The Age of Talking

Giuseppe Attardi
Dipartimento di Informatica
Università di Pisa

Firenze, November30, 2016

- Children reach the age of talking at 3 years
- When will computers reach the age of talking?
- Are we making progress?
- What are the promising directions?
- How to exploit large processing capabilities and big data?
- Can we take inspiration from biology?

Motivation

- Language is the most distinctive feature of human intelligence
- Language shapes thought
- Emulating language capabilities is a scientific challenge
- Keystone for intelligent systems

2001 a space Odyssey: 40 years later

Computer chess

Audio-video communication

On board entertainment

Technology surpassed the vision

Internet
The Web
Smartphones
Genomics
nmanned space exploration
Home computing
Big data

Except for

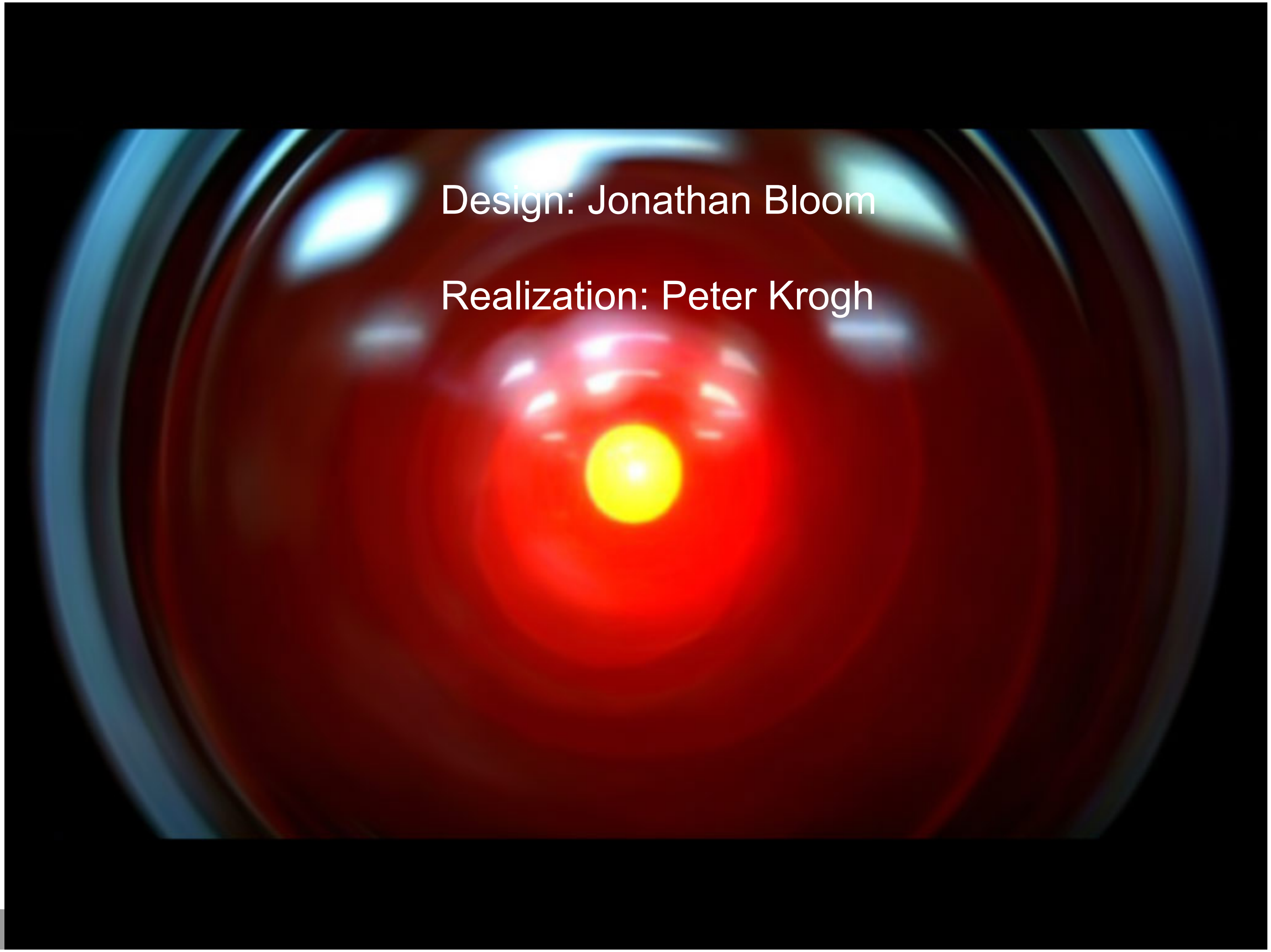
Computer Speech

Computer Vision

Computer cognition

Speech technology in 2001: the vision

Speech technology in 2001: the reality



Machine Translation, circa 2001

Lo spirito è forte ma la carne è debole

tradotto in russo

La vodka è forte ma la bistecca è tenera

apocrifo

Machine Translation Progress

- Gli chiese di riorganizzare Forza Italia

  The churches to reorganize Italy Force (Altavista)

  She asked him to reorganize Forza Italia (Google)

- Il ministro Stanca si è laureato alla Bocconi

  The Minister Stanca graduated at Mouthfuls (Altavista)

  The Minister Stanca is a graduate of Bocconi (Google)

How to learn natural language

- Children learn to speak naturally, by interacting with others
- Nobody teaches them grammar
- Is it possible to let computer learn language in a similarly natural way?

Statistical Machine Learning

- Supervised Training
- Annotated document collections
- Ability to process Big Data
  - If we used same algorithms 10 years ago they would still be running
- Similar techniques for speech and text

Recent Breakthrough

- Speech
- Question Answering
  - IBM Watson
  - battuti i campioni d
    televisivo Jeopardy

Quiz Bowl Competition

- Iyyer et al. 2014: A Neural Network for Factoid Question Answering over Paragraphs
- QUESTION:

  He left unfinished a novel whose title character forges his father's signature to get out of school and avoids the draft by feigning desire to join.

  One of his novels features the jesuit Naptha and his opponent Settembrini, while his most famous work depicts the aging writer Gustav von Aschenbach.

  Name this German author of The Magic Mountain and Death in Venice.
- ANSWER: Thomas Mann

QANTA vs Ken Jennings

- QUESTION:

  Along with Evangelista Torricelli, this man is the namesake of a point the minimizes the distances to the vertices of a triangle.

  He developed a factorization method …

  ANSWER: Fermat

- QUESTION:

  A movie by this director contains several scenes set in the Yoshiwara Nightclub.

  In a movie by this director a man is recognized by a blind beggar because he is wistlin
  In the hall of the mountain king.

