

Security and Fairness of Deep Learning

Course Summary

Anupam Datta

CMU

Spring 2018

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 October 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 November 15, 2016. The blog post features a large yellow background with the title text and a blue share icon at the bottom right.

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 February 02, 2017. The page also displays the article's DOI (10.1038/nature21056), its Altmetric score (2665), and its citation count (85). The article is available for download and citation. The page includes a search bar, a navigation menu, and a sidebar with social media sharing options and associated content.

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](#)

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

Diagnosis Machine learning
Skin cancer

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

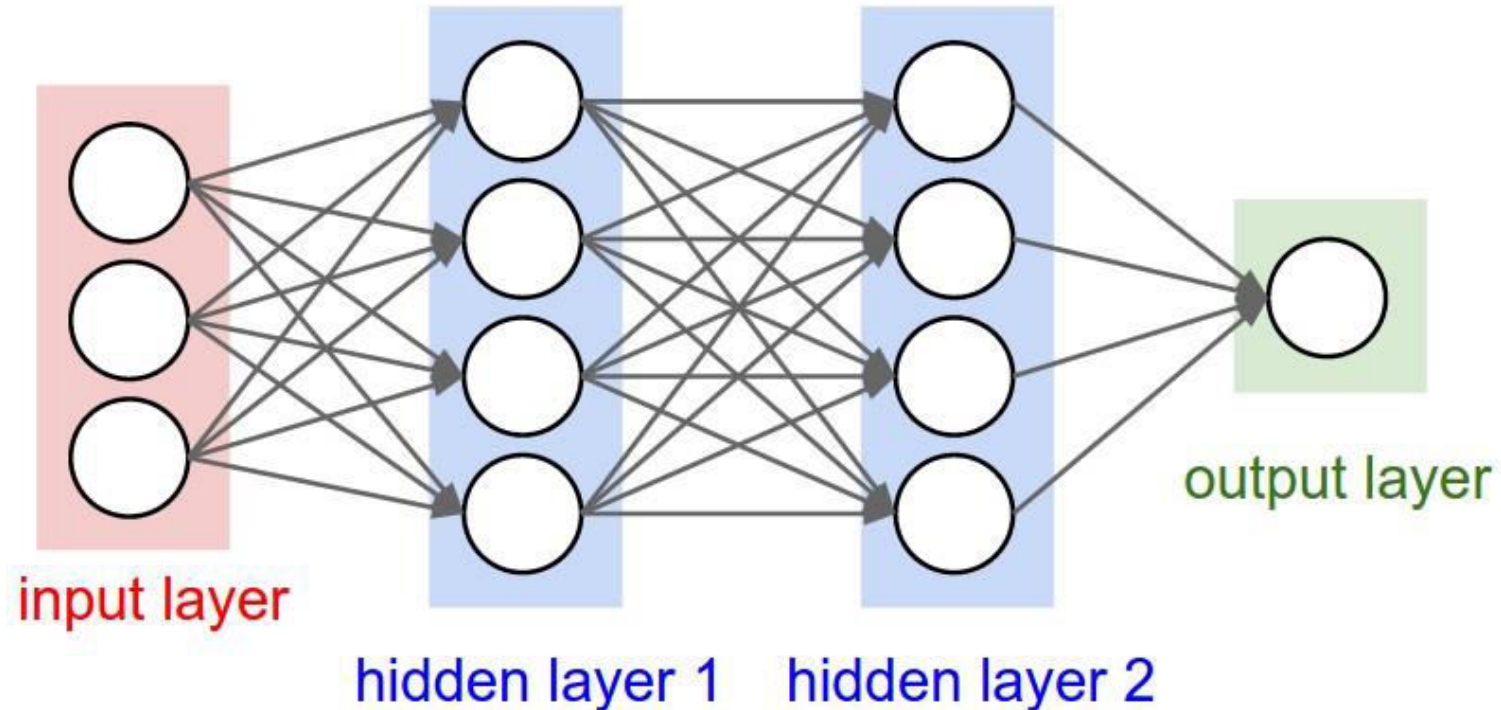


05	02	21	97	38	15	00	40	00	95	04	05	07	78	52	12	50	77	31	21
49	49	99	40	17	81	18	57	60	87	17	40	98	43	69	45	04	56	62	00
81	49	31	73	55	79	14	29	93	71	40	67	57	03	30	03	49	13	36	65
52	70	95	23	04	60	11	42	69	44	69	56	01	32	56	71	37	02	36	91
22	31	16	71	51	67	02	59	41	92	36	54	22	40	40	28	66	33	13	80
24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

