(Short) Deep Dive into Deep Nets

Outline

- Part A: Glass is half-full
  - Deep nets can do much
- Part B: Glass is half-empty
  - There is always more work to do

Make deep nets accessible to masses (including non-programmers)

Advocate combo of Part A with
1. AI Knowledge Representation
2. Linguistics and Philosophy

- GFT Code,
- 100s of examples,
- slides, videos, papers

Standard 3-Step Recipe

Most users should not do this themselves

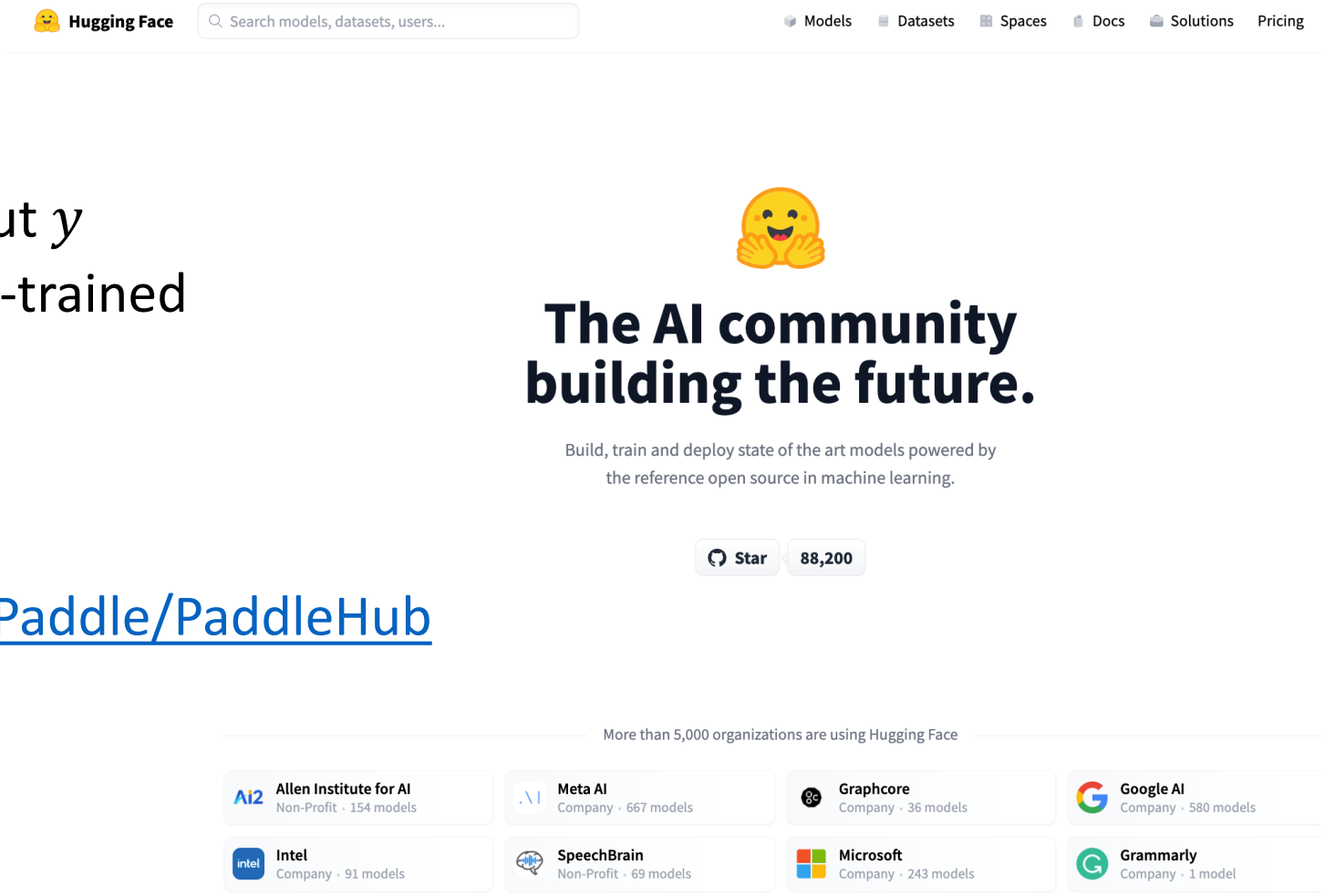
Step	gft Support	Description	Time	Hardware
1		Pre-Training	Days/Weeks	Large GPU Cluster
2	*gft_fit*	Fine-Tuning	Hours/Days	1+ GPUs
3	*gft_predict*	Inference	Seconds/Minutes	0+ GPUs

Most users should not invest in pretraining because growth (& costs) are out of control

