

Security and Fairness of Deep Learning

# Course Overview

Anupam Datta

CMU

Spring 2019

# Course staff

- Instructor: Anupam Datta
  - Office: B23 221 (SV)
  - Email: [danupam@cmu.edu](mailto:danupam@cmu.edu)
  - Office hours: Thur 12-1pm Pacific
  - Google hangouts: link on Piazza
  
- TA: Klas Leino
  - Office: GHC 7004 (Pittsburgh)
  - Email: [kleino@andrew.cmu.edu](mailto:kleino@andrew.cmu.edu)
  - Office hours: TBA
  - Google hangouts: link on Piazza



# Recent successes of deep learning

The image shows two overlapping browser windows. The top window displays a TechNewsWorld article titled "Microsoft AI Beats Humans at Speech Recognition" by Richard Adhikari, dated Oct 20, 2016. The article is categorized under "EMERGING TECH" and includes social media sharing options for Facebook, Twitter, LinkedIn, Google+, and RSS. The bottom window shows a Google Translate blog post titled "Found in translation: More accurate, fluent sentences in Google Translate" by Barak Turovsky, Product Lead at Google Translate, dated Nov 15, 2016. The blog post features a large yellow background with the title in white text. Below the title, it mentions that in 10 years, Google Translate has supported 103 languages and helped connect people across language barriers.

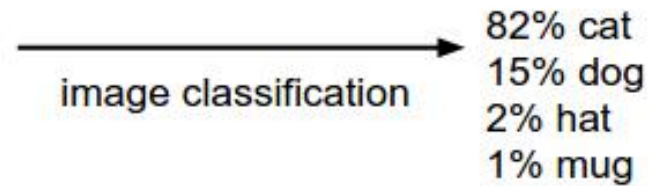
The image shows a screenshot of a Nature journal article page. The article is titled "Dermatologist-level classification of skin cancer with deep neural networks" and is categorized as a "Letter". The authors listed are Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. The article was published in Nature 542, pages 115-118, on 02 February 2017. The DOI is 10.1038/nature21056. The article was received on 28 June 2016, accepted on 14 December 2016, and published online on 25 January 2017. A corrigendum was published on 28 June 2017. The article has an Altmetric score of 2665 and 85 citations. The page also includes a search bar, a navigation menu, and a sidebar with an "Editorial Summary" section titled "Neural network identifies skin cancers" by Andre Esteva et al., which describes the use of a deep convolutional neural network to classify skin images. The sidebar also includes an "Associated Content" section with a link to "Medicine: The final frontier in cancer diagnosis" by Sancy A. Leachman & Glenn Merlino.

