Building NLP classifiers

Vlado Boža
FMFI UK
Central Europe AI (CEAI)

Job salary prediction | Kaggle

#	Δrank	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	lazylearner *	3464.55935	18	Fri, 29 Mar 2013 20:05:33
2	—	**Vlado Boza ***	**3612.65985**	**56**	**Fri, 29 Mar 2013 14:49:15**
3	—	LR *	3862.80974	19	Tue, 02 Apr 2013 21:33:00
4	—	dejavu	4051.09000	48	Fri, 29 Mar 2013 20:45:58

Stress Engineer Glasgow	Stress Engineer Glasgow We re currently looking for talented engineers to join our growing Glasgow team at a variety of levels. The roles are ideally suited to high calibre engineering graduates with any level of appropriate experience, so that we can give you the opportunity to use your technical skills to provide high quality input to our aerospace projects, spanning both aerostructures and aeroengines. In return, you can expect good career opportunities and the chance for advancement and personal and professional development, support while you gain Chartership and some opportunities to possibly travel or work in other offices, in or outside of the UK. The Requirements You will need to have a good engineering degree that includes structural analysis (such as aeronautical, mechanical, automotive, civil) with some experience in a professional engineering environment relevant to (but not limited to) the aerospace sector. You will need to demonstrate experience in at least one or more of the following areas: Structural/stress analysis Composite stress analysis (any industry) Linear and nonlinear finite element analysis Fatigue and damage tolerance Structural dynamics Thermal analysis Aerostructures experience You will also be expected to demonstrate the following qualities: A strong desire to progress quickly to a position of leadership Professional approach Strong communication skills, written and verbal Commercial awareness Team working, being comfortable working in international teams and self managing PLEASE NOTE SECURITY CLEARANCE IS REQUIRED FOR THIS ROLE Stress Engineer Glasgow	30000

Not so good solution attempts

- Measuring job similarity
  - For kNN
  - As SVM kernel

- Very tempting (so much stuff to play with), but also not very good

- Also usually awfully slow during prediction, unless you use something like: https://github.com/searchivarius/NMSLIB

Reasonable baseline

- Linear regression on top of bag of bigrams
- Text: “C++ programmer in London”
- Dictionary: {“c++ programmer”: 3, “programmer in”: 6, “in London”: 1, …}
- Input for regression:
  - [0, 1, 0, 1, 0, 0, 1, ….]

Regression on top of bigrams

- Pros:
  - Stupidly easy to implement (even from scratch in C++) (5 lines in scikit-learn)
  - Fast during prediction

