Security and Fairness of Deep Learning

Course Overview

Spring 2020

Today

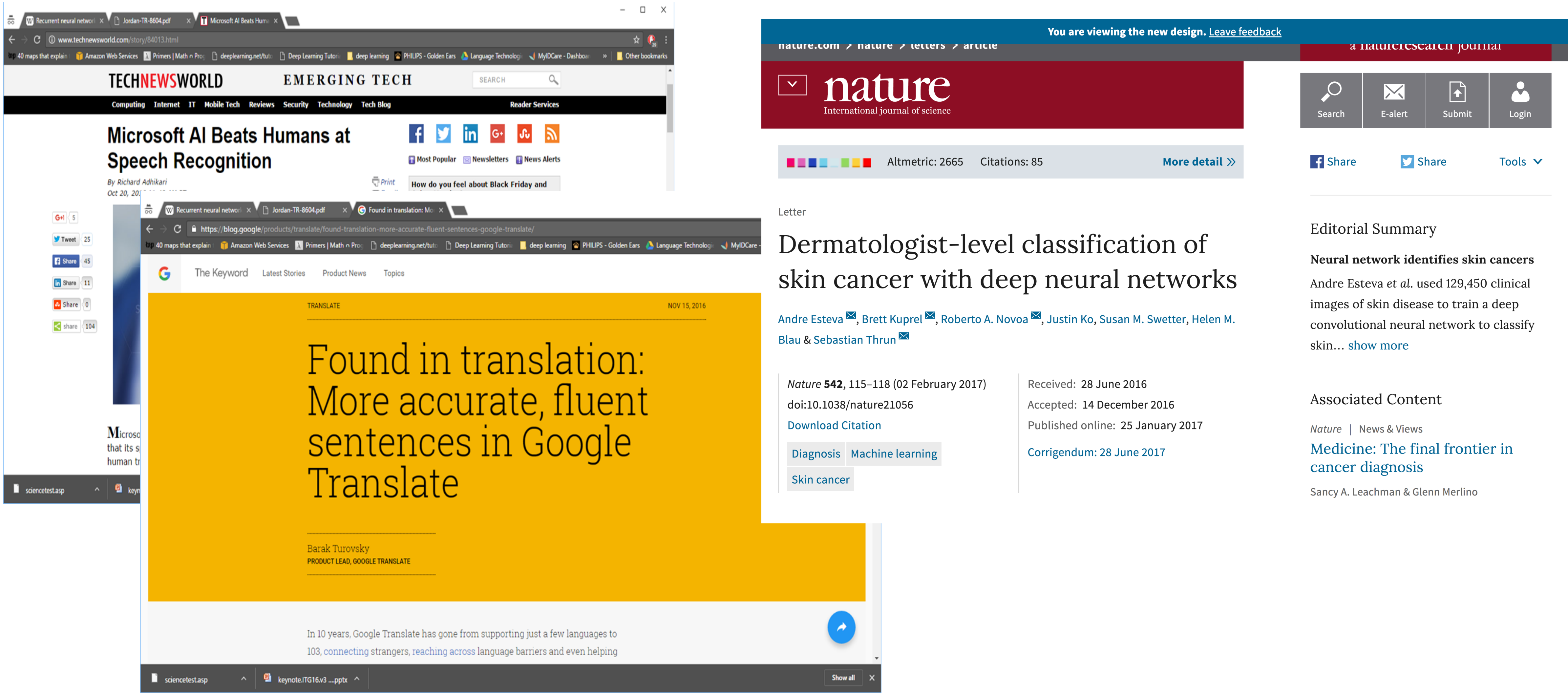
- Goals
- Modules

- Prerequisites
- Logistics
- Grading
- Policies

Course staff

- Instructor: Piotr (Peter) Mardziel
  - Email: piotrm@cmu.edu
  - Office hours: Thursdays 12-1pm Pacific
  - Office: B23 114
- TA: Zifan Wang
  - Email: zifan.wang@sv.cmu.edu
  - Office hours: Wednesdays 2:30-3:30pm Eastern
  - Office: CIC 2206
- TA: Caleb Lu
  - Email: kaijil@andrew.cmu.edu
  - Office hours: TBD
  - Office: TBD
- Office hours available remotely:
  - Zoom meeting links on website

Continuing successes of deep learning



TECHNEWSWORLD
EMERGING TECH
SEARCH
Computing Internet IT Mobile Tech Reviews Security Technology Tech Blog Reader Services
Most Popular Newsletters News Alerts
Print
How do you feel about Black Friday and

Microsoft AI Beats Humans at Speech Recognition

By Richard Adhikari
Oct 20, 20

The Keyword Latest Stories Product News Topics
TRANSLATE
NOV 15, 2016

Found in translation: More accurate, fluent sentences in Google Translate

Barak Turovsky
PRODUCT LEAD, GOOGLE TRANSLATE

In 10 years, Google Translate has gone from supporting just a few languages to 103, connecting strangers, reaching across language barriers and even helping

You are viewing the new design. Leave feedback
nature.com > nature > letters > article
nature
International journal of science
Search E-alert Submit Login
Share Share Tools