  ANSWER: Fritz Lang

Early history of NLP: 1950s

- Early NLP (Machine Translation) on machines less powerful than pocket calculators
- Foundational work on automata, formal languages, probabilities, and information theory
- First speech systems (Davis et al., Bell Labs)
- MT heavily funded by military – a lot of it was just word substitution programs but there were a few seeds of later successes, e.g., trigrams
- Little understanding of natural language syntax, semantics, pragmatics
- Problem soon appeared intractable

Recent Breakthroughs

- Watson at Jeopardy! Quiz:
  - http://www.aaai.org/Magazine/Watson/watson.php
  - Final Game
  - PBS report
- Google Translate on iPhone
  - http://googleblog.blogspot.com/2011/02/introducing-google-translate-app-for.html
- Apple SIRI

Smartest Machine on Earth

- IBM Watson beats human champions at TV quiz Jeopardy!
- State of the art Question Answering system

Tsunami of Deep Learning

- AlphaGo beats human champion at Go.
- RankBrain is third most important important factor in the ranking algorithm along with links and content at Google
- RankBrain is given batches of historical searches and learns to make predictions from these
- It learns to deal also with queries and words never seen before
- Most likely it is using Word Embeddings

Apple SIRI

- ASR (Automated Speech Recognition) integrated in mobile phone
- Special signal processing chip for noise reduction
- SIRI ASR
- Cloud service for analysis
- Integration with applications

Google Voice Actions

- Google: what is the population of Rome?
- Google: how tall is Berlusconi
- How old is Lady Gaga
- Who is the CEO of Tiscali
- Who won the Champions League
- Send text to Gervasi Please lend me your tablet
- Navigate to Palazzo Pitti in Florence
- Call Antonio Cisternino
- Map of Pisa
- Note to self publish course slides
- Listen to Dylan

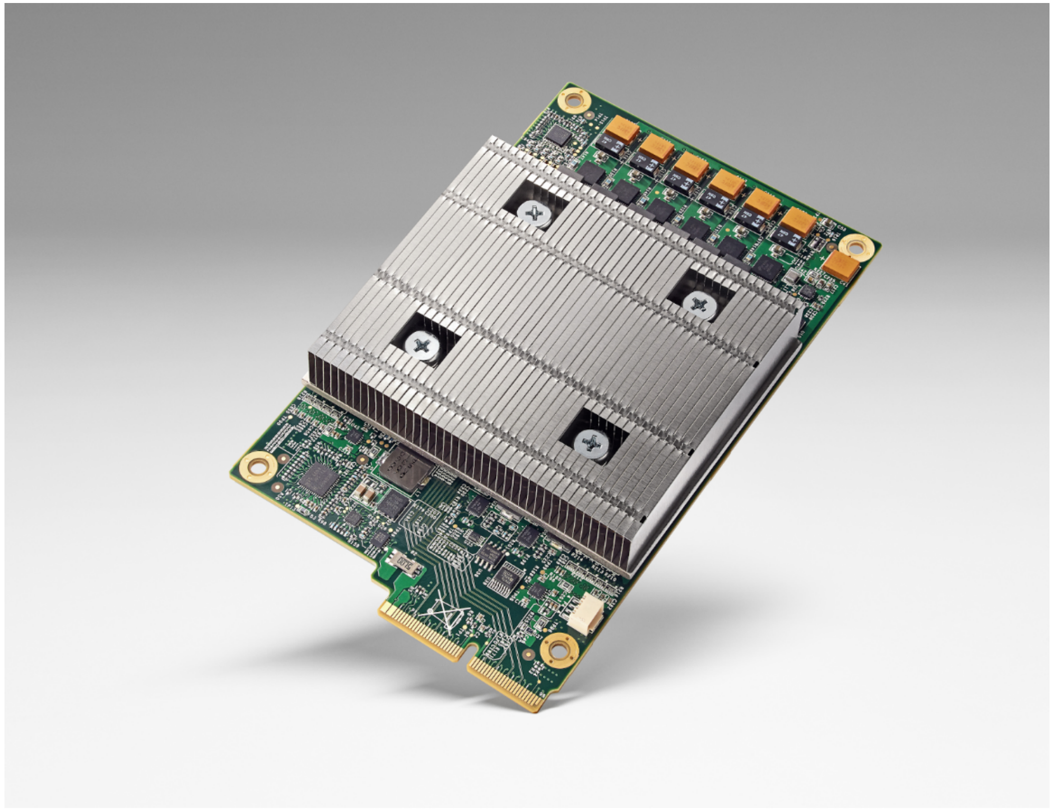
Technological Breakthrough

Technological Breakthrough

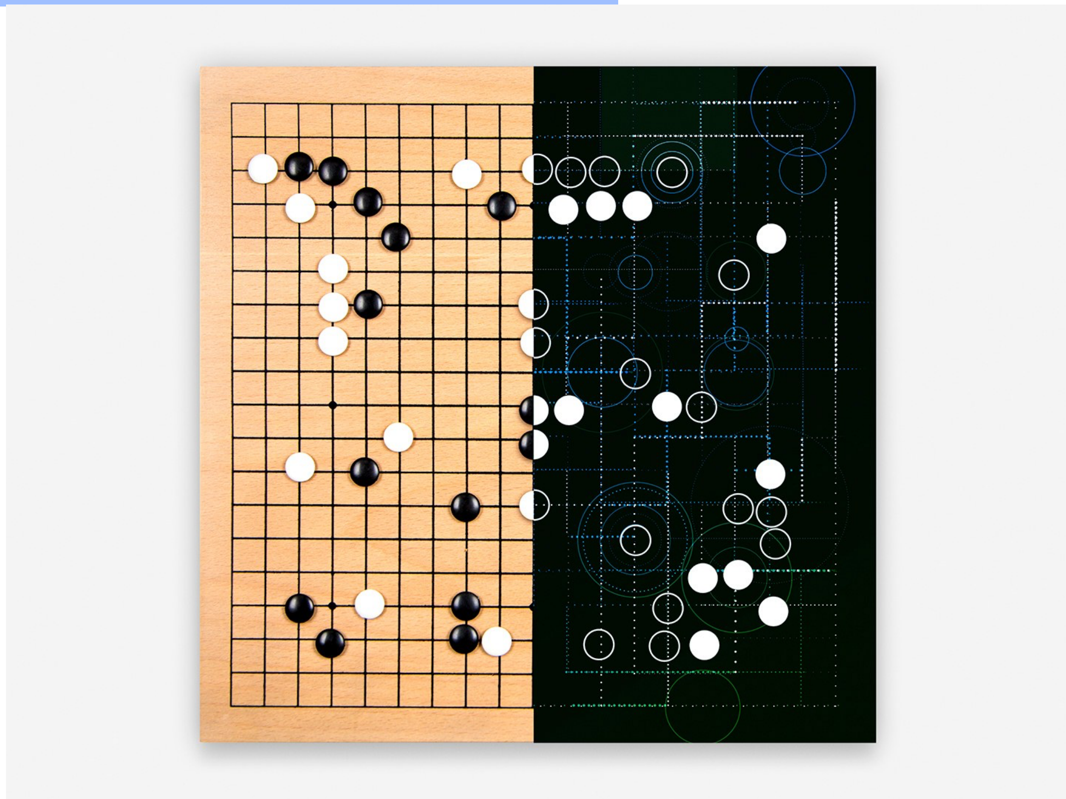
- Machine learning
- Huge amount of data
- Large processing capabilities

Big Data & Deep Learning

- Requires high speed computing
- Typical using GPUs
  - Eg. nVIDIA TESLA
- Google custom chip TPU
- 10x more power per watt
- Half precision floats



Google Deep Mind



- January 2016: AlphaGo beats world champion at Go

Unreasonable Effectiveness of Data

- Halevy, Norvig, and Pereira argue that we should stop acting as if our goal is to author extremely elegant theories, and instead embrace complexity and make use of the best ally we have: the unreasonable effectiveness of data.
- A simpler technique on more data beat a more sophisticated technique on less data.
- Language in the wild, just like human behavior in general, is messy.

