real' features of the data, regardless of noise.
- E.g. we might want to do picture labeling with images that are more or less blurry. The system has to *abstract away* from details.



DAEs: what for?

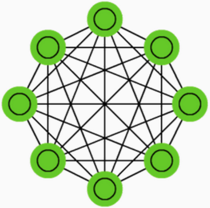


Fevry and Fang, 2018: https://arxiv.org/pdf/1809.02669

Summarisation: since the AE has learnt to abstract away from detail in the course of denoising, it becomes good at summarising.

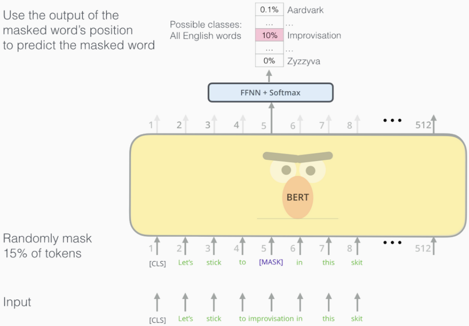
Markov chains

- **Markov chains:** given a node, what are the odds of going to any of the neighbouring nodes?
- No memory (see Markov assumption from language modeling): every state depends solely on the previous state.
- Not necessarily fully connected.
- Not quite neural networks, but they form the theoretical basis for other architectures.



Markov chains: what for?

- We will talk more about Markov chains in the context of Reinforcement Learning!
- For now, let's note that BERT is a little Markov-like...
  Wang and Cho, 2019: https://arxiv.org/pdf/1902.04094



https://jalammar.github.io/illustrated-bert/

What you need to find out about your network

1. Architecture: make sure you can draw it, and describe each component!
2. Shape of input and output layer: what kind of data is expected by the system?
3. Objective function.
4. Training regime.
5. Evaluation measure(s).
6. What is your network used for?