Altmetric: 2665 Citations: 85 More detail »

Letter

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun

Nature **542**, 115–118 (02 February 2017)
doi:10.1038/nature21056
Download Citation

Diagnosis Machine learning Skin cancer

Received: 28 June 2016
Accepted: 14 December 2016
Published online: 25 January 2017
Corrigendum: 28 June 2017

Editorial Summary

Neural network identifies skin cancers

Andre Esteva *et al.* used 129,450 clinical images of skin disease to train a deep convolutional neural network to classify skin… show more

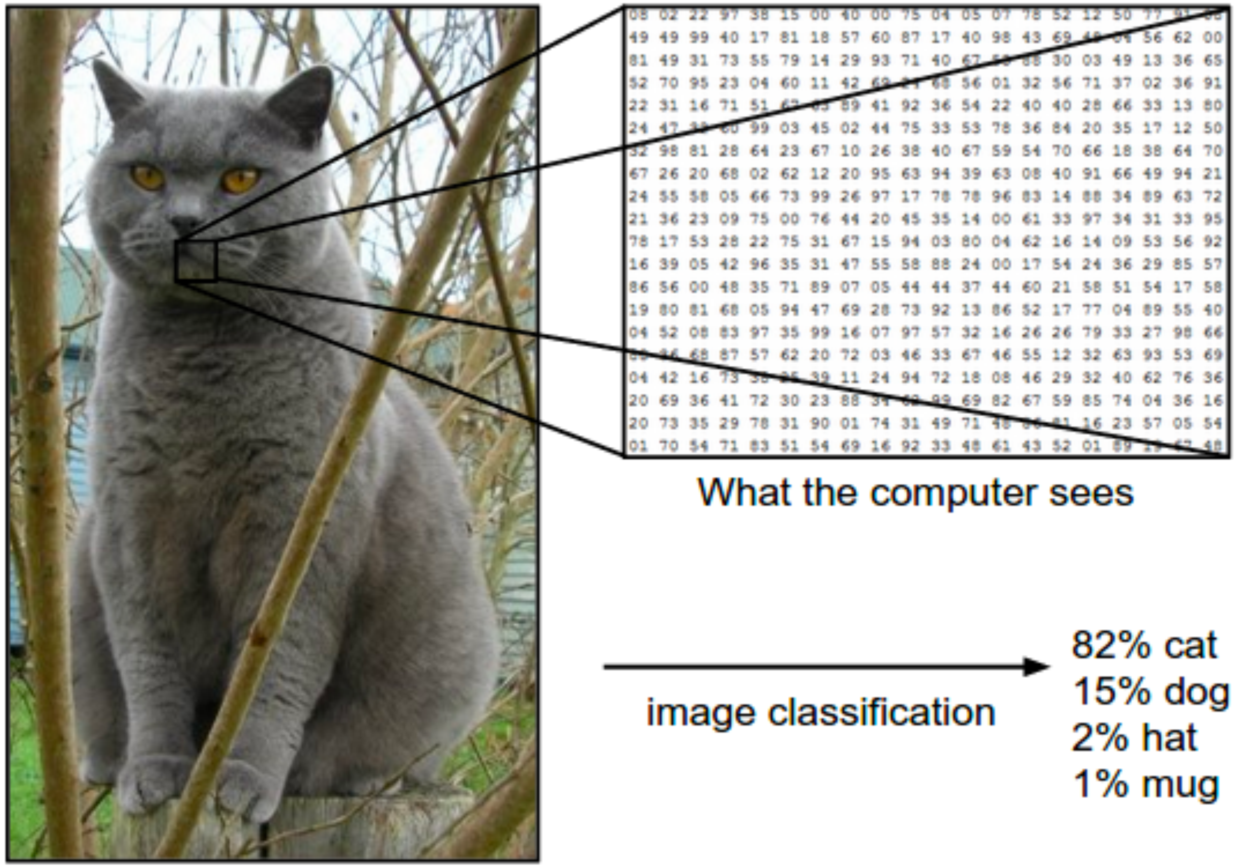
Associated Content

Nature | News & Views

Medicine: The final frontier in cancer diagnosis

Sancy A. Leachman & Glenn Merlino

Image classification



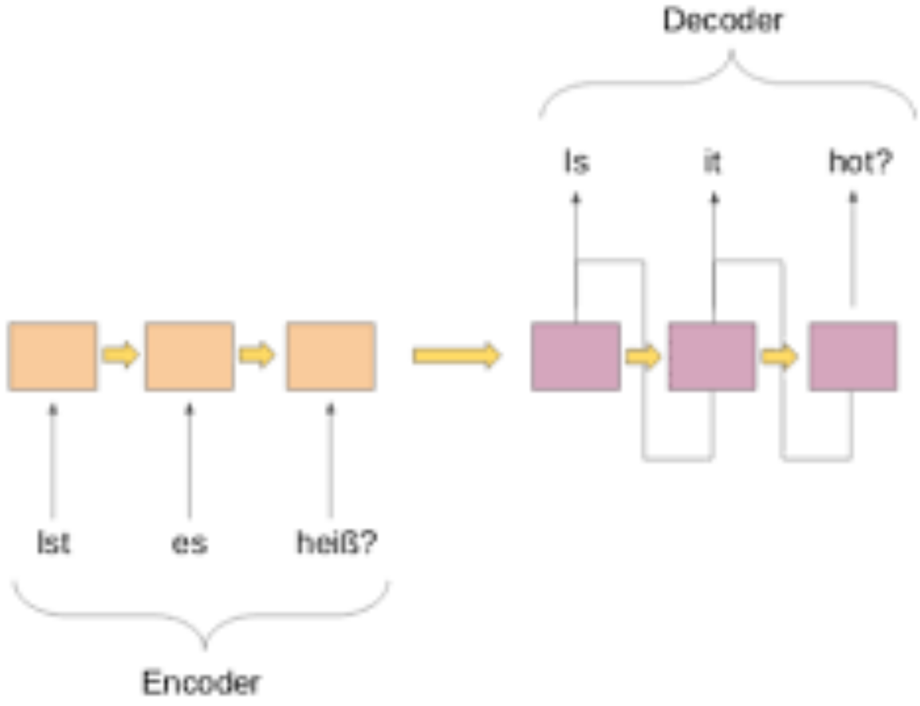
08	02	22	97	38	15	00	40	00	75	04	05	07	78	52	12	50	77	91	08
49	49	99	40	17	81	18	57	60	87	17	40	98	43	69	48	04	56	62	00
81	49	31	73	55	79	14	29	93	71	40	67	53	88	30	03	49	13	36	65
52	70	95	23	04	60	11	42	69	24	68	56	01	32	56	71	37	02	36	91
22	31	16	71	51	67	63	89	41	92	36	54	22	40	40	28	66	33	13	80
24	47	32	60	99	03	45	02	44	75	33	53	78	36	84	20	35	17	12	50
32	98	81	28	64	23	67	10	26	38	40	67	59	54	70	66	18	38	64	70
67	26	20	68	02	62	12	20	95	63	94	39	63	08	40	91	66	49	94	21
24	55	58	05	66	73	99	26	97	17	78	78	96	83	14	88	34	89	63	72
21	36	23	09	75	00	76	44	20	45	35	14	00	61	33	97	34	31	33	95
78	17	53	28	22	75	31	67	15	94	03	80	04	62	16	14	09	53	56	92
16	39	05	42	96	35	31	47	55	58	88	24	00	17	54	24	36	29	85	57
86	56	00	48	35	71	89	07	05	44	44	37	44	60	21	58	51	54	17	58
19	80	81	68	05	94	47	69	28	73	92	13	86	52	17	77	04	89	55	40
04	52	08	83	97	35	99	16	07	97	57	32	16	26	26	79	33	27	98	66
88	36	68	87	57	62	20	72	03	46	33	67	46	55	12	32	63	93	53	69
04	42	16	73	38	25	39	11	24	94	72	18	08	46	29	32	40	62	76	36
20	69	36	41	72	30	23	88	34	62	99	69	82	67	59	85	74	04	36	16
20	73	35	29	78	31	90	01	74	31	49	71	48	86	81	16	23	57	05	54
01	70	54	71	83	51	54	69	16	92	33	48	61	43	52	01	89	19	67	48