- From first principles to state-of-the-art

Course modules

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

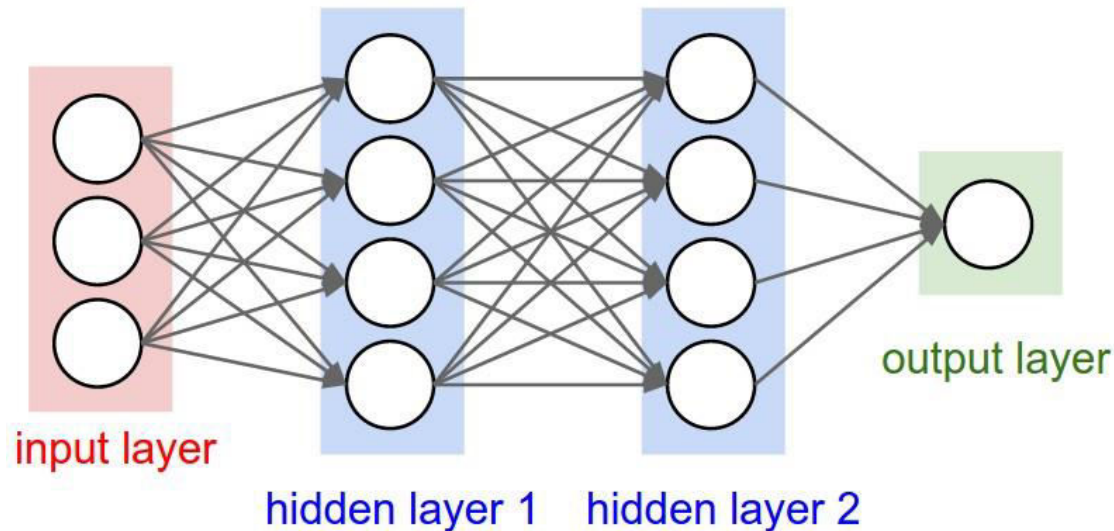
Organized around

- computer vision tasks with convolutional neural networks
- natural language processing tasks with recurrent neural networks

Course modules

1. Fundamentals of deep networks

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

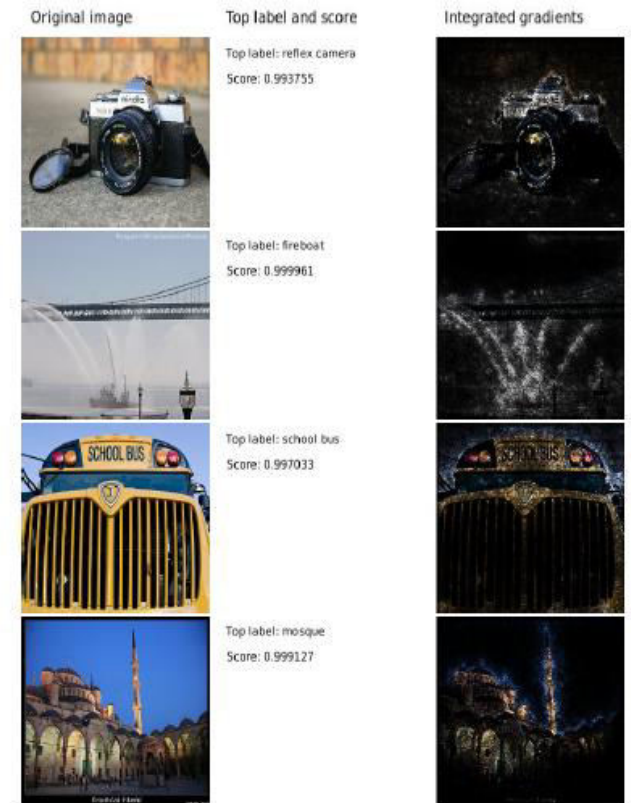
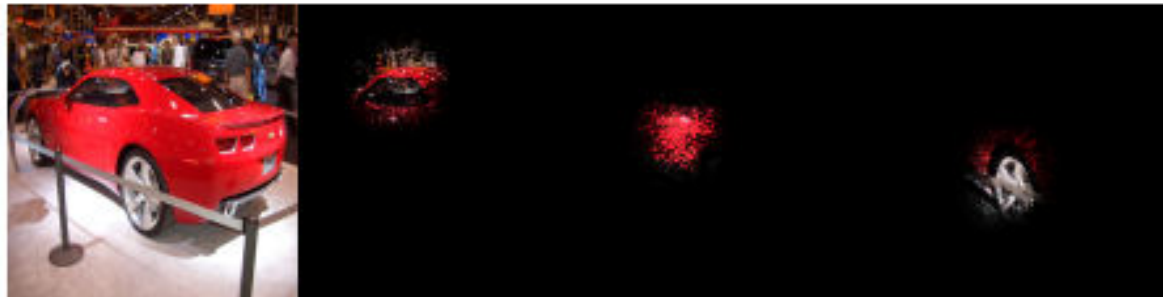