Year	Deep nets	Billions of parameters
2016	ResNet-50 (He *et al.*, 2016)	0.023
2019	BERT (Devlin *et al.*, 2019)	0.34
2019	GPT-2 (Radford *et al.*, 2019)	1.5
2020	GPT-3 (Brown *et al.*, 2020; Dale 2021)	175
2022	PaLM (Chowdhery *et al.*, 2022)	540

Resources: Hubs

- Repository for sharing
  - Datasets: input x and output y
  - Models: pre-trained & post-trained
  - Tutorials
- Examples
  - https://huggingface.co/
  - https://github.com/PaddlePaddle/PaddleHub



gft (general fine-tuning):
A Little Language for Deep Nets

Standard 3-Step Recipe

Step	gft Support	Description	Time	Hardware
1		Pre-Training	Days/Weeks	Large GPU Cluster
2	*gft_fit*	Fine-Tuning	Hours/Days	1+ GPUs
3	*gft_predict*	Inference	Seconds/Minutes	0+ GPUs

- Terminology borrowed from regression:
  - *fit*: $f_{pre} + data \rightarrow f_{post}$
  - *predict*: $f(x) \rightarrow \hat{y}$
- *fit* and *predict* are all you need
  - *gft* programs are short (1-line)
  - No (not much) programming required
    - No python in this tutorial
    - Examples on hubs are (unnecessarily) long/complicated

gft (general fine-tuning):
A Little Language for Deep Nets

Standard 3-Step Recipe

Step	gft Support	Description	Time	Hardware
1		Pre-Training	Days/Weeks	Large GPU Cluster
2	*gft_fit*	Fine-Tuning	Hours/Days	1+ GPUs
3	*gft_predict*	Inference	Seconds/Minutes	0+ GPUs

- Terminology borrowed from regression:
  - *fit*: $f_{pre} + data \rightarrow f_{post}$
  - *predict*: $f(x) \rightarrow \hat{y}$
- *fit* and *predict* are (almost) all you need
  - *gft* programs are short (1-line)
  - No (not much) programming required
    - No python in this tutorial
    - Examples on hubs are (unnecessarily) long/complicated

Examples of 1-line GFT Programs

Step 2: *gft_fit*

```
gft_fit --eqn 'classify: label ~ text' \
    --model H:bert-base-cased \
    --data H:emotion \
    --output_dir $outdir
```

Data

f_{pre}: Pre-trained Model

f_{post}: Post-trained Model

Step 3: *gft_predict*

```
# text-classification: sentiment analysis
echo 'I love you.' | gft_predict --task text-classification
# I love you.   POSITIVE    0.9998705387115479
```

x $\hat{y}$ Score

GFT makes Deep Nets look like Regression

- Terminology borrowed from regression:
  - *fit*: $f_{pre} + data \rightarrow f_{post}$
  - *predict*: $f(x) \rightarrow \hat{y}$
- Demystify deep nets
  - No one would suggest regression is ``intelligent''

Simple Equations Cover Many Cases of Interest
GLUE: A Popular Benchmark

Subtask	Dataset	Equation
COLA	H:glue,cola	$classify: label \sim sentence$
SST2	H:glue,sst2	$classify: label \sim sentence$
WNLI	H:glue,wnli	$classify: label \sim sentence$
MRPC	H:glue,mrpc	$classify: label \sim sentence1 + sentence2$
QNLI	H:glue,qnli	$classify: label \sim sentence1 + sentence2$
QQP	H:glue,qqp	$classify: label \sim question + sentence$
SSTB	H:glue,sstb	$regress: label \sim question1 + question2$
MNLI	H:glue,mnli	$classify: label \sim premise + hypothesis$

More Standard Benchmarks

Task	Subtask	Dataset	Equation
GLUE	COLA	H:glue,cola	$classify: label \sim sentence$
SQuAD 1.0		H:squad	$classify\_spans: answers \sim question + context$
SQuAD 2.0		H:squad_v2	$classify\_spans: answers \sim question + context$
CONLL2003	POS	H:conll2003	$classify\_tokens: pos\_tags \sim tokens$
	NER	H:conll2003	$classify\_tokens: ner\_tags \sim tokens$
TIMIT		H:timit_asr	$ctc: text \sim audio$
Amazon Reviews		H:amazon_reviews_multi	$classify: label \sim question + sentence$

Tasks

- classify, text-classification — $y \in \{0,1,2, \ldots\}$
- regress — $y \in \mathbb{R}$ or $y \in \mathbb{R}^N$
- QA, Question Answering, classify spans — y for each start/end of span
- token classification — y for each token
  - NER (Named Entity Recognition)
  - POS (Part of Speech Tagging)
- translation, MT
- ASR, Automatic Speech Recognition, ctc — y for each phoneme

Prompts (Questions)

- Prompt: *What capital in Europe has the highest elevation?*

The capital city with the highest elevation in Europe is La Paz, Bolivia, which is located at an elevation of over 3,600 meters (11,800 feet) above sea level. However, La Paz is not in Europe, it is in South America. The capital city with the highest elevation in Europe is Andorra la Vella, the capital of Andorra, which is located at an elevation of 1,023 meters (3,356 feet) above sea level. It is followed by Bern, the capital of Switzerland, which is located at an elevation of 554 meters (1,818 feet) above sea level.

