

Security and Fairness of Deep Learning

Course Overview

Anupam Datta

CMU

Spring 2018

Course staff

- Instructor: Anupam Datta
 - Office: B23 221
 - Email: danupam@cmu.edu
 - Office hours: Thur 12-1pm Pacific
 - Google hangouts: link on Piazza

- TA: Caleb Lu
 - Office: Bldg 19 1031
 - Email: caleb.lu@sv.cmu.edu
 - Office hours: Mon 10am-12pm Pacific
 - Google hangouts: link on Piazza



Recent successes of deep learning

The image shows two overlapping browser windows. The top window is 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 "Tech Blog". The bottom window is a Google 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 graphic with the text "Found in translation: More accurate, fluent sentences in Google Translate".

The image shows a screenshot of a Nature journal article. The header includes the Nature logo and the text "International journal of science". The article title is "Dermatologist-level classification of skin cancer with deep neural networks". The authors listed are Andre Esteva, Brett Kuperl, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. The article is categorized as a "Letter". The publication date is 02 February 2017, with a doi of 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 is associated with the keywords "Diagnosis", "Machine learning", and "Skin cancer". The article is also associated with the content "Medicine: The final frontier in cancer diagnosis" by Sancy A. Leachman & Glenn Merlino.

Image classification

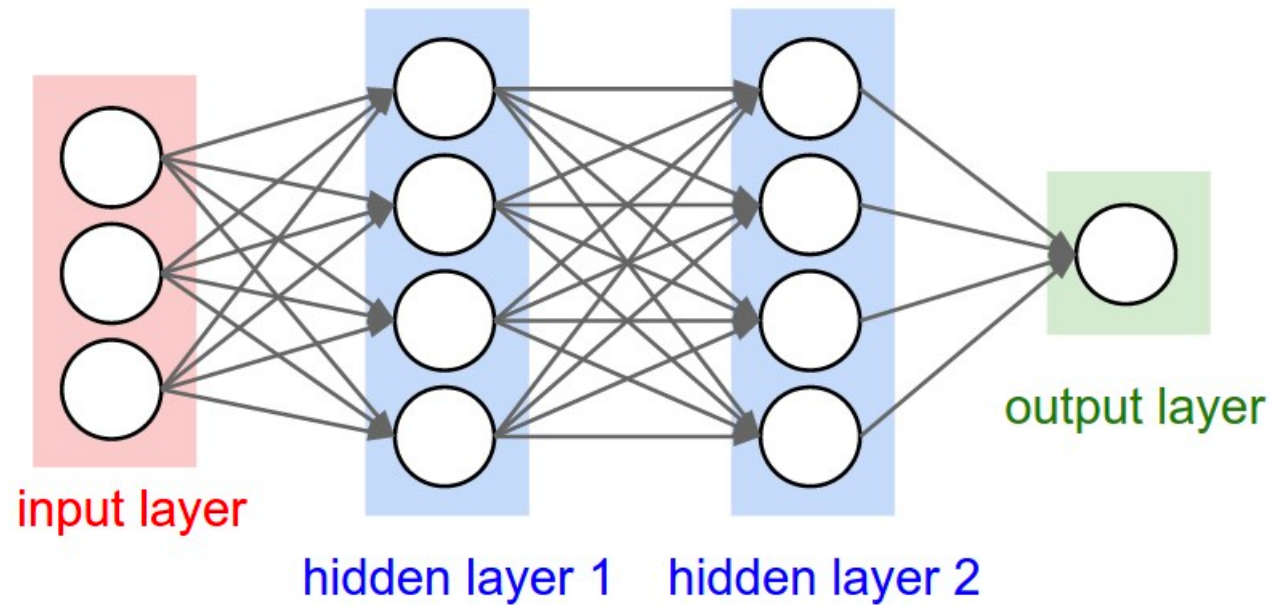


08	02	22	97	38	15	00	40	00	75	04	05	07	78	52	12	50	77	81	02
49	49	99	40	17	81	18	57	60	87	17	40	98	43	69	47	04	56	62	00
81	49	31	73	55	79	14	29	93	71	40	67	57	88	30	03	49	13	36	65
52	70	95	23	04	60	11	42	68	21	68	56	01	32	56	71	37	02	36	91
22	31	16	71	51	67	85	89	41	92	36	54	22	40	40	28	66	33	13	80
24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