- HW1: Intro to ML and libraries with logistic regression
- HW2: Training CNNs

Course modules

2. Unlocking the black box

- Explaining behavior of deep neural networks



- HW3: Explanations for CNNs

Course modules

3. Security of deep learning models

- Attacks on classifiers and defenses

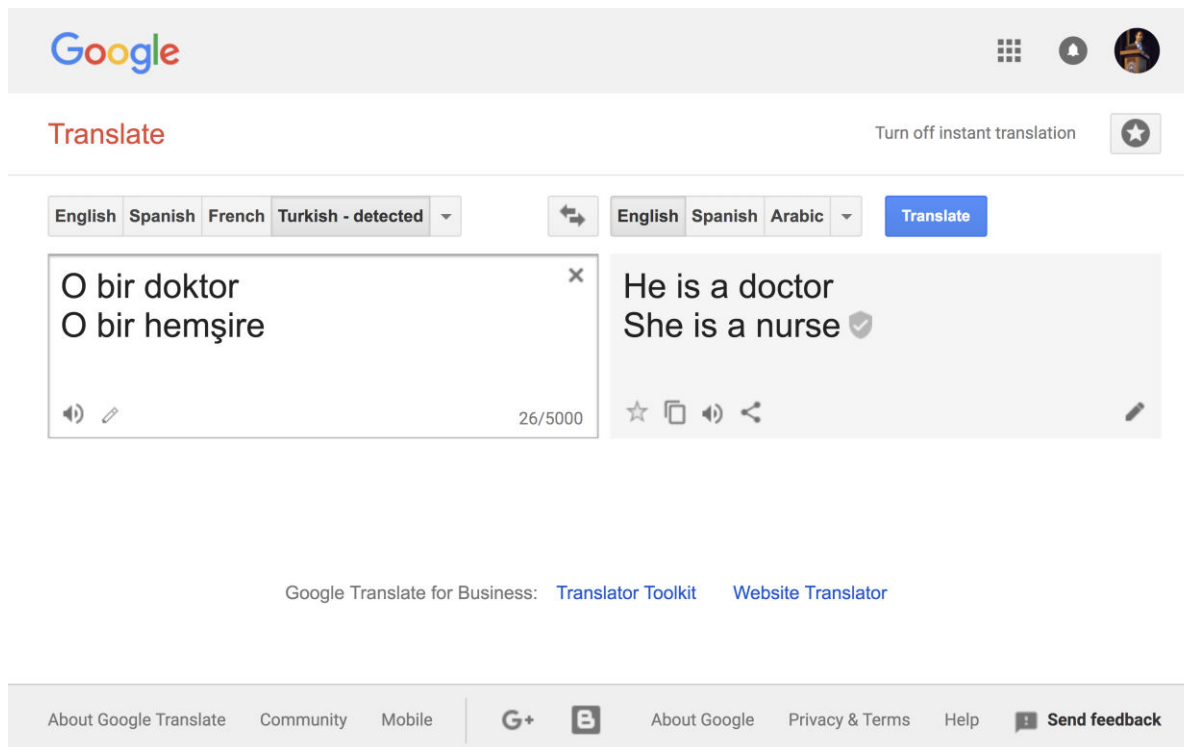


- HW4: Adversarial learning with GANs and attacks on CNNs

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 "Turn off instant translation" and a star icon. Below this, there are two dropdown menus for language selection. The first dropdown is set to "Turkish - detected" and the second is set to "English". A blue "Translate" button is positioned to the right of the second dropdown. The main content area is split into two panels. The left panel contains the Turkish text "O bir doktor" and "O bir hemşire". The right panel contains the English translation "He is a doctor" and "She is a nurse". Below the text in both panels are icons for audio playback, a star, a copy icon, and a share icon. At the bottom of the page, 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".

