Deep Learning for Time-Series

Introduction

Laurent El Ghaoui

MFE 230T-2
Haas Business School
UC Berkeley

8/13/2020

Outline

Context: explosion of research

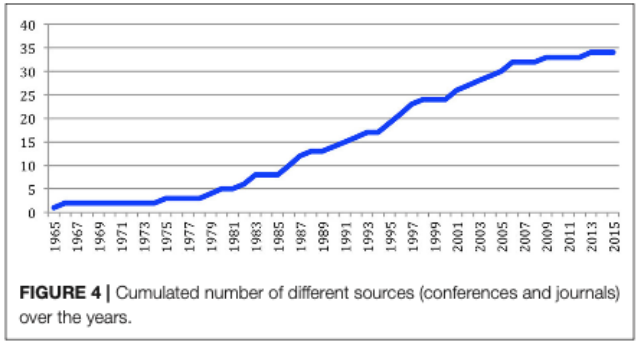


FIGURE 4 | Cumulated number of different sources (conferences and journals) over the years.

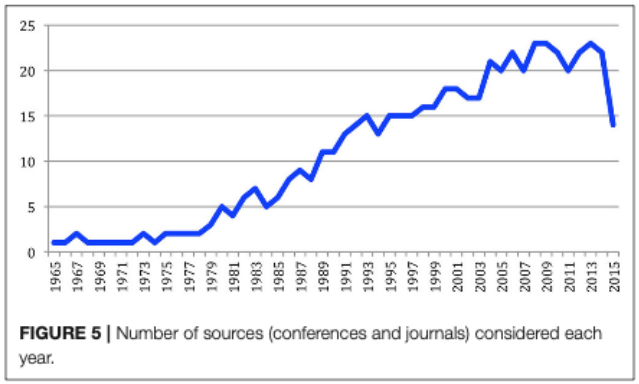


FIGURE 5 | Number of sources (conferences and journals) considered each year.

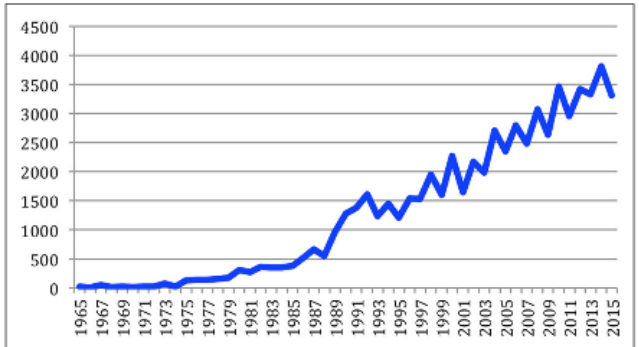


FIGURE 6 | Number of papers each year.

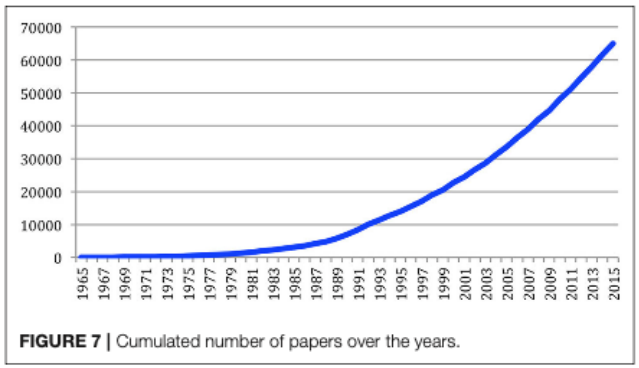
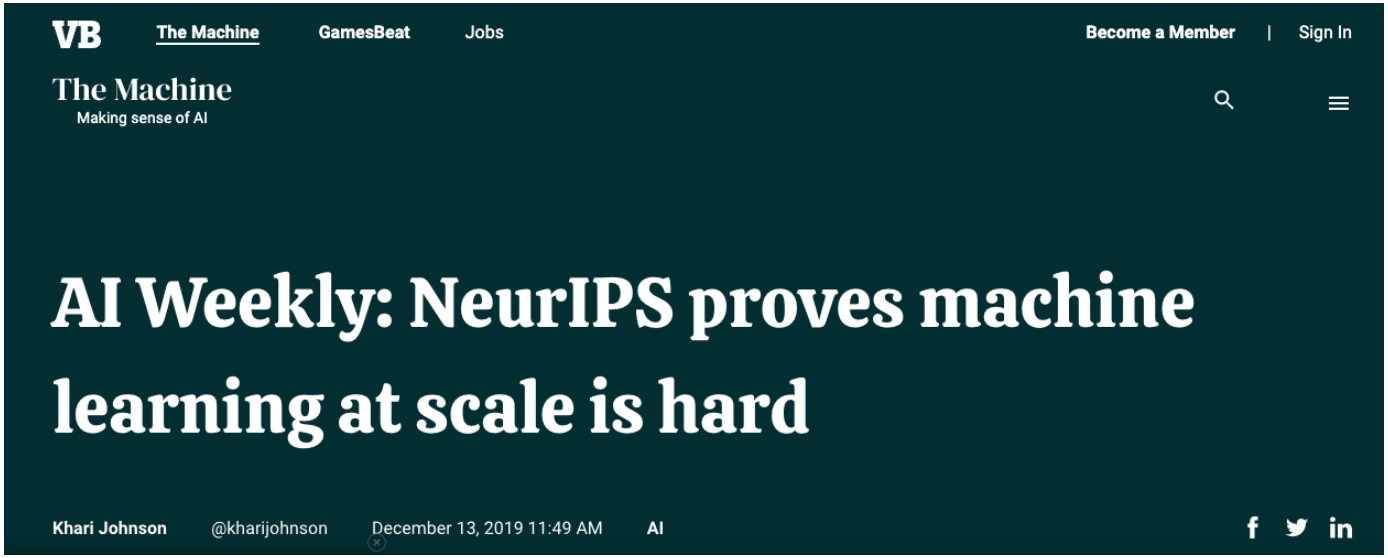


