GAČR EXPRO NEUREM^3

Studying Representations

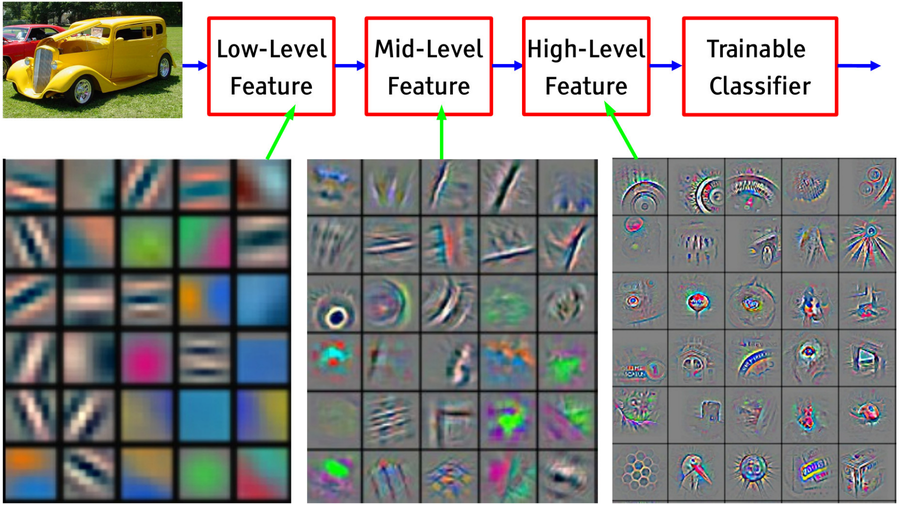
Ondřej Bojar

Sept 16, Pec pod Sněžkou

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Deep NNs for Image Classification

- It's deep if it has more than one stage of non-linear feature transformation



Low-Level Feature

Mid-Level Feature

High-Level Feature

Trainable Classifier

Caveat on Evaluation (1/2)

Consider word2vec “comprehensive” test set (Mikolov et al., 2013):

- 8.8k “semantic” and 10.6k “syntactic” questions,
- w2v “accuracy is quite good” (eyeballing)
  - The authors do mention that exact-match is “only about 60%”).

Kocmi and Bojar (2016) carefully examined the test set:

- “Semantic” questions cover only 3 question types:
  - country→city, country→currency, masculine family member→ feminine
  - Vylomova et al. (2016) test many other relations, e.g. walk-run, dog-puppy, bark-dog, cook-eat.
- “Syntactic” questions constructed by combinations:
  - starting from only 313 distinct word pairs,
  - (leading to only 35 different pairs per question on average),
  - And of the 313 pairs, 286 are formed regularly.

Caveat on Evaluation (2/2)

	Test Set by	
Accuracy on “Synt Qs”	Mikolov et al.	Kocmi et al.
word2vec as released	62.5%	43.5%
word2vec on our data	42.5%	9.7%
SubGram on our data	42.3%	22.4%

Caveat on Evaluation (2/2)

	Test Set by	
Accuracy on “Synt Qs”	Mikolov et al.	Kocmi et al.
word2vec as released	62.5%	43.5%
word2vec on our data	42.5%	9.7%
SubGram on our data	42.3%	22.4%
Nine rules	**71.9%**	**66.4%**

Caveat on Ultimate Evaluation

Kocmi and Bojar (2016):

- submitted to TSD on March 22, 2016.
- appeared in TSD in September 2016.

… cited by 4.

Bojanowski et al. (2017):

- submitted to arxiv on July 15, 2016.
- appeared in TACL 2017.

… cited by 1024.

Caveat on Ultimate Evaluation

Kocmi and Bojar (2016):

- submitted to TSD on March 22, 2016.
- appeared in TSD in September 2016.

… cited by 4.

- No code released, no fast code implemented at all.

Bojanowski et al. (2017):

- submitted to arxiv on July 15, 2016.
- appeared in TACL 2017.

… cited by 1024.

- This is the FastText paper.

ÚFAL People in NEUREM^3

- Ondřej Bojar
- Pavel Pecina
- Jindra Helcl (non-autoregressive MT, i.a.)
- Ivana Kvapilíková (unsupervised MT)
- Michal Auersperger (document representations)
- (Jindřich Libovický) (MT with images, i.a.)
- (Petra Galuščáková) (something with video?)

Expected Outcomes of NEUREM^3

- Insight into what the representations look like (for ASR and NMT).
- Tools for diagnosing:
  - Which tasks are learned implicitly with the main one.
  - Why is the network making some particular types of errors.
  - Which generalizations has the network learned and which not.
- Methods for:
  - semi-supervised and unsupervised learning.
  - pre-training, reuse of model parts, combining larger models, model interfacing,
  - successful multi-task training,
    all esp. in the areas of ASR and NMT.

Expected Outcomes of NEUREM^3

- Insight into what the representations look like (for ASR and NMT).
- Tools for diagnosing:
  - Which tasks are learned implicitly with the main one.
  - Why is the network making some particular types of errors.
  - Which generalizations has the network learned and which not.
- Methods for:
  - semi-supervised and unsupervised learning.
  - pre-training, reuse of model parts, combining larger models, model interfacing,
  - successful multi-task training,

    all esp. in the areas of ASR and NMT.
- Good papers, good papers, good papers...

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tom Kocmi and Ondřej Bojar. 2016. SubGram: Extending Skip-gram Word Representation with Substrings. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, number 9924 in Lecture Notes in Computer Science, pages 182–189, Cham / Heidelberg / New York / Dordrecht / London. Masaryk University, Springer International Publishing.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2016. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1671–1682, Berlin, Germany, August. Association for Computational Linguistics.