What the computer sees

image classification → 82% cat
15% dog
2% hat
1% mug

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

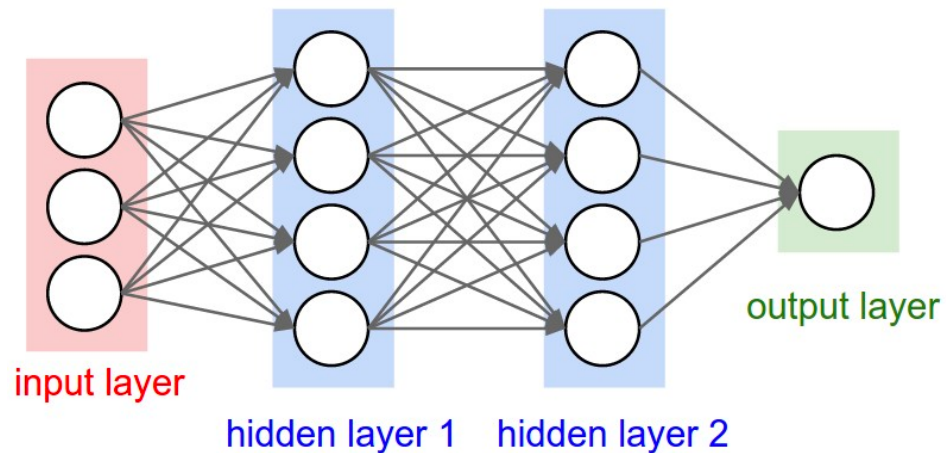
Course modules

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



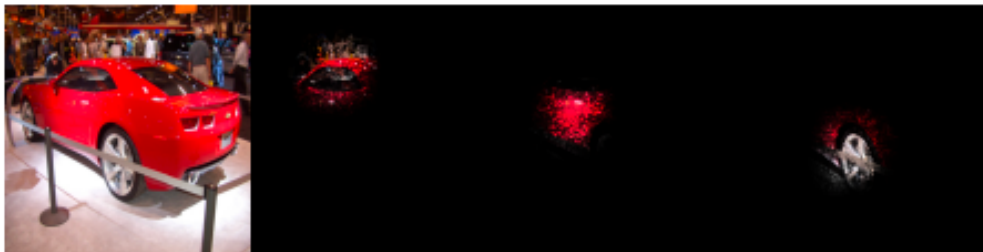
theano




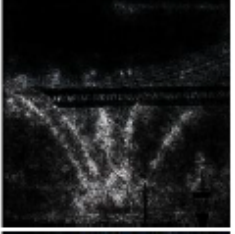






Course modules

2. Unlocking the black box

- Explaining behavior of deep neural networks



Original image	Top label and score	Integrated gradients
	Top label: reflex camera Score: 0.993755	
	Top label: fireboat Score: 0.999961	
	Top label: school bus Score: 0.997033	
	Top label: mosque Score: 0.999127	

Course modules

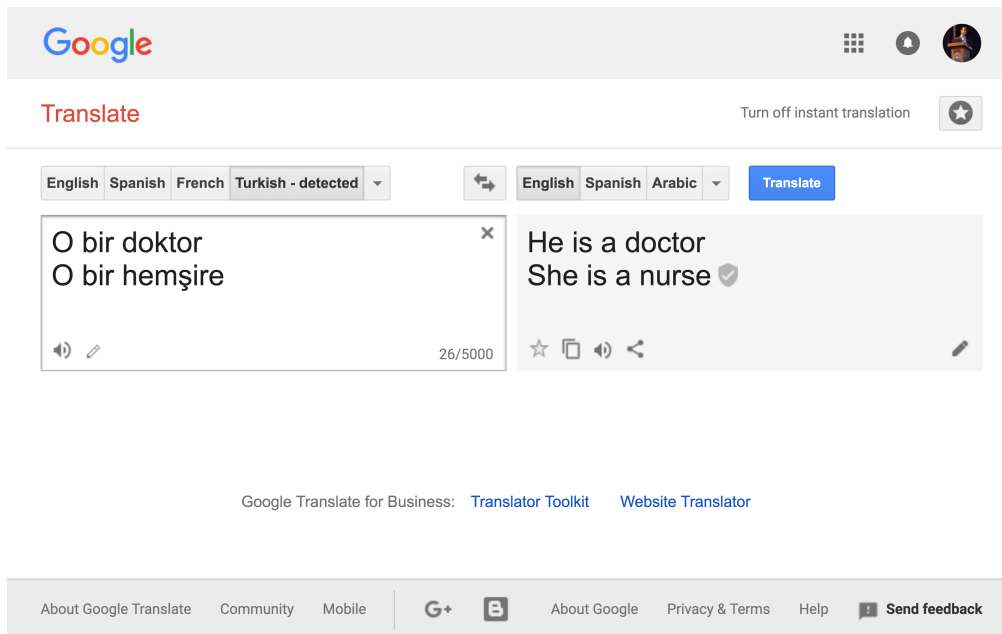
3. Security of deep learning models
 - Attacks on classifiers and defenses