Scientific Dispute: is it science?

Prof. Noam Chomsky, Linguist, MIT

Peter Norvig, Director of research, Google

Statistical Machine Learning

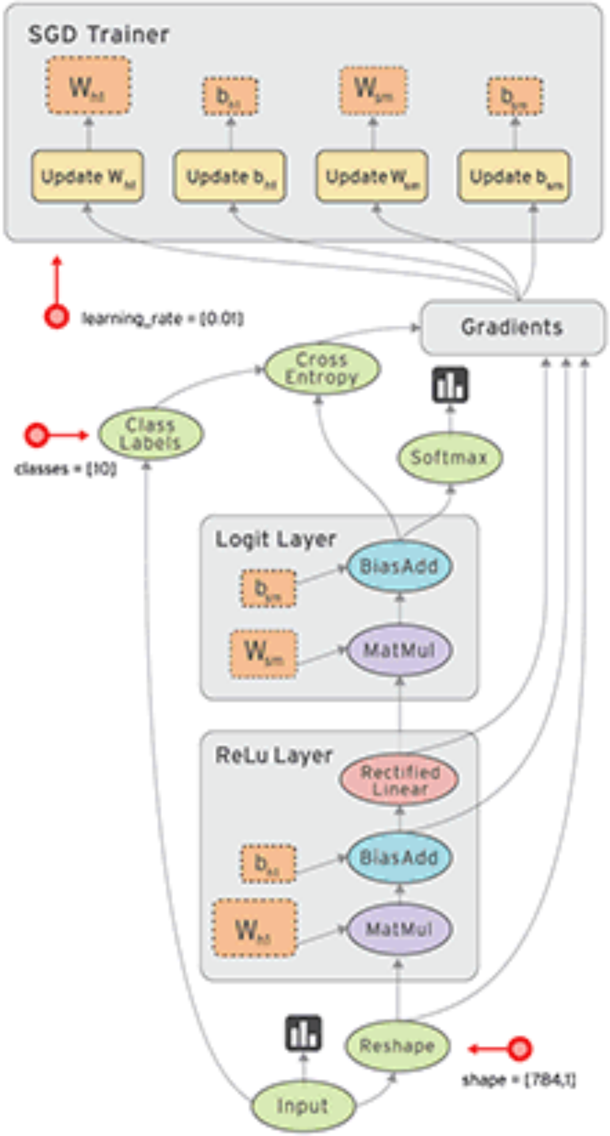
- Training on large amounts of data
- Requires ability to process Big Data
  - If we used same algorithms 10 years ago they would still be running
- The Unreasonable Effectiveness of Big Data
  - Norvig vs Chomsky controversy

Supervised Statistical ML Methods

- Learn from training examples
- Freed us from devising rules or algorithms
- Requires creation of annotated training corpora
- Imposed the tyranny of feature engineering

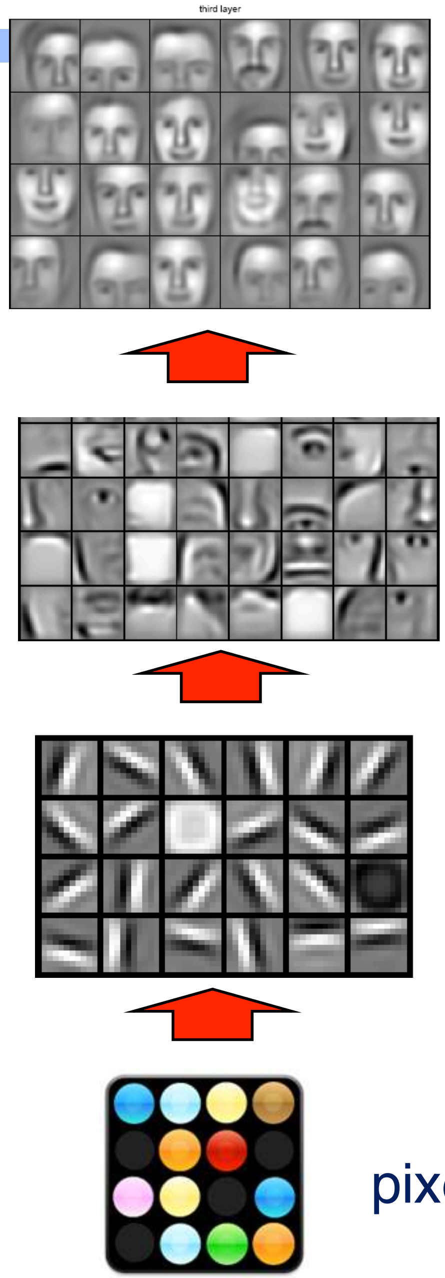
Deep Learning

- Design a model architecture
- Define a loss function
- Run the network letting the parameters and the data representations self-organize as to minimize this loss
- End-to-end learning: no intermediate stages nor representations



Deep Neural Network

Training set:
aligned images of faces



object models

object parts
(combination
of edges)