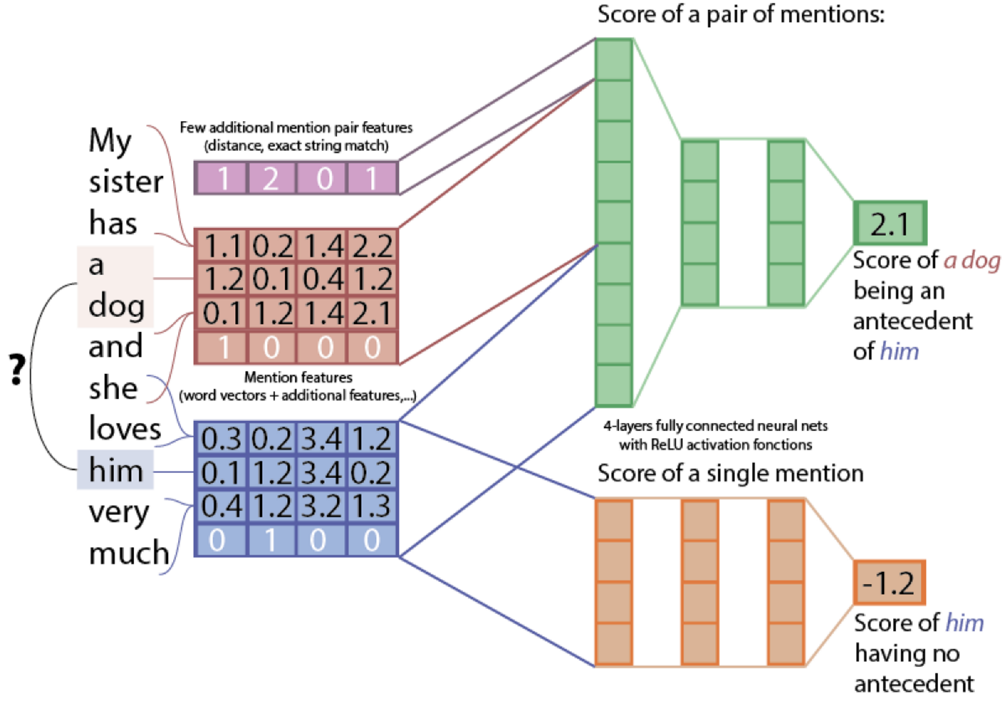
What the computer sees

image classification

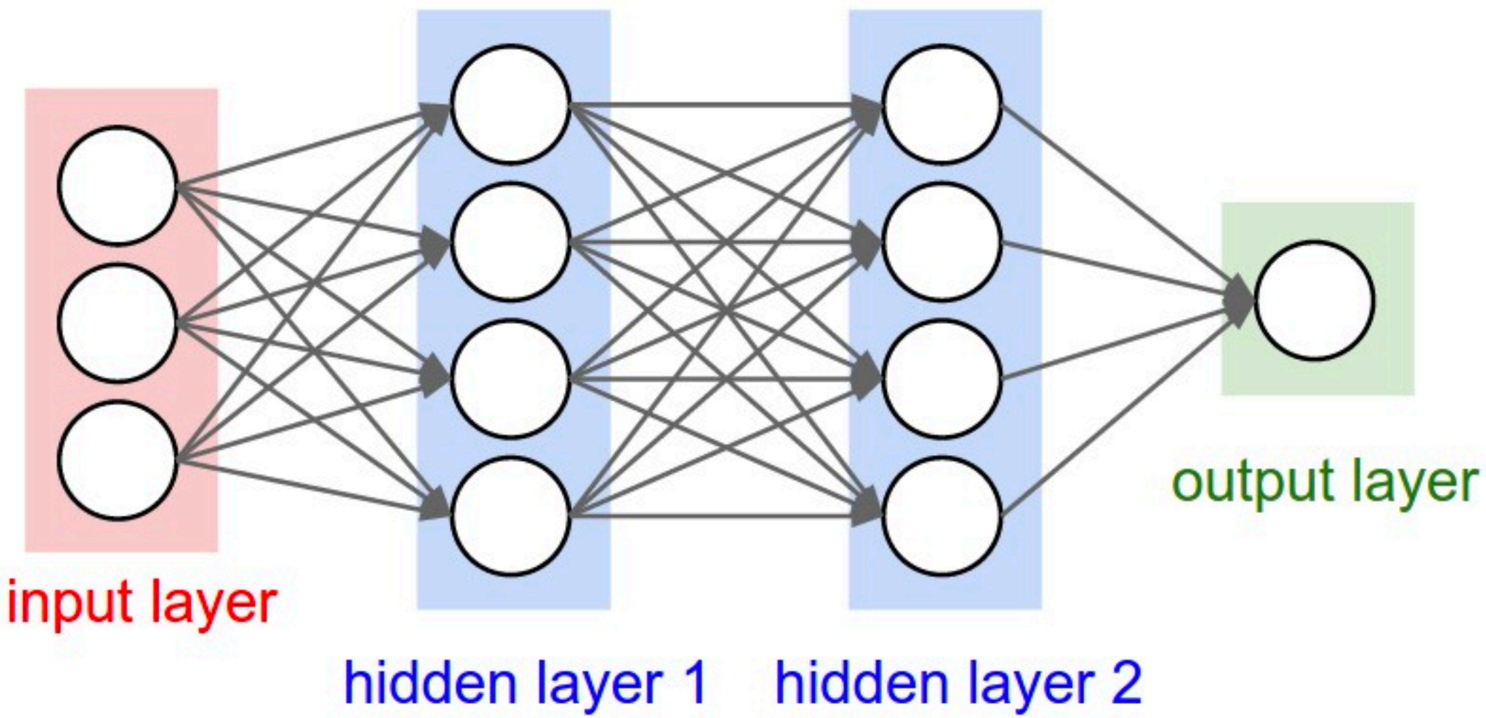
82% cat
15% dog
2% hat
1% mug

NLP: translation, etc.




Score of a pair of mentions:
My
sister
has
a
dog
and
she
loves
him
very
much
?
Few additional mention pair features
(distance, exact string match)
1 2 0 1
1.1 0.2 1.4 2.2
1.2 0.1 0.4 1.2
0.1 1.2 1.4 2.1
1 0 0 0
Mention features
(word vectors + additional features,...)
0.3 0.2 3.4 1.2
0.1 1.2 3.4 0.2
0.4 1.2 3.2 1.3
0 1 0 0
2.1
Score of a dog being an antecedent of him
4-layers fully connected neural nets
with ReLU activation fonctions
Score of a single mention
-1.2
Score of him having no antecedent

Deep neural networks learn representations



Deeper layers learn progressively more abstract representations:
pixels, edges, motifs, parts of objects, objects