# Image classification

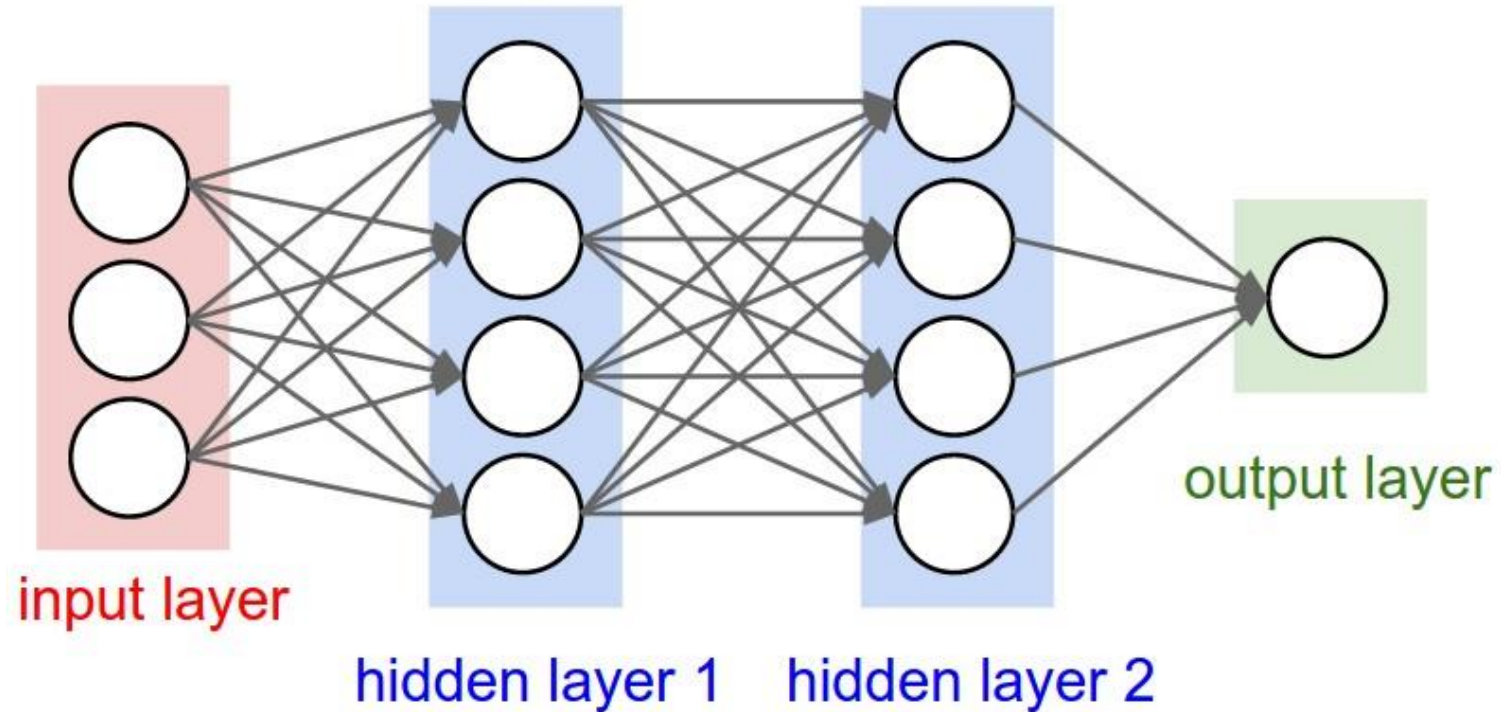


05	02	22	97	38	15	00	40	00	75	04	05	07	78	52	12	50	77	91	20
49	49	99	40	17	81	18	57	60	87	17	40	98	43	69	45	04	56	62	00
81	49	31	73	55	79	14	29	93	71	40	67	57	88	30	03	49	13	36	65
92	70	95	23	04	60	11	42	69	84	68	56	01	32	56	71	37	02	36	91
22	31	16	71	51	67	83	59	41	92	36	54	22	40	40	28	66	33	13	80
24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

What the computer sees



# 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

# Course modules

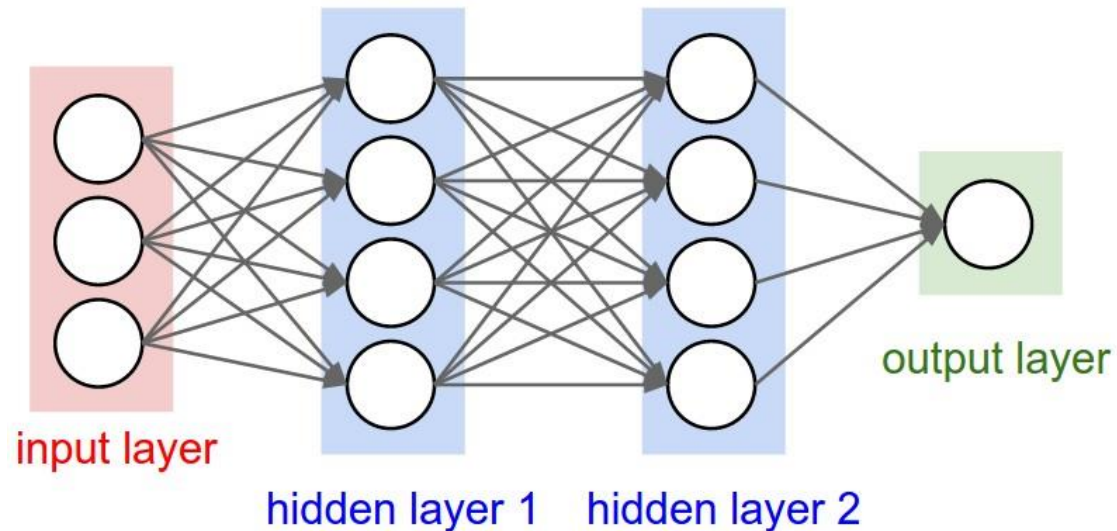
1. Fundamentals of deep networks
2. Unlocking the black box
3. Security of deep learning models
4. Fairness of deep learning



# Course modules

## 1. Fundamentals of deep networks

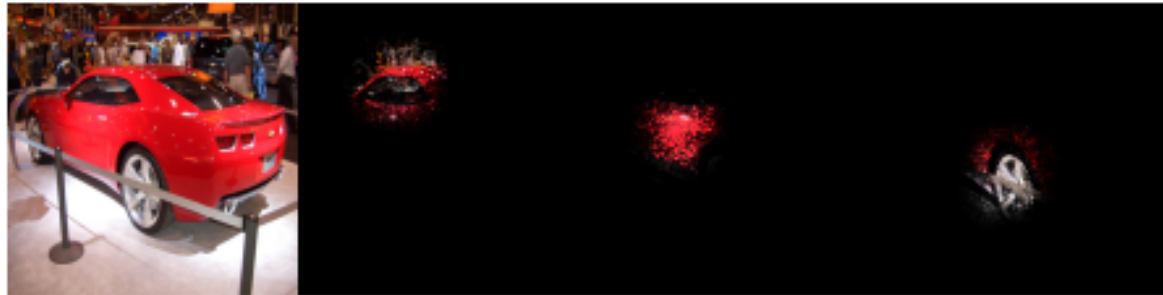
- Background on machine learning
- Architectures, training, platforms
- Focus on convolutional and recurrent neural networks