edges

pixels

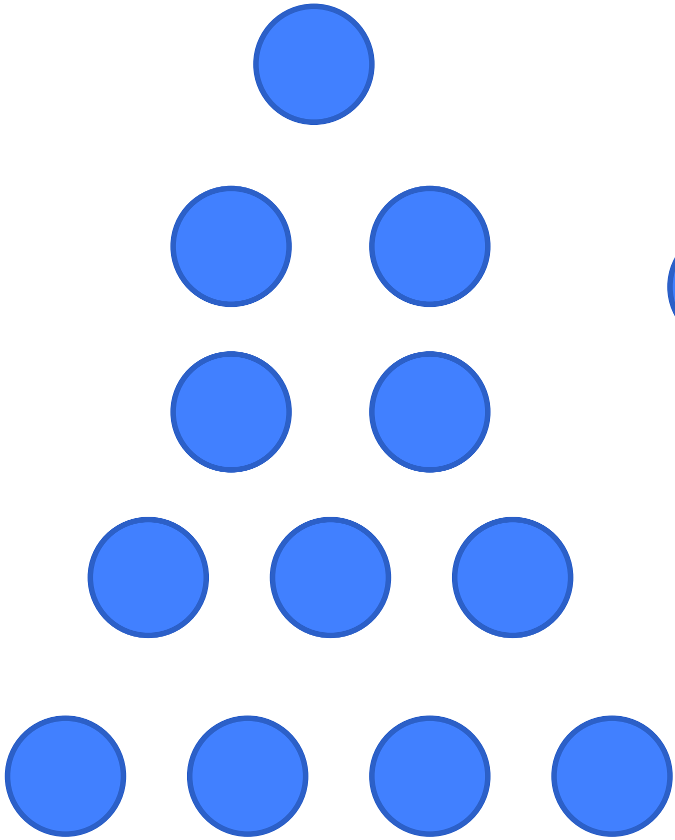
slide from Honglak Lee

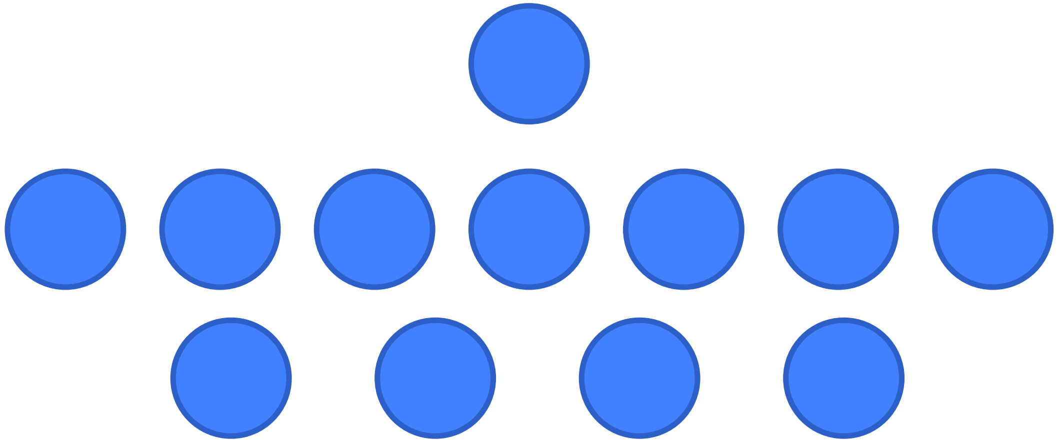
Deep vs Shallow

- Given the same number of non-linear units, a deep architecture is more expressive than a shallow one (Bishop 1995)
- Two layer (plus input layer) neural networks have been shown to be able to approximate **any** function
  - RNNs are Turing-complete
- However, functions compactly represented in k layers may require ***exponential*** size when expressed in 2 layers

Deep Network

Shallow Network





Shallow (2 layer) networks need a lot more hidden layer nodes to compensate for lack of expressivity

In a deep network, high levels can express combinations between features learned at lower levels

Problem

- Training deep network faces the ***vanishing gradient problem***
- Gradients tend to get smaller while propagating backwards through the hidden layers
- Neurons in the bottom layers learn much more slowly

2006: The Deep Breakthrough

- Hinton, Osindero & Teh « A Fast Learning Algorithm for Deep Belief Nets », *Neural Computation*, 2006
- Bengio, Lamblin, Popovici, Larochelle « Greedy Layer-Wise Training of Deep Networks », *NIPS'2006*
- Ranzato, Poultney, Chopra, LeCun « Efficient Learning of Sparse Representations with an Energy-Based Model », *NIPS'2006*

Slide credit : Yoshua Bengio

Deep Learning Breakthrough: 2006

- Unsupervised learning of shallow features from large amounts of unannotated data
- Features are tuned to specific tasks with second stage of supervised learning

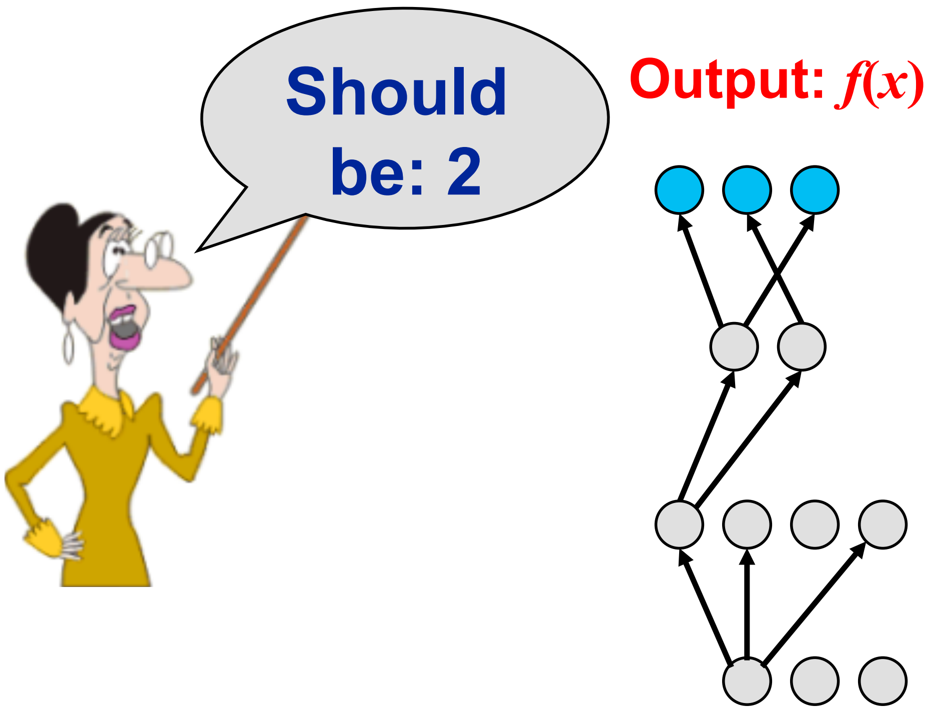
Unsupervised Training

- Far more un-labeled data in the world (i.e. online) than labeled data:
  - Websites
  - Books
  - Videos
  - Pictures
- Deep networks take advantage of unlabelled data by learning **good representations** of the data **through unsupervised learning**
- Humans learn initially from unlabelled examples
- Babies learn to talk without labeled data

Unsupervised Feature Learning

- Learning features that represent the data allows them to be used to train a supervised classifier
- As the features are learned in an unsupervised way from a different and larger dataset, **less risk of over-fitting**
- **No need for manual feature engineering**
  - (e.g. Kaggle Salary Prediction contest)
- Latent features are learned that attempt to explain the data

Supervised Fine Tuning



End to End

- Deep layers mean that there is no need to split artificially problem into subtasks
- For example:
  - no need for POS
  - even no need for tokenization!!!
- **No need for manual feature engineering**
  - (e.g. Kaggle Salary Prediction contest)
- Latent features are learned that attempt to explain the data

Enabling Factors

- Training of deep networks was made computationally feasible by:
  - Faster CPU's
  - Parallel CPU architectures
  - Advent of GPU computing
- Neural networks are often represented as a matrix of weight vectors
- GPU's are optimized for very fast matrix multiplication
- 2008 - Nvidia's CUDA library for GPU computing is released

Application Areas

- Typically applied to image and speech recognition, lately also NLP
- Each are non-linear classification problems where the inputs are highly hierarchal in nature (language, images, etc)
- The world has a hierarchical structure – Jeff Hawkins – On Intelligence
- Problems that humans excel in and machine do very poorly

State of the Art in Many Areas

- Speech Recognition (2010, Dahl et al)
- MNIST hand-written digit recognition (Ciresan et al, 2010)
- Image Recognition (GoogLeNet won ILRSRVC 2014 challenge with a 27-layer net)
- Andrew Ng, Stanford:

  "I've worked all my life in Machine Learning, and I've never seen one algorithm knock over benchmarks like Deep Learning"

SOTA besides Image and Speech

- activity of potential drug molecules
- analysing particle accelerator data
- reconstructing brain circuits
- predicting the effects of mutations in non-coding DNA on gene expression and disease
- natural language understanding:
  - topic classification
  - sentiment analysis
  - question answering
  - language translation

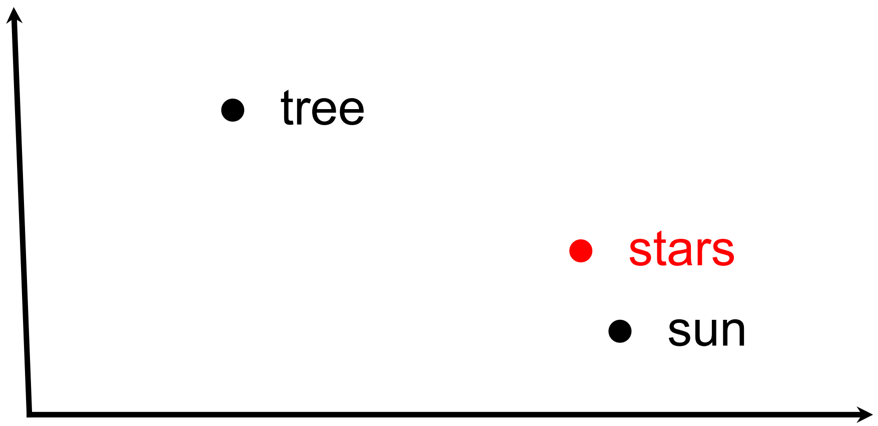
DL for NLP

Distributional Semantics

- Co-occurrence counts

	shining	bright	trees	dark	look
stars	38	45	2	27	12

- High dimensional sparse vectors
- Similarity in meaning as vector similarity?

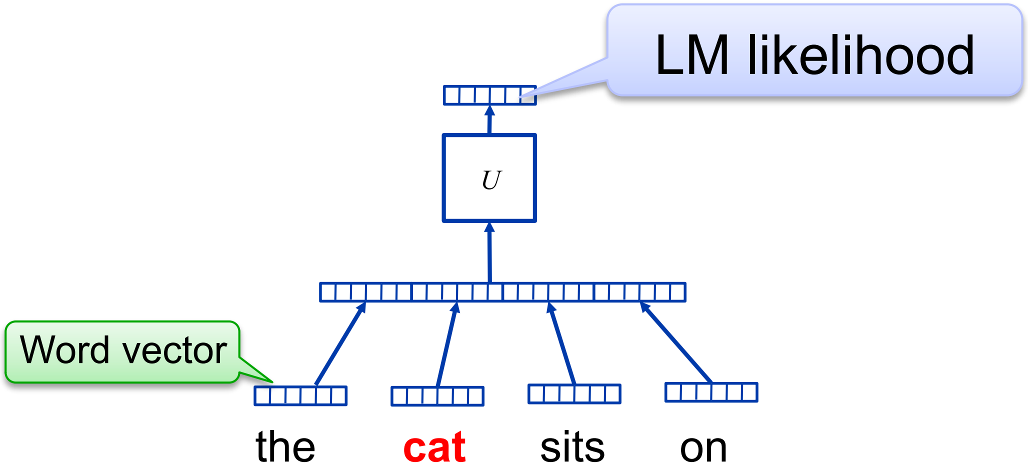


Co-occurrence Vectors

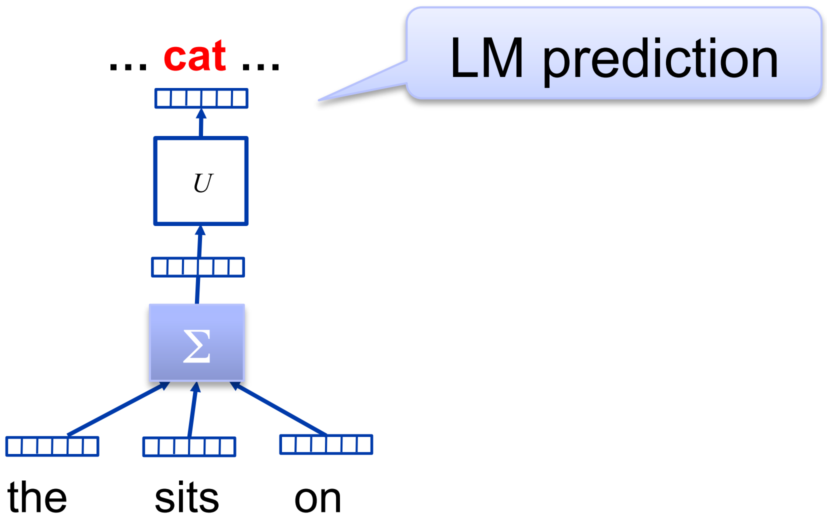
FRANCE 454	JESUS 1973	XBOX 6909	REDDISH 11724	SCRATCHED 29869	MEGABITS 87025
PERSUADE	THICKETS	DECADENT	WIDESCREEN	ODD	PPA
FAW	SAVARY	DIVO	ANTICA	ANCHIETA	UDDIN
BLACKSTOCK	SYMPATHETIC	VERUS	SHABBY	EMIGRATION	BIOLOGICALLY
GIORGI	JFK	OXIDE	AWE	MARKING	KAYAK
SHAFFEED	KHWARAZM	URBINA	THUD	HEUER	MCLARENS
RUMELLA	STATIONERY	EPOS	OCCUPANT	SAMBHAJI	GLADWIN
PLANUM	GSNUMBER	EGLINTON	REVISED	WORSHIPPERS	CENTRALLY
GOA'ULD	OPERATOR	EDGING	LEAVENED	RITSUKO	INDONESIA
COLLATION	OPERATOR	FRG	PANDIONIDAE	LIFELESS	MONEO
BACHA	W.J.	NAMSOS	SHIRT	MAHAN	NILGRIS

neighboring words are **not** semantically related

Neural Network Language Model



Expensive to train:
- 3-4 weeks on Wikipedia



Quick to train:
- 40 min. on Wikipedia
- tricks:
  - parallelism
  - avoid synchronization

Word Embeddings

FRANCE 454	JESUS 1973	XBOX 6909	REDDISH 11724	SCRATCHED 29869	MEGABITS 87025
AUSTRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
BELGIUM	SATI	PLAYSTATION	BLUISH	SMASHED	MB/S
GERMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
ITALY	SATAN	IPOD	PURPLISH	POPPED	BAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
SWEDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
NORWAY	VISHNU	HD	GRAYISH	SCREWED	MEGAHERTZ
EUROPE	ANANDA	DREAMCAST	WHITISH	SECTIONED	MEGAPIXELS
HUNGARY	PARVATI	GEFORCE	SILVERY	SLASHED	GBIT/S
SWITZERLAND	GRACE	CAPCOM	YELLOWISH	RIPPED	AMPERES