- Surprise
  - L2 regularization worked better than L1

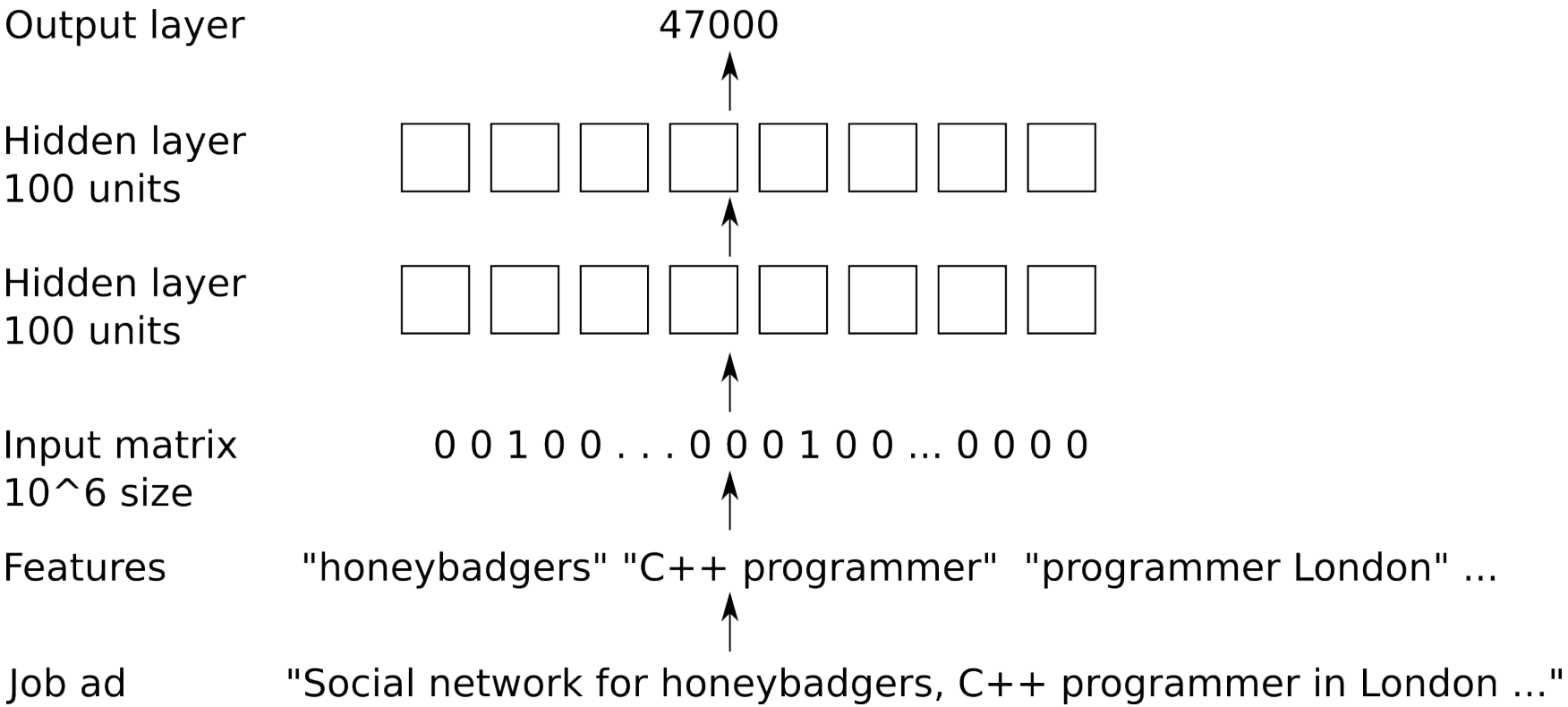
Pushing it further

- Neural nets are just extension of linear regression

- Input has high dimension but it is sparse
  - Need support for sparse matrices for efficient implementation of backprop
  - Easy to get in 2017, hard in 2013 (used custom c++ implementation during that time)
  - Now e.g. embedding layer from Keras does the trick

- FastText from facebook does very similar thing
  - https://github.com/facebookresearch/fastText

Neural architecture



More tricks

- Feature hashing
  - Saves memory for dictionary
  - Instead of {“c++ programmer” -> 4742, ...}
  - We use hash function from strings (features) to numbers
  - Collisions might happen, but ML algos don’t care

  - Scikit-learn: HashingVectorizer/FeatureHasher
  - Not used in kaggle competition, but later in prod