# Course modules

## 2. Unlocking the black box

- Explaining behavior of deep neural networks



Original image



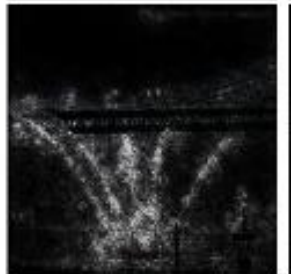
Top label and score

Top label: reflex camera  
Score: 0.993755

Integrated gradients



Top label: fireboat  
Score: 0.999961



Top label: school bus  
Score: 0.997033



Top label: mosque  
Score: 0.999127



# Course modules

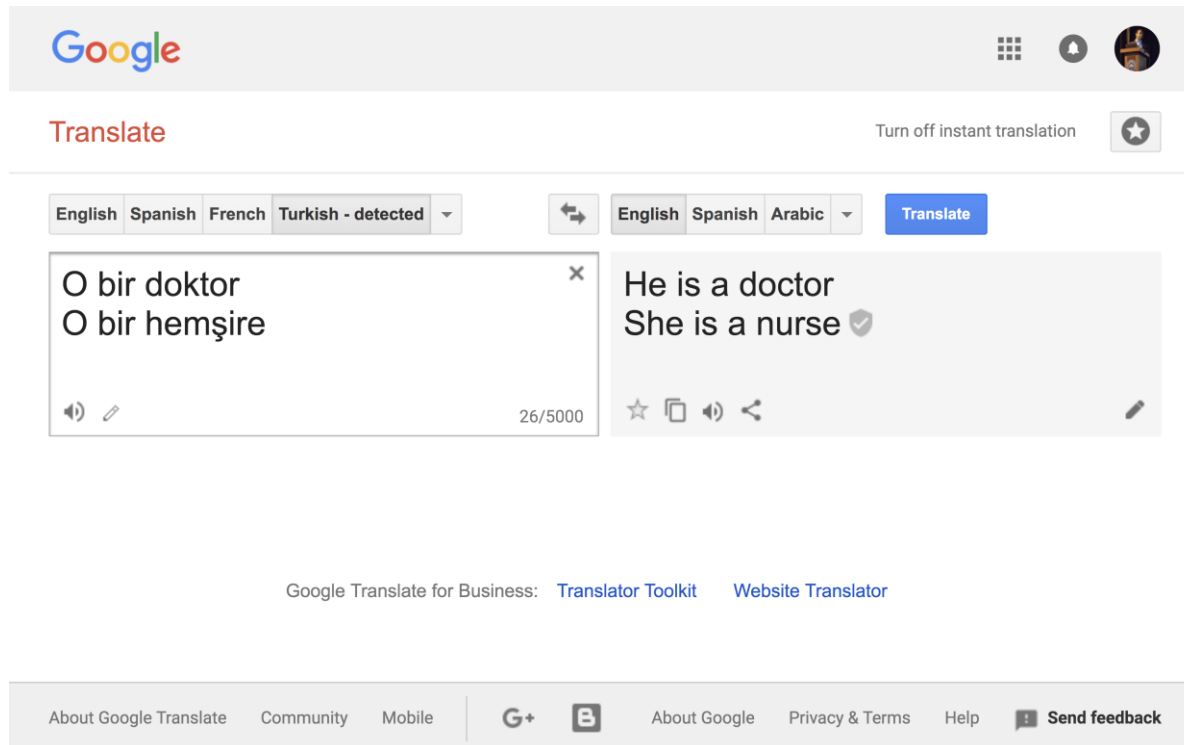
3. Security of deep learning models
  - Attacks on classifiers and defenses



