Advanced Machine Learning

András Kornai

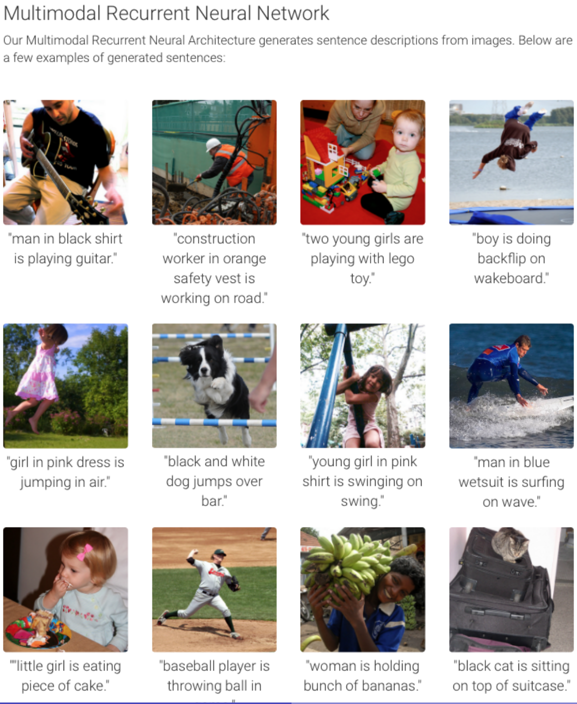
AML 2024/10/09

NLP (classic ML for lg data)

- Strict separation (typically 80-10-10) of *train*, *dev* and *test* data
- Train is used for building the model, dev for finetuning, test typically hidden from the model builder
- A model optimizes some figure of merit (e.g. word error rate in speech recognition)
- Strong culture of shared tasks (each team working on the same data)
- Generally requires large datasets (gigaword is now typical)
- Supervised methods rule – unsupervised learning still in its infancy
- `https://aclweb.org/aclwiki/POS_Tagging_(StateOfTheArt)`

NLP from a ML perspective

- There are more specific NLP courses such as https://python-nlp.github.io or https://github.com/elte-nlp/elte-nlp-course
- If you assume 128k words, at worst you'd need 17 bits to encode a word (in practive 12-15 bits suffice since word frequencies are not uniform. This assumption is compatible with the idea that there are infinitely many words.)
- Word vectors give you 300 dimensional real encodings (higher dimensions and higher precision are often used these days) for 9.6 kilobits/word
- But the word vectors are nice: similar words will have similar vectors (we measure this by cosine similarity more than Euclidean distance, because vector length are proportional to log word frequency)
- What does it mean for words to be similar? That they appear in the same or similar contexts (similar sentences, docs, . . .)



Multimodal Recurrent Neural Network

Our Multimodal Recurrent Neural Architecture generates sentence descriptions from images. Below are a few examples of generated sentences:

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"boy is doing backflip on wakeboard."

"girl in pink dress is jumping in air."

"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

"man in blue wetsuit is surfing on wave."

""little girl is eating piece of cake."

"baseball player is throwing ball in

"woman is holding bunch of bananas."

"black cat is sitting on top of suitcase."

From Visual Question-Answering to Visual Reasoning

Together with the increased performance on the typical computer vision tasks, computer vision transitions into more holistic reasoning systems. One such study is visual question answering, where the visual system is exposed to questions about images. However, even the most challenging existing tasks can still be handled by systems with limited reasoning capabilities. For instance, the state-of-the-art on VQA, the most popular visual question answering dataset, relies heavily on pre-trained visual features. Yet, other elements that are associated with human intelligence like memory, step-by-step planning, compositional thinking, or symbolic manipulation, are often ignored. We want to close the gap between that "fast thinking", which is often impulsive and in this context responsible for a quick interpretation of the visual scene, and "slower thinking" that is more algorithmic. To achieve such goals, we need to think, build, think again, and build suitable benchmarks, architectures and algorithms. Can you create the first system that connects vision with the reasoning in the next three or four years?

What color are her eyes?
What is the mustache made of?