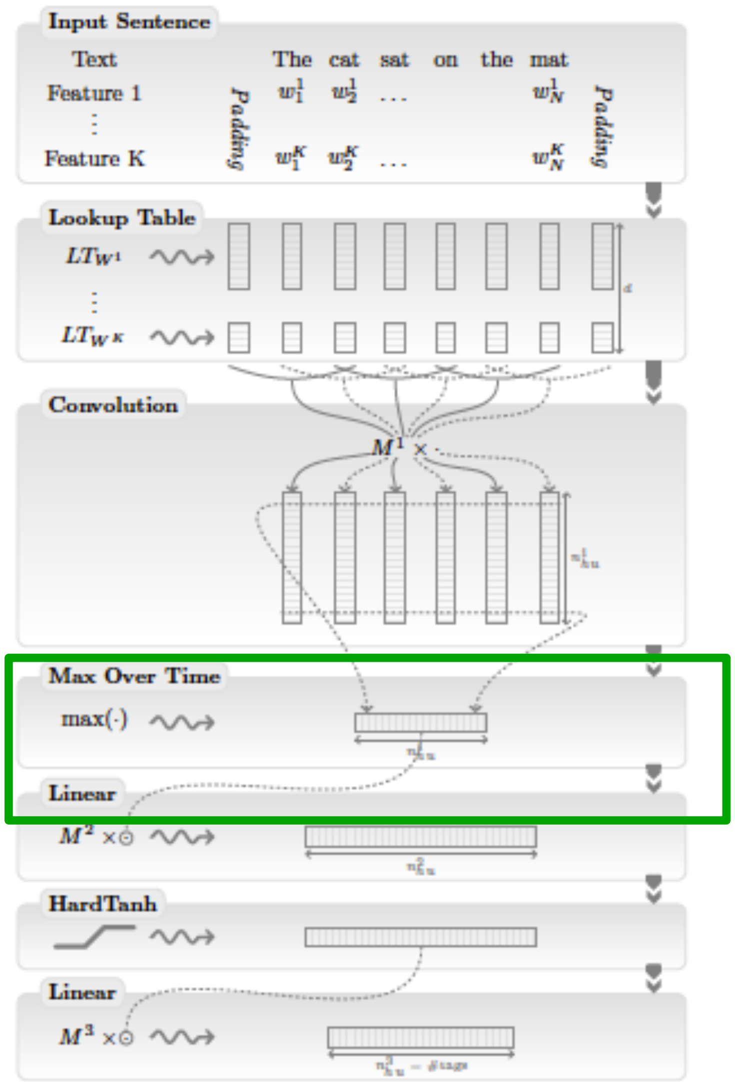
neighboring words **are** semantically related

Deep Learning for NLP

Convolutional Network

Convolution over whole sentence

$$\left[f_\theta^l\right]_i = \max_t \left[f_\theta^{l-1}\right]_{i,t} \qquad 1 \leq i \leq n_{hu}^{l-1}$$



demo

Parser Online Demo

Question?

- Do we need multiple tasks, aka NLP pipelines?
- Some tasks are artificial:
  - e.g are POS tags useful
- End-to-end training allows avoiding splitting into tasks
- Let the layer learn abstract representation
- Example:
  - dependency parsing with clusters of words performs similarly to using POS

Question

- Do we need linguists at all?
- Children learn to talk with no linguistic training

- This is what Manning calls

The tsunami of DL over NLP

The tsunami

- High percent of papers at ACL 2015 using DL
- Neil Lawrence:
  - NLP is kind of like a rabbit in the headlights of the Deep Learning machine, waiting to be flattened.
- Geoff Hinton:
  - In a few years time we will put DL on a chip that fits into someone’s ear is just like a real Babel fish
- Michael Jordan:
  - “I’d use the billion dollars to build a NASA-size program focusing on natural language processing, in all of its glory (semantics, pragmatics, etc.).”

But…

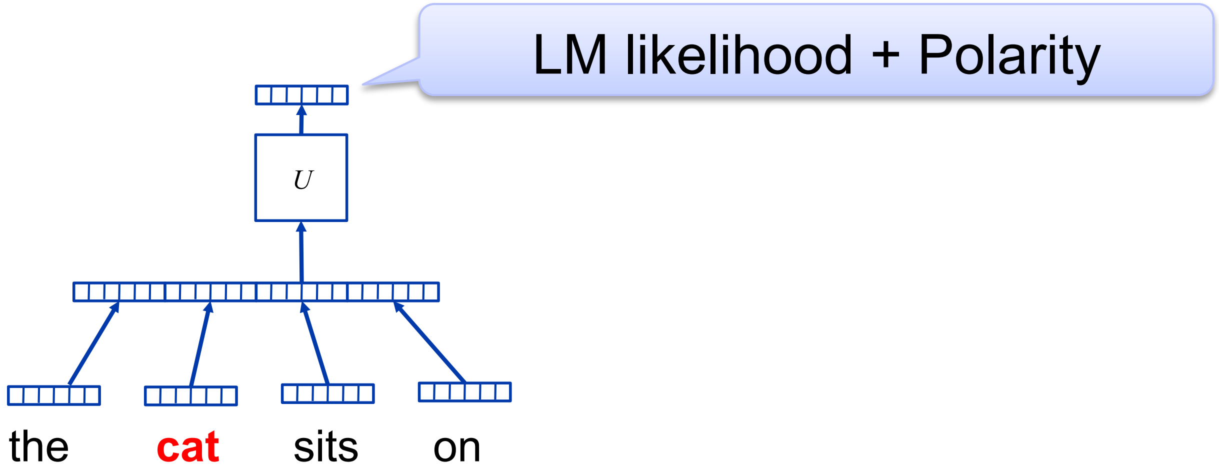
- Isn’t fascinating that we can develop systems that can deal with disparate tasks using a unified learning architecture?
- After all, children learn to speak in 3 years without much study

Biological Inspiration

- This is **NOT** how humans learn
- Humans first learn simple concepts, and then learner more complex ideas by combining simpler concepts
- There is evidence though that the cortex has a **single learning algorithm**:
  - Inputs from optic nerves of ferrets was rerouted to into their audio cortex
  - They were able to learn to see with their audio cortex instead
- If we want a general learning algorithm, it needs to be able to:
  - Work with any type of data
  - Extract it's own features
  - Transfer what it's learned to new domains
  - Perform multi-modal learning – simultaneously learn from multiple different inputs (vision, language, etc)