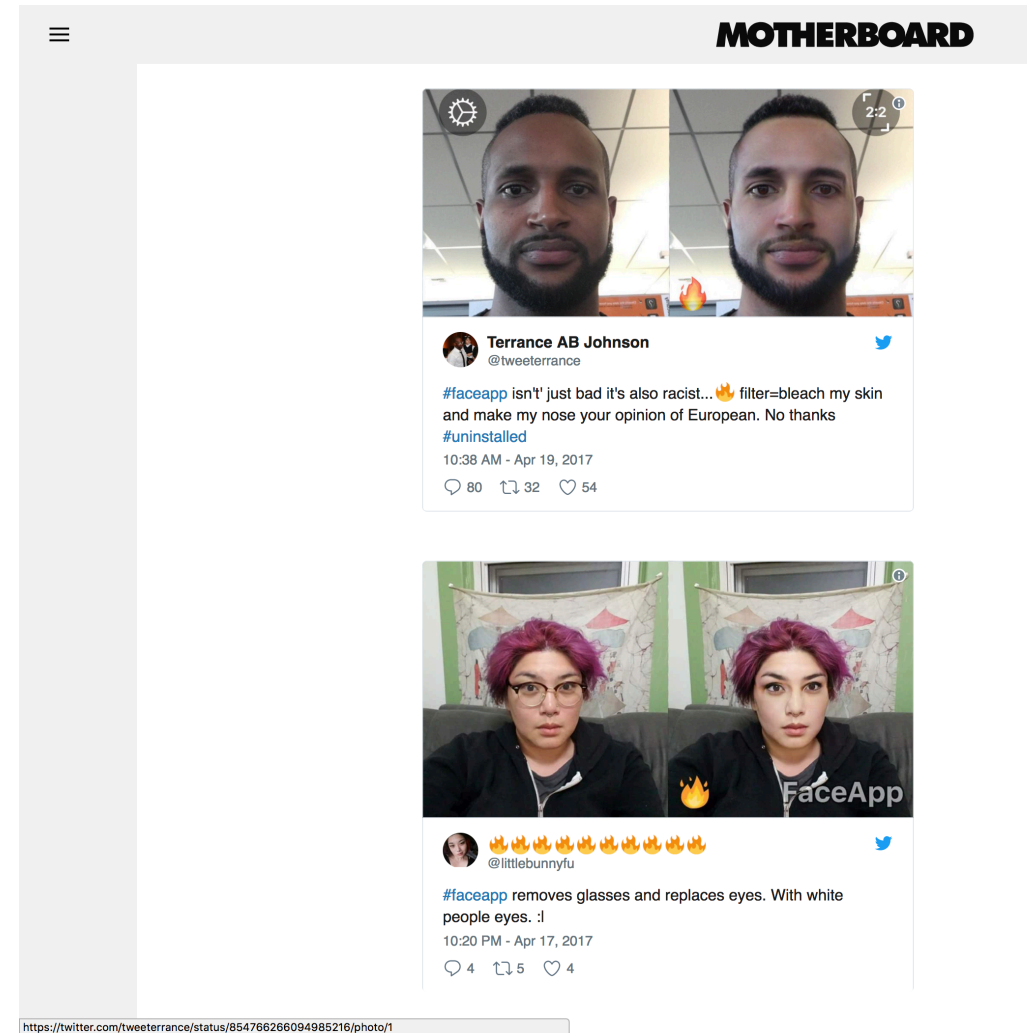
Course modules

4. Fairness of deep learning

- Bias and de-biasing



The screenshot shows the Google Translate web interface. At the top left is the Google logo. Below it, the word "Translate" is displayed in red. To the right of "Translate" is a link that says "Turn off instant translation" and a star icon. Below this, there are two language selection boxes. The first box shows "English", "Spanish", "French", and "Turkish - detected" with a dropdown arrow. The second box shows "English", "Spanish", and "Arabic" with a dropdown arrow. A blue "Translate" button is positioned between the two boxes. Below the language boxes, there are two text boxes. The left box contains the Turkish text "O bir doktor" and "O bir hemşire" with a close button (x) and a character count "26/5000". The right box contains the English translation "He is a doctor" and "She is a nurse" with a checkmark icon. Below the text boxes are icons for voice input, copy, and share. At the bottom of the interface, there are links for "Google Translate for Business: Translator Toolkit" and "Website Translator". The footer contains 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 right corner has the "MOTHERBOARD" logo. The first tweet is from Terrance AB Johnson (@tweeterrance) and includes a side-by-side comparison of a man's face before and after using a FaceApp filter. The text of the tweet 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 tweet is dated "10:38 AM - Apr 19, 2017" and has 80 replies, 32 retweets, and 54 likes. The second tweet is from @littlebunnyfu and includes a side-by-side comparison of a woman's face before and after using a FaceApp filter. The text of the tweet reads: "🔥🔥🔥🔥🔥🔥🔥🔥🔥 #faceapp removes glasses and replaces eyes. With white people eyes. :l". The tweet is dated "10:20 PM - Apr 17, 2017" and has 4 replies, 5 retweets, and 4 likes. At the bottom of the screenshot, there is a URL: "https://twitter.com/tweeterrance/status/854766266094985216/photo/1".

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/>
- Temp web page: <http://www.andrew.cmu.edu/user/kaijil/class-18739/>

- 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: 80%
 - 4 x 20%
- Paper summaries: 10%
 - 5 x 2%
- 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](#)