Enabling trends

- Large volumes of training data
- Computation power
  - GPUs,…

Course objective

Understand deeply how and why deep networks work and their weaknesses

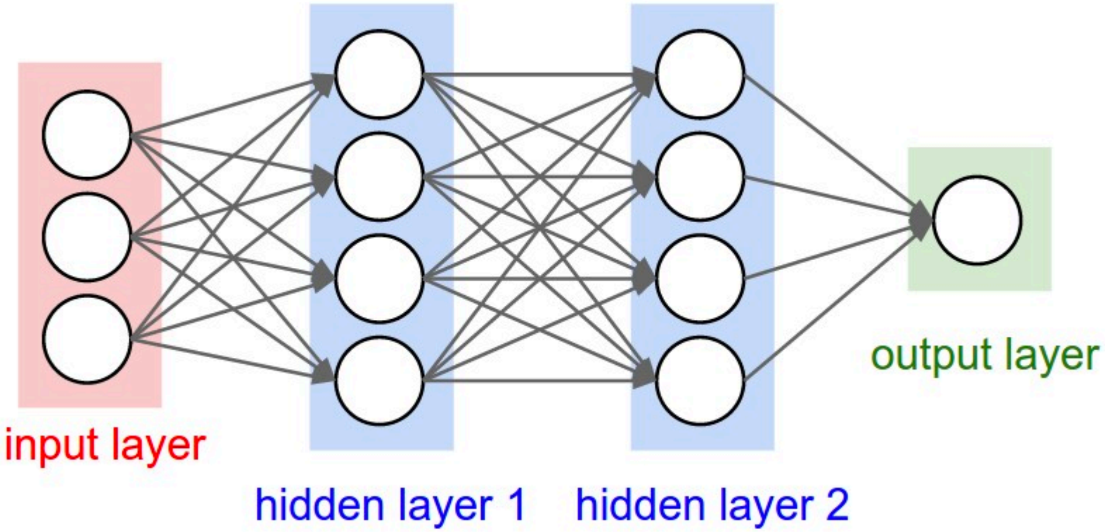
Become informed: what can go wrong (other than poor performance)?

Course modules

1. Fundamentals of deep learning
2. Explanations for deep learning
3. Security of deep learning
4. Privacy and Fairness in deep learning

Course modules

1. Fundamentals of deep learning
   - Background on machine learning
   - Architectures, training, platforms
   - Focus on convolutional and recurrent neural networks