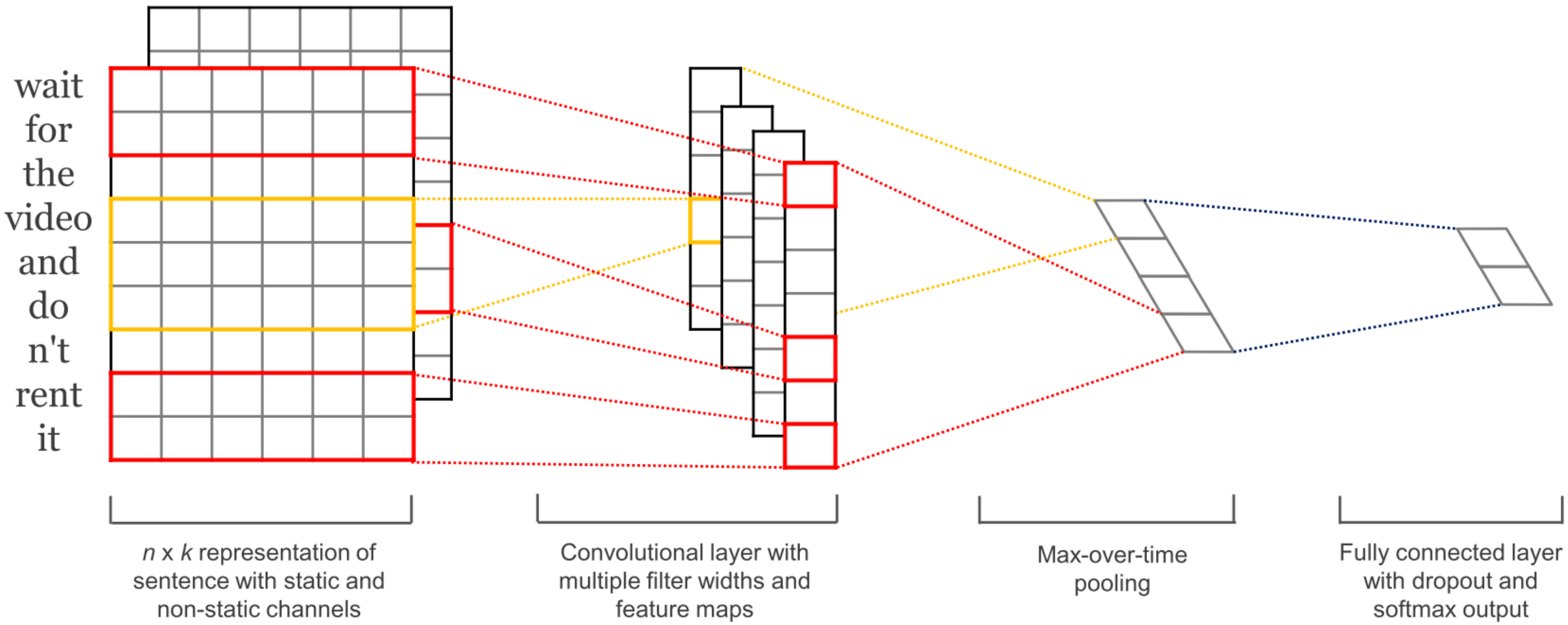
Discriminative Word Embeddings

- Sentiment Specific Word Embeddings



- Uses an annotated corpus with polarities (e.g. tweets)
- SS Word Embeddings achieve SOTA accuracy on tweet sentiment classification

Deep Convolutional Network



Semeval 2015 Sentiment on Tweets

Team	Phrase Level Polarity	Tweet
Attardi (unofficial)		**67.28**
Moschitti	**84.79**	64.59
KLUEless	84.51	61.20
IOA	82.76	62.62
WarwickDCS	82.46	57.62
Webis		64.84

Social Sensing

- Detecting reports of natural disasters (e.g. floods, earthquakes) on Twitter

System	Precision	Recall	F-1
Baseline	86.87	70.96	78.11
Discrim. Embeddings	85.94	75.05	80.12
Convolutional	96.65	95.52	**96.08**