FIGURE 7 | Cumulated number of papers over the years.

Growth of NLP. Source: https://www.frontiersin.org/articles/10.3389/frma.2018.00036/full

- Enormous growth of the amount of publications in machine learning / AI / data science
- Very hard to keep up with the firehose

Conferences: the case of NeurIPS



caption

"All told, NeurIPS 2019 welcomed 13,000 attendees, up 40 % from the prior year. In all, the conference has quadrupled in size in the past five years. Organizers started to use a lottery system this year, after 9,000 tickets to 2018 NeurIPS sold out in 12 minutes."

It's not research as usual anymore

- Big corporate actors now play a massive role:
  - Setting up large AI research labs
  - Big recruiting efforts
  - Large internal resources (data sets, computing clouds)
  - Publicizing activities, *e.g.* "Google at NeurIPS 2019".
  - Most research follows / sets specific trends (*e.g.*, transformers in NLP)
- Research is highly marketed (on social media)
- See "Troubling Trends in Machine Learning Scholarship" (Lipton & Steinhardt, 2019)

How to cope with the firehose

- Use NLP!
- Learn how to quickly critically evaluate the quality of papers:
  - Quickly try the proposed methods
  - Don't necessarily always trust the published results

Outline

Course goals and topics

Goal: “A self-directed introduction to deep learning models for time-series, with a focus on finance applications”.

Some major themes are, within finance and time-series:

- Completion problems;
- Outlier detection;
- NLP;
- RL.

Other topics, as suggested by student groups, are possible.

One of the goals of this course is to learn how to deal with the explosion of scientific publications.

Outline

Course Approach

- Students form groups of **5**. Due to scheduling constraints, and a high enrollment, smaller groups are not possible.
- Each group is responsible for developing a project based on a published paper, an in-depth study that involves numerical experiments on real data, for example comparing two approaches to sentiment classification.
- Deliverables: three 15 min **presentations** and a report.
  - *8/13/2020:* You must have formed your group and chosen your paper by then.
  - *8/20/2020:* Introduce the paper, why you chose it, and describe your project.
  - *8/27/2020:* Updates on project, describe the challenges you encounter, and next steps.
  - *9/3/2020:* final presentation, presenting the results of your study, recommendations for further explorations.
- The **report** should be concise and contain an introduction to the paper, main findings, critical review of the paper, and a numerical experiment, with comments on the results.
- Each group is welcome to propose a specific paper; the teaching staff will suggest a few choices.

Example: NLP for finance

- *Paper:* "FinBERT: pre-trained model on SEC filings for financial natural language tasks", at https://www.researchgate.net/publication/334974348_FinBERT_pre-trained_model_on_SEC_filings_for_financial_natural_language_tasks
- *Project:* based on the above paper, evaluate the model on a real data set; using FinBERT for sentiment analysis, maybe on earnings calls data if possible; compare to a baseline such as Naive Bayes.

Deliverables

- The **three 15-min presentation**s, in powerpoint format, each to be submitted by the end of day of the presentation, in order to allow teams to incorporate minor changes in response to questions or suggestions from the instructor.
- A **final report** (10 pages max) is due Tuesday 9/8/2020, 8am PST.

Logistics

- *Instructor:* Laurent El Ghaoui (elghaoui@berkeley.edu).
- *TAs:*
  - Vinicio de Sola (vinicio.desola@gmail.com),
  - Victor Lu (victor_lu@berkeley.edu).
- *Groups:* Put down the names on the Google sheet: https://docs.google.com/spreadsheets/d/1ABDlQdLsATzhyY81drfWD-UYtut70l1XofiYf1ULGIo/edit?usp=sharing.
- *Discussion sections:* See the calendar.
- *Homeworks:* None. Instead, each group will be responsible for developing three presentations and a report.
- *Grading:* 15 % for each presentation, 55 % final report.

How do we communicate?

- Preferred way: Slack.
- Email to me. Always cc Vinicio and Victor!