python

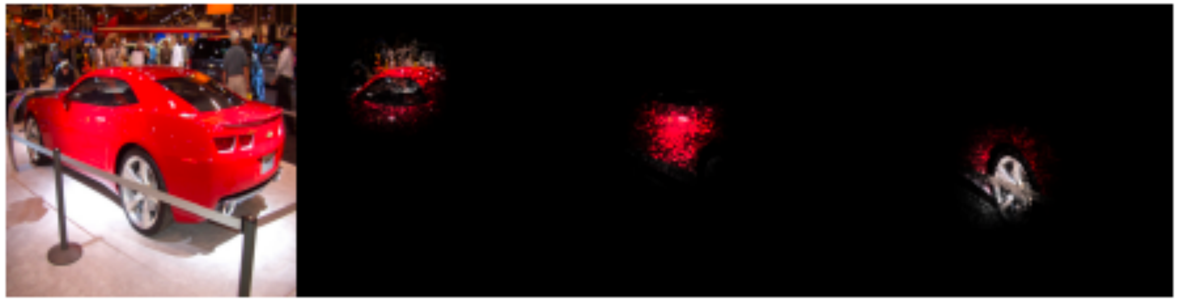
NumPy

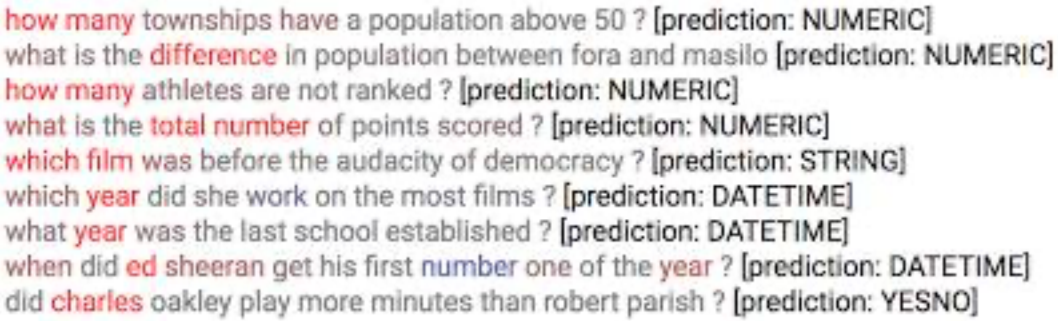
TensorFlow

Keras

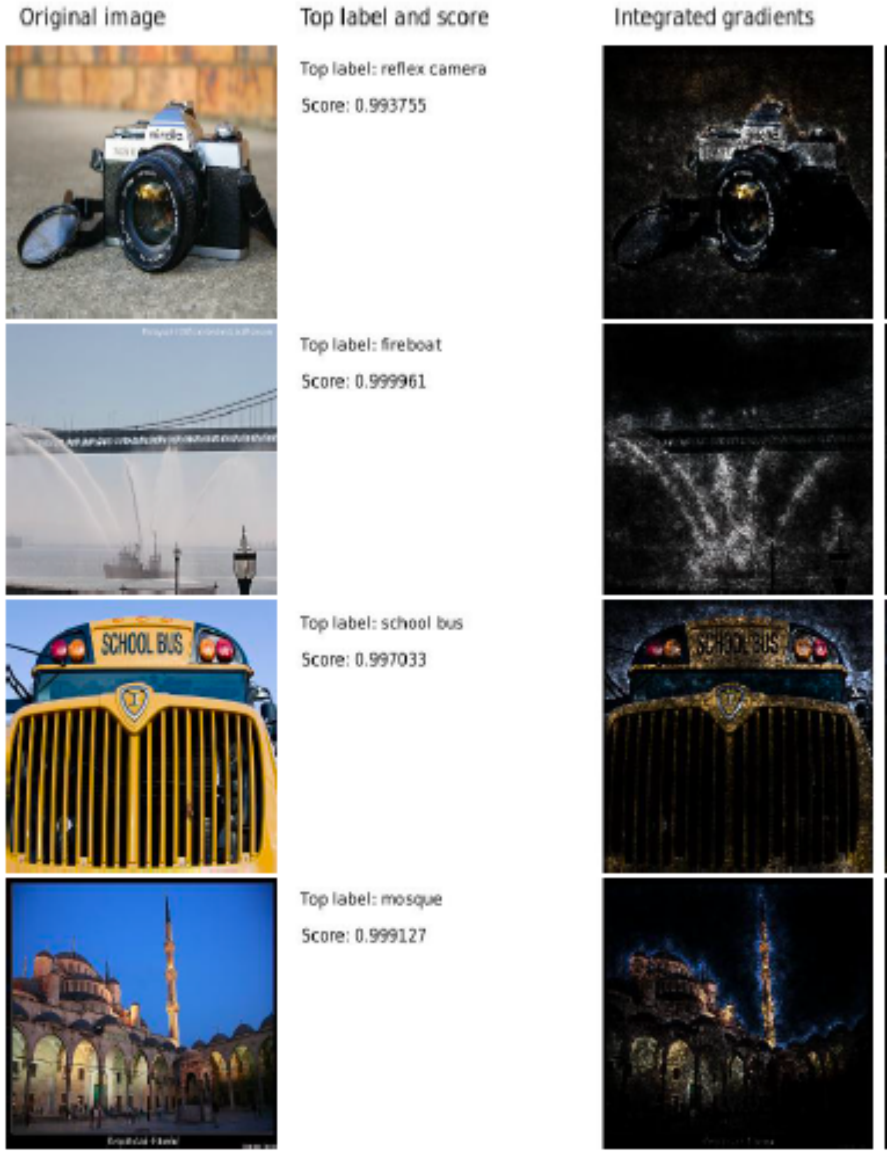
Course modules

2. Explanations for deep learning
   - Feature importance and visualization



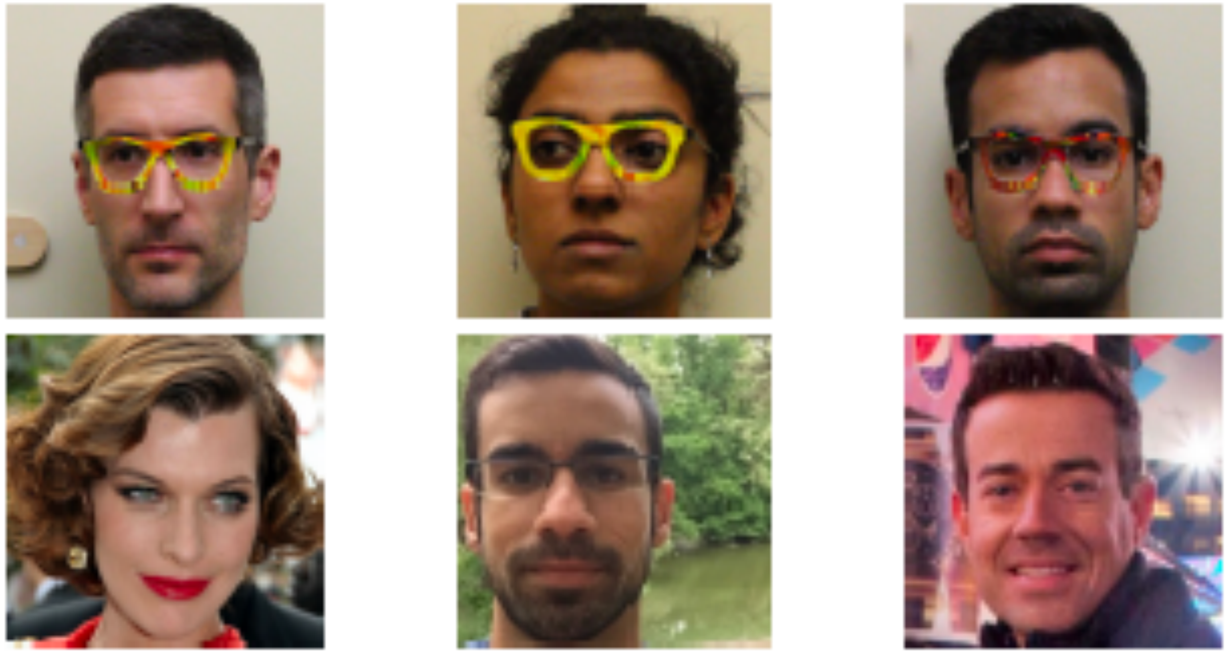


how many townships have a population above 50 ? [prediction: NUMERIC]
what is the difference in population between fora and masilo [prediction: NUMERIC]
how many athletes are not ranked ? [prediction: NUMERIC]
what is the total number of points scored ? [prediction: NUMERIC]
which film was before the audacity of democracy ? [prediction: STRING]
which year did she work on the most films ? [prediction: DATETIME]
what year was the last school established ? [prediction: DATETIME]
when did ed sheeran get his first number one of the year ? [prediction: DATETIME]
did charles oakley play more minutes than robert parish ? [prediction: YESNO]