- Prompt Engineering:
  - Find valuable prompts
    - (prompts that solve an information need)
  - Prompt engineering is like fine-tuning
    - (but does not require as much access to model)

Part A: Glass is half full (deep nets can do much)
Part B: Glass is half empty (always more to do)

- Advocate combo of Part A with

1. AI Knowledge Representation
2. Linguistics and Philosophy

- Pendulum Swung Too Far
  - There have been many AI Winters
  - Often, after ``irrational exuberance''
  - (like current excitement with nets)

- We tend to be impressed by people speak/write well
  - Fluency → well-read → success → smart
- Machines are better than people on many tasks (spelling),
  - Now that machines are more fluent than people
  - are they smarter?
- Fear: AI Winter
  - there will be disappointment when public figures out that fluency ≠intelligence

A Pendulum Swung Too Far

Linguistic Issues in Language Technology 6 (2011).

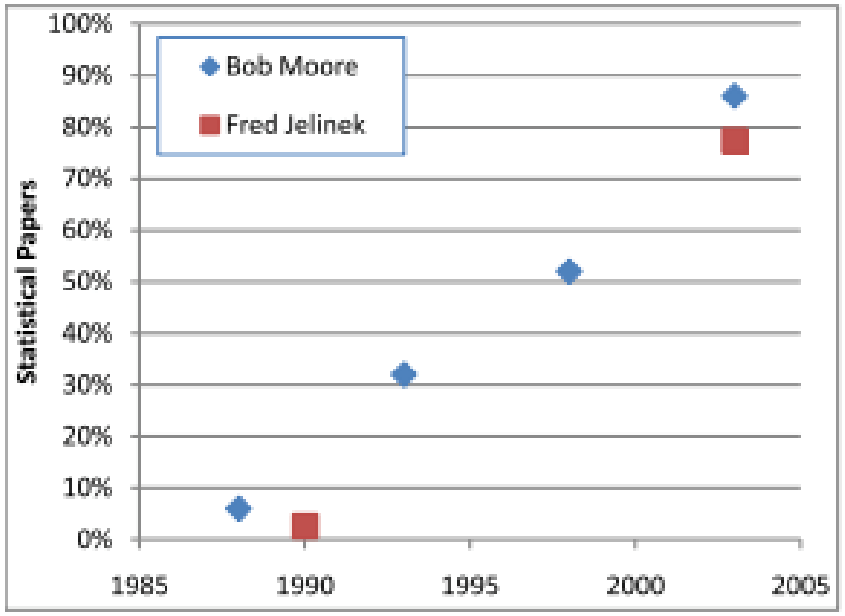
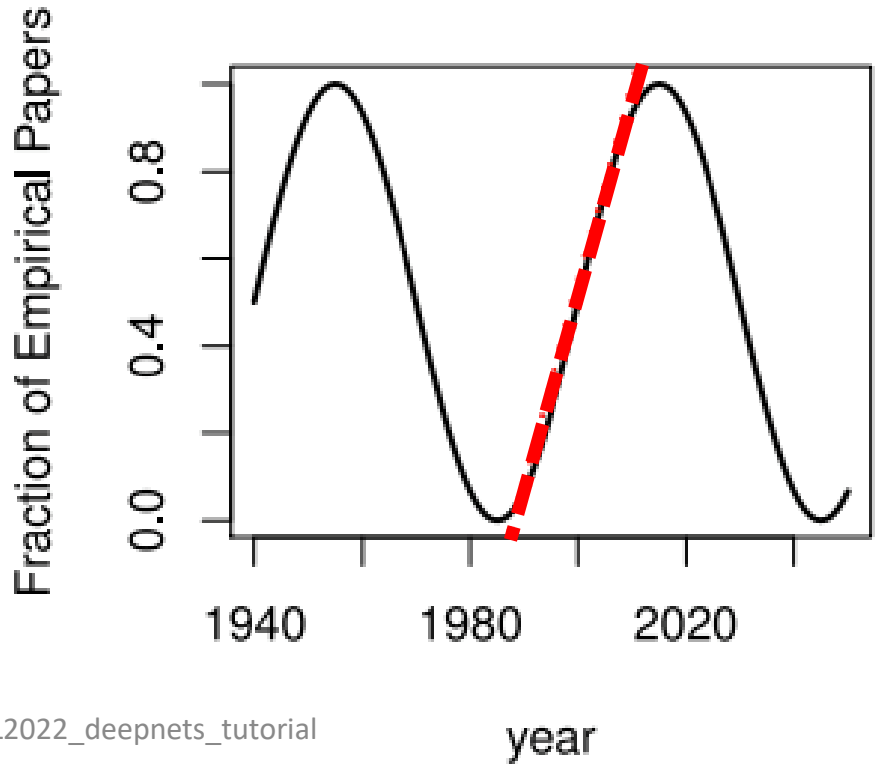