More tricks

- Dropout
  - Only used on input
  - Think about it as dataset augmentation

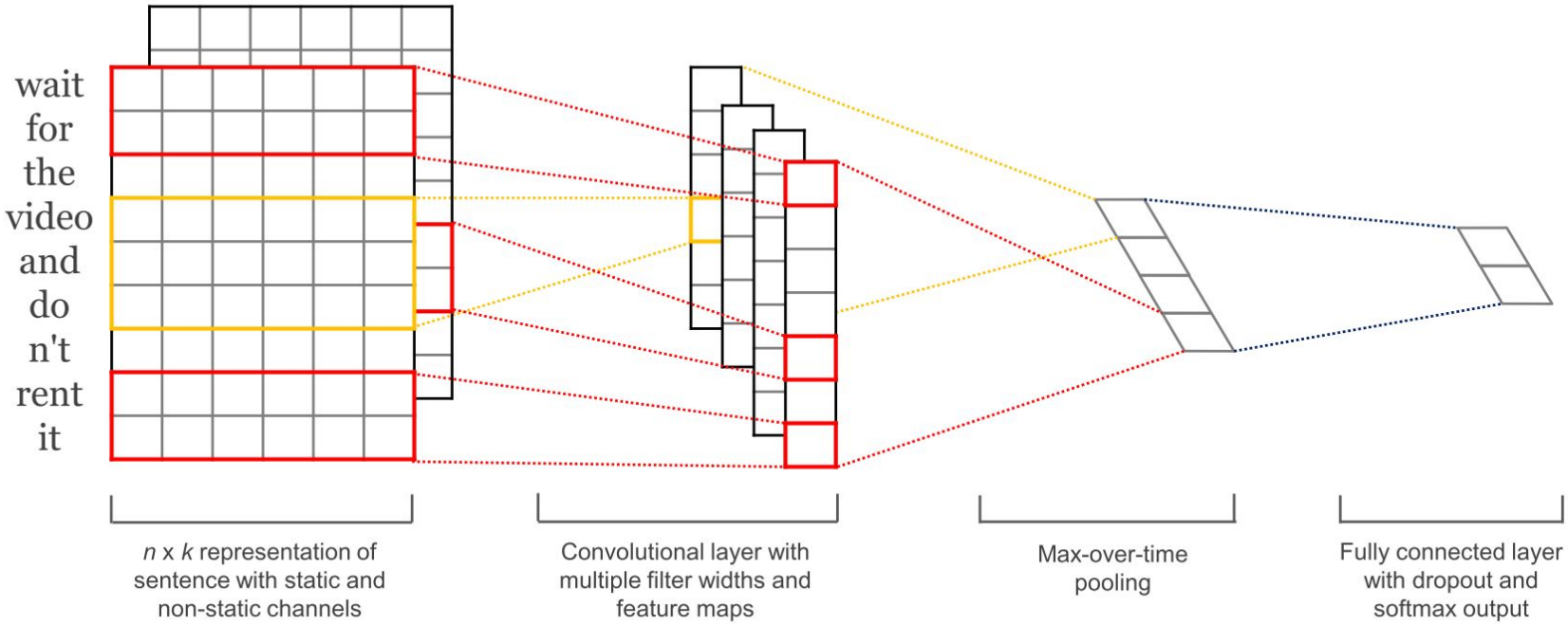
- Ensembling (averaging multiple predictions)
  - Stabilizes model output
  - Boosts performance
  - Standard way:
    - Train multiple neural nets with different
      - Initializations / order of samples / bootstrapped datasets
  - Poor man version: average output of neural nets, after each k iterations of training (after you get reasonable convergence)

Winner and 3rd place solution

- Winner
  - Big neural net on top of bag of words (15000 words)
  - 3 layers with 5000-1000-1000 hidden units

- 3rd place
  - Classification to several buckets of salary
  - Models distribution of salaries better

2017s solution outlook - convnets



Glossary

- DCVC - USA VC fund focused on AI

- CEAI - Central Europe AI Incubator (backed by DCVC)

- Raptor - one of CEAI startups

Deutsche Bank chiefs quit after rate-fixing and cover-up scandals

Wells Fargo CEO resigns from advisory role with the Federal Reserve

Raptor

KYC

MEDIUM

Customer data

Name:	Khalifa bin Zayed Al Nahyan	Occupation:	President of the United Arab Emirates, Emir of Abu Dhabi, commander of the Union Defence Force
Customer type:	CTI 3	Products/Services:	Foreign exchange, Commodities, Deposits, Real-estate investments
Date of Birth:	Wed, Feb 25, 1948	Purpose of acc.:	Transactional
Identification:	8343664	Anticipated activity:	Money services

Documentary

Passport

HIGH

Non documentary

Authorization call

MEDIUM

Financial statement

MEDIUM

Aggregated ID data from core banking, document storage, and KYC vendors

CDD

MEDIUM

BFWL: Partial address match(4.3%) and name match(42%)

CDD watchlists, etc

EDD

HIGH

Panama papers Detail

British Virgin Islands companies used to buy luxury real estate and other properties ...

The New York Times: In Aiding Dubai, Abu Dhabi Tightens Its Grip Detail

... Mohammed bin Rashid al-Maktoum, and his cousin Sheik Khalifa bin Zayed al-Nahyan, who is the president... by marriage. The daughter of Sheik Mohammed of Dubai recently married Sheik Mansour bin Zayed al-Nahyan, the half brother of Sheik Khalifa and a member of the federal cabinet. It became clear that what ...