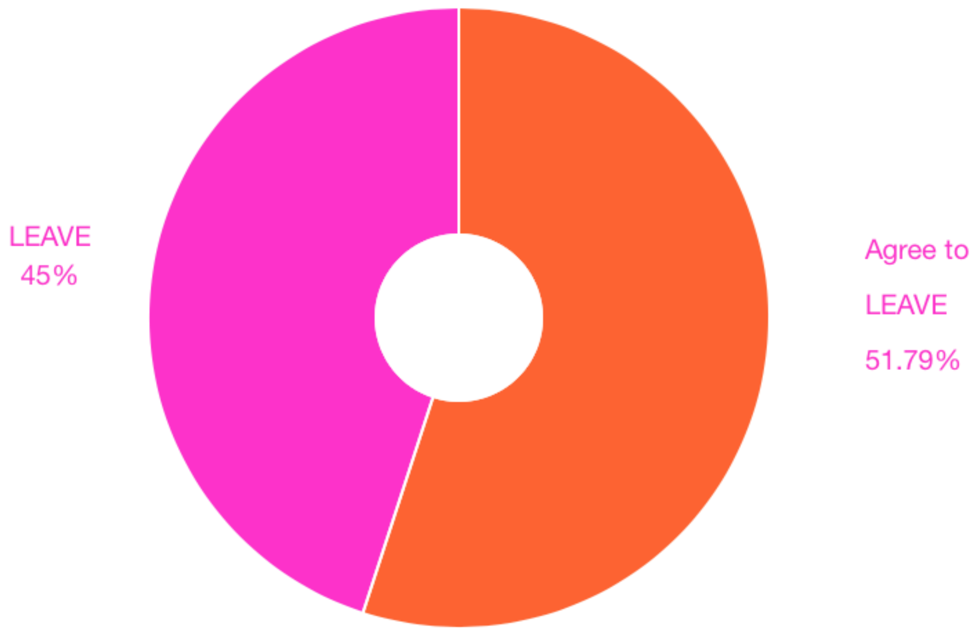
Sentiment Analysis

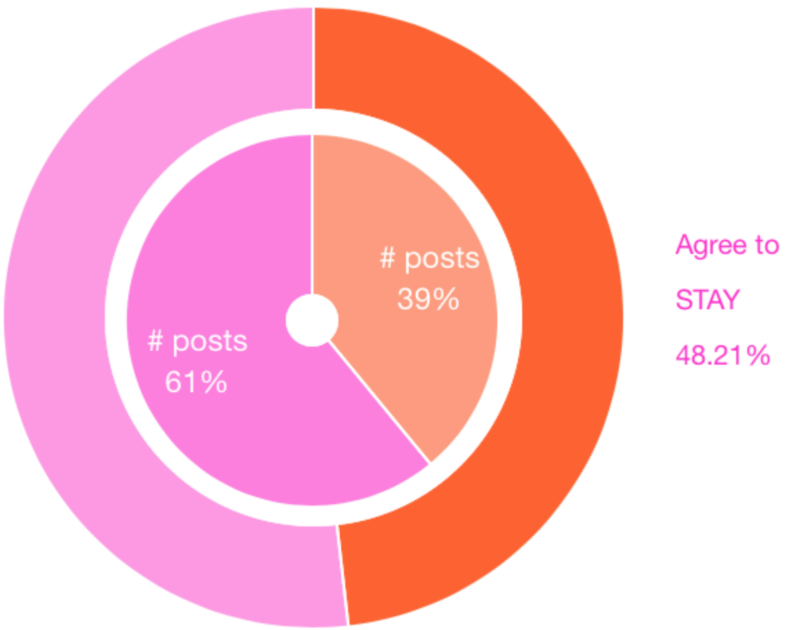
WebSays + Tiscali

17/1/2013

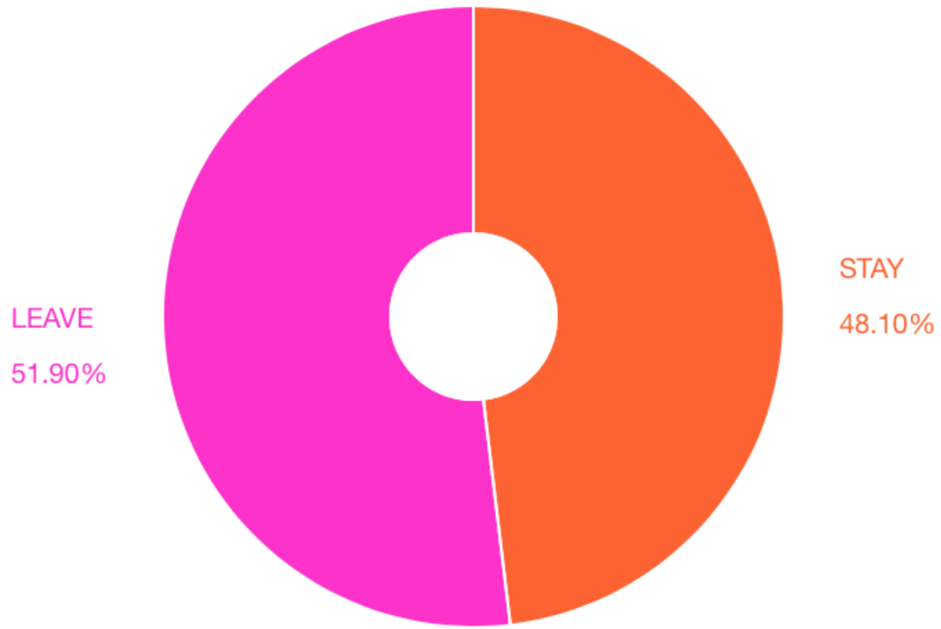
Brexit

Exit Polls	Predictions





Brexit - Results

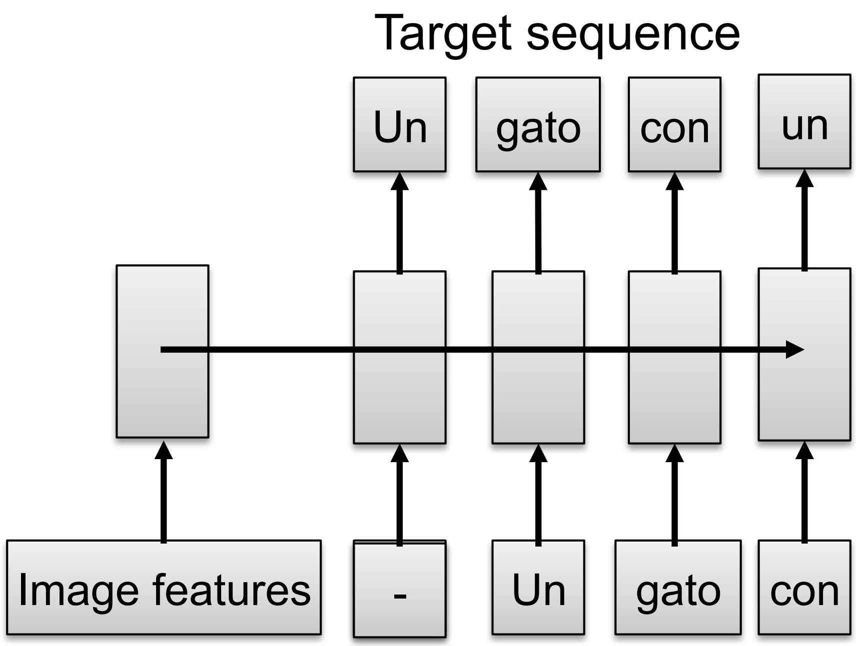


Security

- Symantec uses DL for identifying and defending against zero-day malware attacks
- CIA uses a tool capable of predicting the arising of riots 3 days in advance

Image Captioning

- Extract features from images with CNN
- Input to LSTM
- Trained on MSCOCO
  - 300k images, 6 caption/image



"little girl is eating piece of cake."

"baseball player is throwing ball in game."

"woman is holding bunch of bananas."

"a young boy is holding a baseball bat."

"a cat is sitting on a couch with a remote control."

"a woman holding a teddy bear in front of a mirror."

Semtence Compression

- Alan Turing, known as the father of computer science, the codebreaker that helped win World War 2, and the man tortured by the state for being gay, is given a pardon nearly 60 years after his death.
- *Alan Turing is given a pardon*.

- Gwyneth Paltrow and her husband Chris Martin, are to separate after more than 10 years of marriage.
- *Gwyneth Paltrow* ***are*** *to separate.*

Natural Language Inference

Reasoning in Question Answering

- Reasoning is essential in a QA task
- Traditional approach: **rule-based** reasoning
  - Mapping natural languages to logic form
  - Inference over logic forms

not easy

- **Dichotomy:**
  - ML for NL analysis
  - symbolic reasoning for QA
- **DL perspective**:
  - distributional representation of sentences
  - remember facts from the past
  - … so that it can suitably deal with **long-term dependencies**

Episodes

- From Facebook BaBI data set:

  I: Jane went to the hallway

  I: Mary walked to the bathroom

  I: Sandra went to the garden

  I: Sandra took the milk there

  Q: Where is the milk?

  A: garden

Tasks

- Path Finding:

I: The bathroom is south of bedroom

I: The bedroom is east of kitchen

Q: How do you go from bathroom to kitchen?

A: north, west

- Positional Reasoning:

I: The triangle is above the rectangle

I: The square is to the left of the triangle

Q: Is the rectangle to the right of the square?

A: Yes

Neural Reasoner

- Layered architecture for dealing with complex logic relations in reasoning:
  - One encoding layer
  - Multiple reasoning layers
  - Answer layer (either chooses answer, or generates answer sentence)
- Interaction between question and facts representations models the reasoning

Quiz Bowl Competition

- Iyyer et al. 2014: A Neural Network for Factoid Question Answering over Paragraphs
- QUESTION:

  He left unfinished a novel whose title character forges his father's signature to get out of school and avoids the draft by feigning desire to join.

  One of his novels features the jesuit Naptha and his opponent Settembrini, while his most famous work depicts the aging writer Gustav von Aschenbach.

  Name this German author of The Magic Mountain and Death in Venice.
- ANSWER: Thomas Mann

QANTA vs Ken Jennings

- QUESTION:

  Along with Evangelista Torricelli, this man is the namesake of a point the minimizes the distances to the vertices of a triangle.

  He developed a factorization method …

  ANSWER: Fermat

- QUESTION:

  A movie by this director contains several scenes set in the Yoshiwara Nightclub.

  In a movie by this director a man is recognized by a blind beggar because he is wistlin
  In the hall of the mountain king.

  ANSWER: Fritz Lang

Conclusions

- DL is having huge impacts in many areas of AI
- Ability to process Big Data with parallel hardware crucial
- Embrace or avoid?