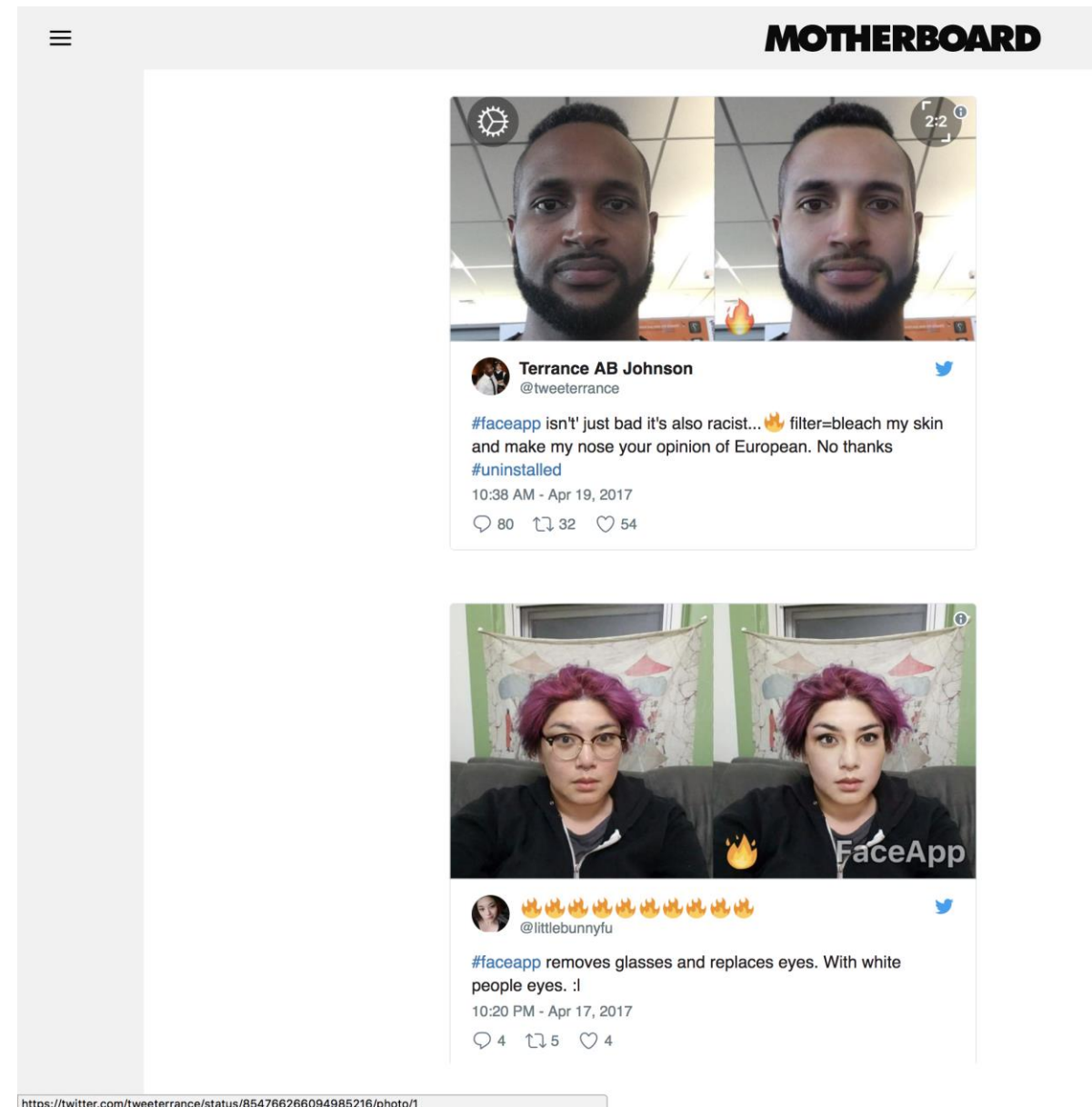
# Course modules

## 4. Fairness of deep learning

- Bias and de-biasing



The screenshot shows the Google Translate web interface. The source text is in Turkish: "O bir doktor" and "O bir hemşire". The target text is in English: "He is a doctor" and "She is a nurse". The interface includes a "Translate" button, a "Turn off instant translation" toggle, and a footer with links for "About Google Translate", "Community", "Mobile", "G+", "B", "About Google", "Privacy & Terms", "Help", and "Send feedback".



The screenshot shows a Twitter thread on the Motherboard website. The top tweet is from Terrance AB Johnson (@tweeterrance) dated 10:38 AM - Apr 19, 2017. It features two side-by-side photos of a man's face, one original and one with a "bleach" filter. The text reads: "#faceapp isn't! just bad it's also racist... filter=bleach my skin and make my nose your opinion of European. No thanks #uninstalled". The bottom tweet is from @littlebunnyfu dated 10:20 PM - Apr 17, 2017. It features two side-by-side photos of a woman's face, one original and one with a "white eyes" filter. The text reads: "#faceapp removes glasses and replaces eyes. With white people eyes. :|". The Motherboard logo is visible in the top right corner.

# 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
- Web page: <http://www.ece.cmu.edu/~ece739/>
- Canvas (for grades, homework)
- Piazza (for all other communication)
  - Please enroll; you should have received invitation
- 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

# Collaboration policy on homework

- You are allowed 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.

[CMU Computing Policy](#)

[CMU Policy on Cheating](#)