FIGURE 1 The shift from Rationalism to Empiricism is striking (and no longer controversial). This plot is based on two independent surveys of ACL meetings by Bob Moore and Fred Jelinek (personal communication).

- 1950s: Empiricism (Shannon, Skinner, Firth, Harris)
- 1970s: Rationalism (Chomsky, Minsky)
- 1990s: Empiricism (IBM, AT&T Bell Labs)



Large Language Models (LLMs)

- History
  - 1950s: Empiricism
    - Shannon, Skinner, Firth, Harris
    - *You shall know a word by the company it keeps* (Firth, 1957)
  - PMI (Church & Hanks, 1990)
  - Word2Vec
  - BERT
- Strengths:
  - fluency
  - word-associations (psychology)
- Collocations (Distributional Hypothesis):
  - find patterns that are more freq than chance
  - no meaning (semantics), syntax, logic, truth

And Weaknesses:

- 1970s: Rationalism
  - Chomsky, Minsky
- Truth:
  - logical form
  - temporal/spatial logic
  - possible worlds
- Meaning
- Purpose
  - Planning/Discourse Structure
- Common Sense Knowledge

Semantics (Negation)

"What BERT is not" (Allyson Ettinger, TACL-2020)

- Given "A robin is a ____." BERT fills in "bird" 
- Given "A robin is not a ____." ditto 

Semantics (Negation)

"What BERT is not" (Allyson Ettinger, TACL-2020)

- Given "A robin is a ____." BERT fills in "bird" 
- Given "A robin is not a ____." ditto 

Collocations, Synonyms & Antonyms

- *You shall know a word by the company it keeps* – Firth (1957)
- By construction, embeddings (PMI, Word2Vec, BERT)
  - find words that appear near one another (more than chance)
  - including both synonyms as well as antonyms
  - (since corpora are full of comparisons & contrasts)
- In Word2Vec, *moral* is closer to *immoral* than to *good*.
  - (since *moral* and *immoral* appear in similar contexts more than near *good*)
  - Collocation ≠ Synonymy

Hallucinations / Possible Worlds

- LMs have a tendency to ``hallucinate'' when summarizing documents
- Output sounds plausible,
  - but may add embellishments to the input
- LMs tend to make up ``alternative facts''
  - faster than they can be fact-checked

- This may well be their most dangerous failing
- People might believe some of these conspiracy theories

Reliability:

Experiment run by Ken Church

- **In 1885, Stanford University was founded** by Leland and Jane Stanford in memory of their only child, recently deceased.
- In 1887, The first successful gasoline powered automobile in the United States was manufactured in Cleveland, Ohio by Charles and Frank Duryea.
- In 1887, The American Federation of Labor was founded.
- In 1887, The Salvation Army was formally incorporated in London, England.
- In 1890, The American Red Cross was established by Clara Barton.

…

In 1893, The first Miss America Pageant was held in New York City.

In 1895, The National Association for the Advancement of Colored People (NAACP) was founded.

In 1895, The National Council of Women of the United States was founded.

In 1897, The Boston Symphony Orchestra gave its first performance.

In 1897, The first winner of the Nobel Peace Prize was born.

In 1899, The first automobile race was held at Narragansett Park in Cranston, Rhode Island.

In 1900, The first automobile race was held at Narragansett Park in Cranston, Rhode Island.

In 1900, The National Association for the Advancement of Colored People (NAACP) was founded.

Looks great!

But:

- **ALL THE DATES ARE WRONG.**
- **SOME ARE INCONSISTENT.**

How far is Kyiv from Russia?

How far is Kyiv from Russia

NATURAL LANGUAGE | MATH INPUT

Input interpretation

distance	from	center of Russia
	to	Kyiv, Kiev, Ukraine

Result

2690 miles

Unit conversions

Linguistic Approaches

Considerable literature on time and space in linguistics

- Temporal logics
- Space (Bloom, 1999)

Rich set of connections between

- Surface form (syntax)
- Deeper structures (semantics)

Challenge for future work

- Combine recent progress on deep nets with
  - Decades of work in AI Representation
  - and centuries of work in Linguistics