Antol et al. VQA: Visual Question Answering. ICCV'15 & CVPR'18

Research directions

In the following research programme, you will

- define what visual reasoning is by building various datasets and tasks
- build architectures that deal with basic reasoning tasks such as analogies, counting, intuitive physics, memory, all grounded in perception
- understand the limitations of the current systems
- draw inspirations from biological systems
- move beyond the standard paradigm of learning from pixels towards reasoning about pixels

Word vector precursors

- Discrete (partial) decomposition of meanings into finite bit vectors is old hat, for example *brother* = '+sibling +male' *sister* = '+sibling –male'
- Continuous begins with Osgood et al. (1975) who asked for judgements on a scale of -3 to +3 and performed PCA on the results
- Next big thing was Landauer, Dumais, etc. who took term-document cooccurrence data
- Create 'term-document matrix' T where T_{ij} counts the number of times term i appears in document j
- Landauer, Dumais etc. applied SVD, reduced T to a few hundred principal components, called it "Latent semantic indexing" and patented it (thereby slowing down developments by 15 years or so – Microsoft didn't get rich on the patents)

Success has many fathers

Embedding (static)

Given a dictionary D, a static embeddig is a function $\vec{v}$ that assigns for each word $w \in D$ a vector $\vec{v}(w) \in \mathbb{R}^n$

- First computational treatment by Schütze, 1993 (but goes back to Firth, 1957)
- First implementation that really worked (Bengio et al., 2003)
- NLP “almost from scratch” POS, CHUNK, NER, role labeling (Collobert et al., 2011)
- Has linear structure (king–queen=man–woman) (Mikolov, Yih, and Zweig, 2013)
- Why? (Pennington, Socher, and Manning, 2014; Arora et al., 2015; Gittens, Achlioptas, and Mahoney, 2017)

Word embeddings

- Let us define "context as "within a window of $\pm n$ words (typically, $n = 5$). We define PMI(x,y)=log $\frac{p(x,y)}{p(x)p(y)}$
- Actually we tend to ignore negative evidence, and use PPMI = max(0, PMI)
- The big thing is that *supervised* data is obtained cheaply (teraword scale)
- T is now term-term cooccurrence (PPMI) matrix, and we again do SVD for dimension reduction. This way we assign a relatively short vector $\vec{word} \in \mathbb{R}^d$ to each word. This assignment is called the (static) embedding
- Dynamic embeddings (ELMO, BERT) don't have this kind of clean math yet
- In fact, the static was not fully understood until Levy and Goldberg 2014a It worked first, made sense later

Groups

- NLP: *Acevedo* Aktan Karoiu Tatrishvili
- TrafficSigns: Bodai Oroszki **Szőke** Szűcs
- Fingerprint: Bárdos-Deák Boros Czakó *Kránitz*
- Flower: *Békési* Sooomro Szecskás Wiederschiz
- MRI: *Hermán* Kovács Nguyen Varga
- Simulation: *Máth* Nemes
- SignLg: Barta Nagy Oroszki Szimonenk
- Speech: *Gedeon* Kövér
- Punctuation: Gómez, *Gallego*
- WP: *Juhász*

- Arora, Sanjeev et al. (2015). “Random Walks on Context Spaces: Towards an Explanation of the Mysteries of Semantic Word Embeddings”. In: *arXiv:1502.03520v1* 4, pp. 385–399. DOI: 10.1162/tacl_a_00106.
- Bengio, Yoshua et al. (2003). “A Neural Probabilistic Language Model”. In: *Journal of Machine Learning Research* 3, pp. 1137–1155. DOI: 10.1162/tacl_a_00059. URL: http://www.jmlr.org/papers/v3/bengio03a.html.
- Collobert, Ronan et al. (2011). “Natural Language Processing (Almost) from Scratch”. In: *Journal of Machine Learning Research (JMLR)*.
- Firth, John R. (1957). “A synopsis of linguistic theory”. In: *Studies in linguistic analysis*. Blackwell, pp. 1–32.
- Gittens, Alex, Dimitris Achlioptas, and Michael W. Mahoney (2017). “Skip-Gram – Zipf + Uniform = Vector Additivity”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 69–76.

DOI: 10.18653/v1/P17-1007. URL:
http://aclweb.org/anthology/P17-1007.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (2013). "Linguistic Regularities in Continuous Space Word Representations". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*. Atlanta, Georgia: Association for Computational Linguistics, pp. 746–751.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL:
http://www.aclweb.org/anthology/D14-1162.

Schütze, Hinrich (1993). “Word Space”. In: *Advances in Neural Information Processing Systems 5*. Ed. by SJ Hanson, JD Cowan, and CL Giles. Morgan Kaufmann, pp. 895–902.