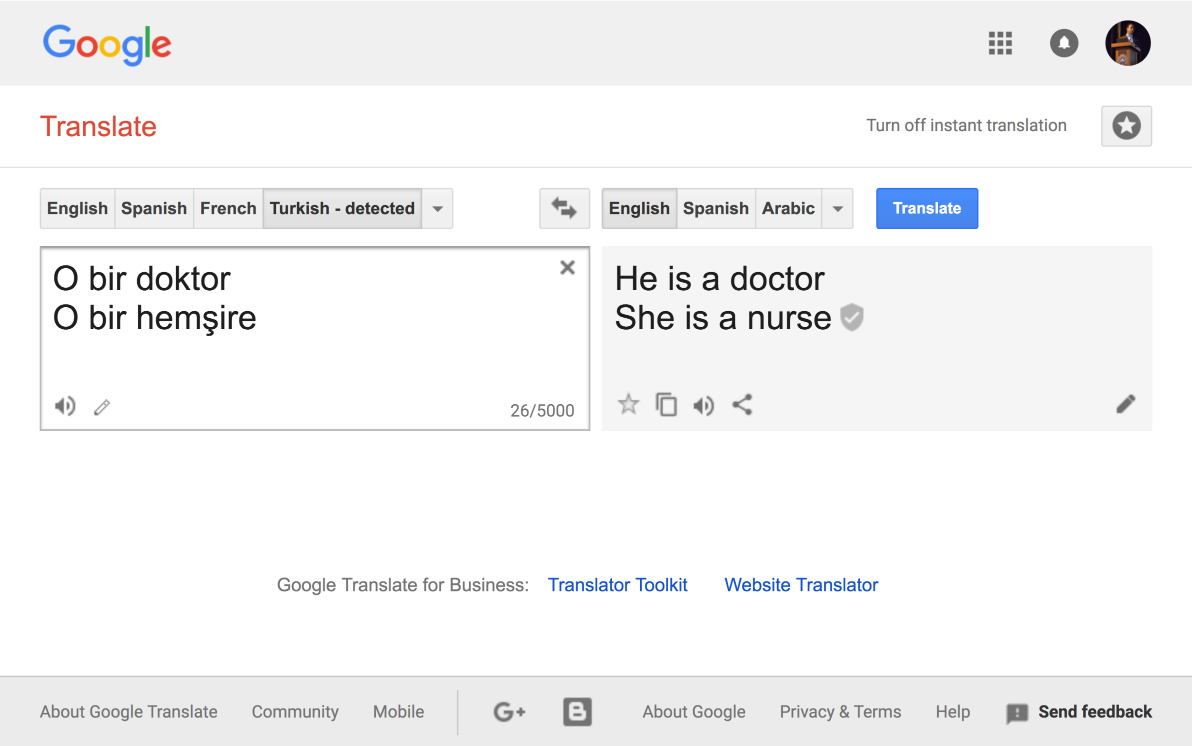
Course modules

3. Security of deep learning models
    - Attacks on classifiers and defenses



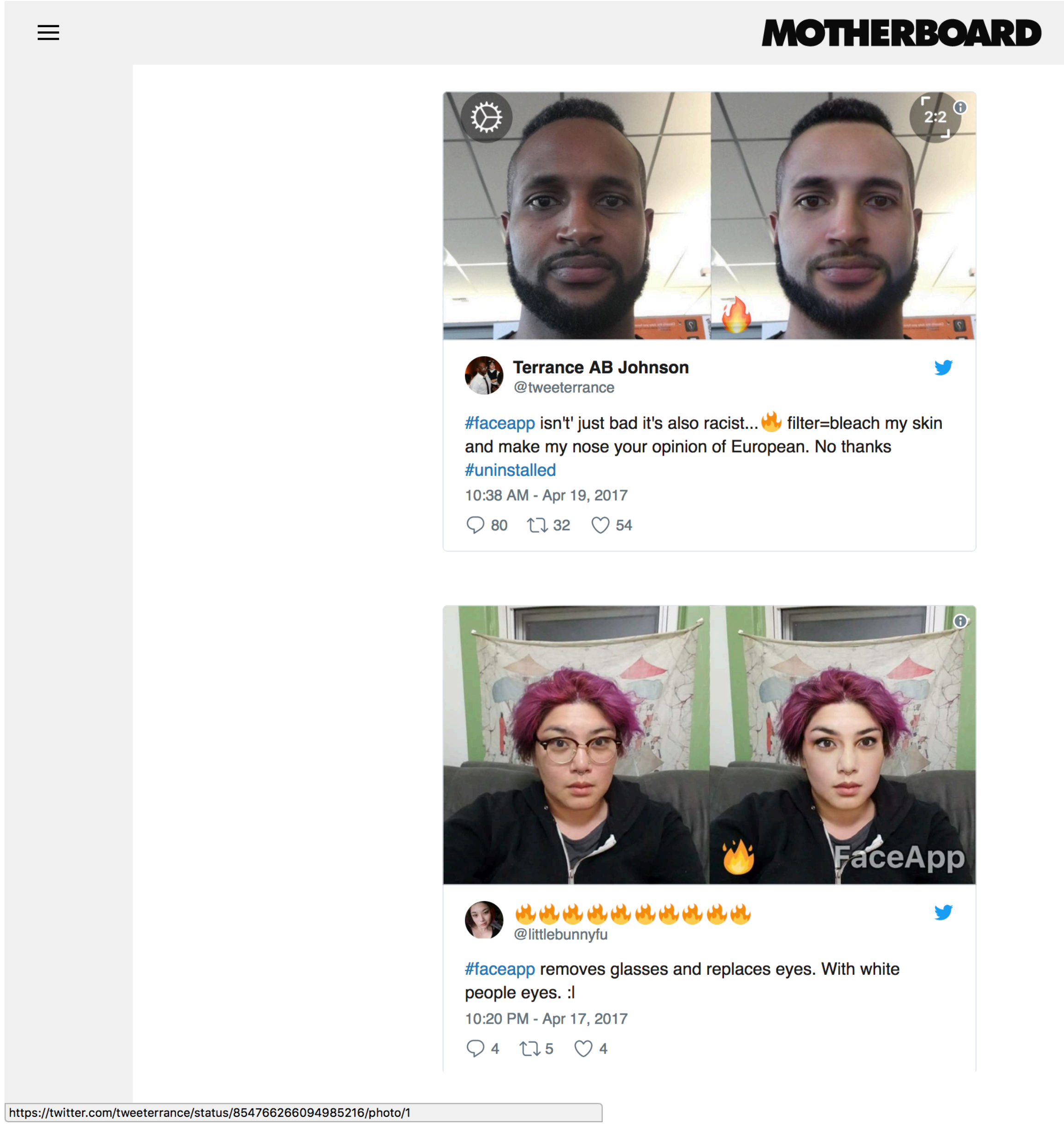
Course modules

4. Privacy and Fairness in deep learning
    - Inferring sensitive information
    - Bias and de-biasing



O bir doktor
O bir hemşire

He is a doctor
She is a nurse



Terrance AB Johnson
@tweeterrance

#faceapp isn't' just bad it's also racist...🔥 filter=bleach my skin and make my nose your opinion of European. No thanks #uninstalled

10:38 AM - Apr 19, 2017

80 32 54

🔥🔥🔥🔥🔥🔥🔥🔥🔥🔥
@littlebunnyfu

#faceapp removes glasses and replaces eyes. With white people eyes. :|

10:20 PM - Apr 17, 2017

4 5 4

https://twitter.com/tweeterrance/status/854766266094985216/photo/1

Course Format

- Lectures covering the background
  - Stanford CS231n (Convolutional Neural Networks for Visual Recognition)
  - Deep Learning textbook
- 1-2 Lecture covering software tools and setup
  - Numpy, Tensorflow, Keras, Jupyter Notebook, Google Computing Services
- **Lectures covering research papers**
  - Occasionally guest lecturers

Prerequisites

- No formal prerequisites

- Basics of linear algebra, probability, multivariate calculus
  - Will review briefly in class and provide resources to learn on your own
  - Roughly Chapters 1-5 of Deep Learning textbook by Goodfellow et al.

- Familiarity with Python
  - Necessary for programming homework

- Quick class poll

Logistics

- Lectures: Tue & Thur, 10:30-11:50am Pacific / 1:30-2:50pm Eastern
- Web page: http://www.ece.cmu.edu/~ece739/

- Gradescope (assignment submission)
- Canvas (grades)
- Piazza (announcements, for all other communication)

- Textbook
  - Deep Learning textbook by Goodfellow, Bengio, Courville

Grading

- Homework: 90%
  - 5 x 18%
- Class participation: 10%
  - Be present and engaged in class and piazza
  - Informed questions for guest lecturers

Collaboration policy on homework

- You are allowed/encouraged to discuss homework problems with other students in the class but are required to write out solutions independently and to acknowledge any collaboration or other source. If you are unsure about something, consult the course staff.

CMU Computing Policy

CMU Policy on Cheating

Acknowledgment

- Based on material from
  - Spring 2019 Course