BBC: UAE executes woman for killing American teacher Detail

... morning after approval was given by the UAE's president, Sheikh Khalifa bin Zayed Al Nahyan...Image copyright AP Image caption Alaa Badr al-Hashemi was filmed by CCTV cameras entering... an American teacher in December 2014. Alaa Badr Abdullah al-Hashemi, a 30-year-old Emirati ...

EDD content from vendors, online news, leak databases, etc

Raptor NLP problems

- Detecting adverse news linked to person (this talk)
- Identity matching
- Ranking results

Detecting adverse news

Simple IR approach (name and keyword in proximity) works pretty well

- Reasonably high recall
- Decent precision

False positives

- Judge **John Smith** sentenced James Doe for money laundering.
- Amy Smith is accused of murder of her brother **John Smith**.
- **John Smith** suffered cardiac arrest.

Make predictor for this.

Other problems

- Must work in crazy languages
  - Chinese
    - No upper case
    - No word boundaries
    - Simplified vs. traditional
  - Arabic
- Regulators

Task definition

- Entity centric sentiment
  - High coverage, one classifier takes it all
  - Vague, not easy to interpret

- Mining specific relations
  - E.g.: (person, legal action, crime)
  - Bad coverage
  - Clearly defined

Metrics specification

- Many metrics which you might optimize
  - AUC, Log likelihood, accuracy, precision, recall, specificity, F1, …
- One reasonable approach
  - Optimize one metric, while having constraints on others
  - Optimize precision, while recall > 99%
- Check Machine Learning Yearnings from Andrew Ng

(Lack of) Training data

- Distant supervision
  - Use heuristics to build training dataset
- Bootstrapping
  - Start with small dataset
  - Use classifier to generate bigger dataset (from areas where classifier is confident)
  - Repeat
- Semisupervised learning
  - Use power of unlabeled data combined with few labels
  - Quite successful in google: https://research.googleblog.com/2016/10/graph-powered-machine-learning-at-google.html

Distant supervision in Raptor

- Positive examples
  - Gather list of persons with adverse news
  - Run names through search engine
  - Gather their mentions alongside bad keywords

- Negative examples
  - Random sentences
    - Does not catch tough sentences
  - Need negative examples which are similar to positive ones (like “Amy murdered **John**”)
    - Hard to get

Negative examples

- Get list of judges and attorneys (usually mentioned around adverse topics, but not affected by them)
- Simple rules like: “X said”
- Syntactic parsing and rules based on that
  - This is usually slow in prediction time, but useful here
- Some sources have different rate of real adverse news vs things looking like adverse news (news pages vs court proceedings)

Distant supervision heuristics

- Combination of multiple rules (even previous version of classifier)

- Inspect random samples of results of heuristics. If they are like 90% accurate, they can be safely used.

Distant supervision caveats

- Don’t run cross validation on distant supervision data directly

- Use validation set with correct labels inputted by hand
  - Size of training set is driven by model size -> this needs to be huge
  - Size of validation set is driven by statistical error -> this can be small, but ...

Modelling approaches

- Features + logistic regression
  - Time spent in featurizing (usually productive)
  - Features
    - Bag of words / bigrams / skipgrams
    - Words around mentioned entity
      - Also skip if there are multiple entities next to each other
        - John, Amy and George were arrested ...
  - Easy to understand and debug
  - Fast prediction time

Modelling approaches

- Deep learning (RNN/CNNs)
  - Saves time on feature engineering
  - Time spent playing with neural architecture (usually not so productive)
  - Pretrained word embeddings help a lot

- Similar performance currently (at 99% recall, we throw out around 70% of false positives from previous IR filter)

Lessons learned

- Discipline needed
  - Playing with neural nets is tempting but not productive
  - Playing with distant supervision heuristics / features is sometimes boring, but brings results
- Classic industry approach works a lot
  - Read a lot of research papers
  - Implement baseline which lots of them compare to
- Real challenge is building models without good training data
- Tricky work is usually outside standard ML course syllabus
  - But right model choice is always important

Other open problems

- Mention detection (especially with weird Arabic names)
- Coreference
- Analysis of formatted text (tables, …)


That's all Folks!