NLP tasks

- word embeddings
- neural language models
- neural machine translation

- HW4: Word2vec and bias in NLP

Course modules

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

Organized around

- computer vision tasks with convolutional neural networks
- natural language processing tasks with recurrent neural networks

Security and Fairness of Deep Learning

Multilingual Neural Machine Translation

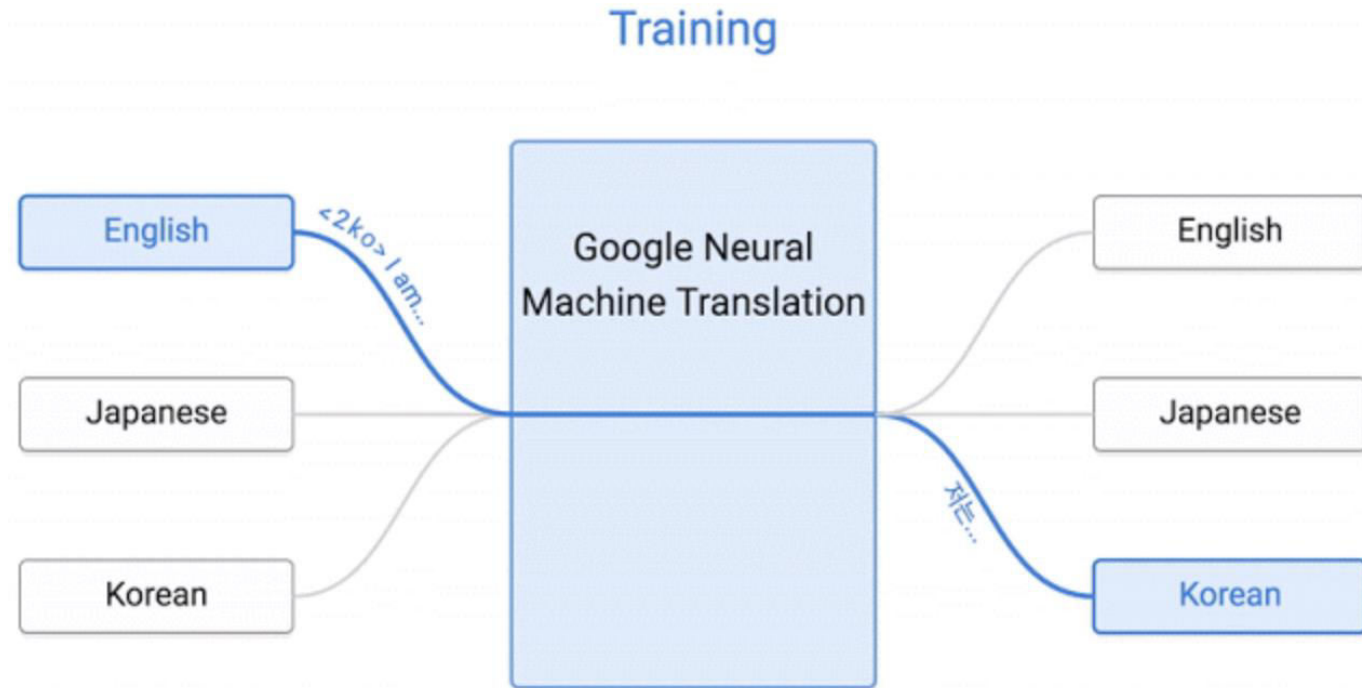
Anupam Datta
CMU

Spring 2018

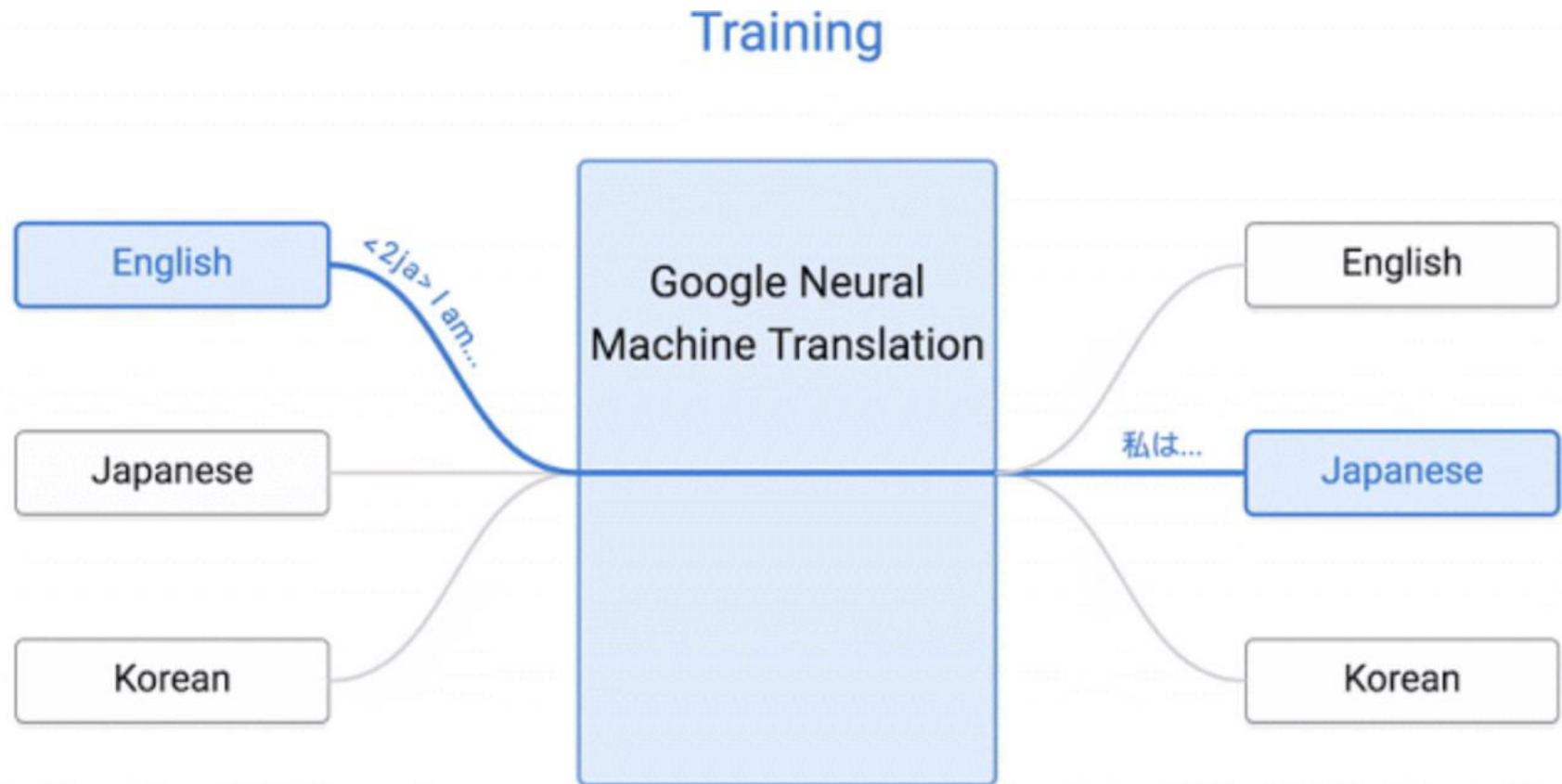
Google's Multilingual Neural Machine Translation System

- Single neural machine translation model to translate between multiple languages
- Used in production
- Key characteristics
 - Simplicity
 - Low-resource language improvements
 - Zero shot learning
 - Universal interlingua representation

Training on some language pairs



Training on some language pairs



Training

Hello, how are you? -> Hola, ¿cómo estás?

It will be modified to:

<2es> Hello, how are you? -> Hola, ¿cómo estás?

- Token indicates target language
- No token for source language – context provides enough language evidence for correct translation
- Over- or under-sampling to adjust for the relative availability of language data
- Same architecture as single language translation

Experimental Results

- Some multilingual models take a little more time to train than single language pair models, likely because each language pair is seen only for a fraction of the training process.
- Depending on the number of languages a full training can take up to 10M steps and 3 weeks to converge (on roughly 100 GPUs).
- Use larger batch sizes with a slightly higher initial learning rate to speed up the convergence of these models.

Many to one translation

Table 1: Many to One: BLEU scores on various data sets for single language pair and multilingual models.

Model	Single	Multi	Diff
WMT German→English (oversampling)	30.43	30.59	+0.16
WMT French→English (oversampling)	35.50	35.73	+0.23
WMT German→English (no oversampling)	30.43	30.54	+0.11
WMT French→English (no oversampling)	35.50	36.77	+1.27
Prod Japanese→English	23.41	23.87	+0.46
Prod Korean→English	25.42	25.47	+0.05
Prod Spanish→English	38.00	38.73	+0.73
Prod Portuguese→English	44.40	45.19	+0.79

Many-to-many

Table 3: Many to Many: BLEU scores on various data sets for single language pair and multilingual models.

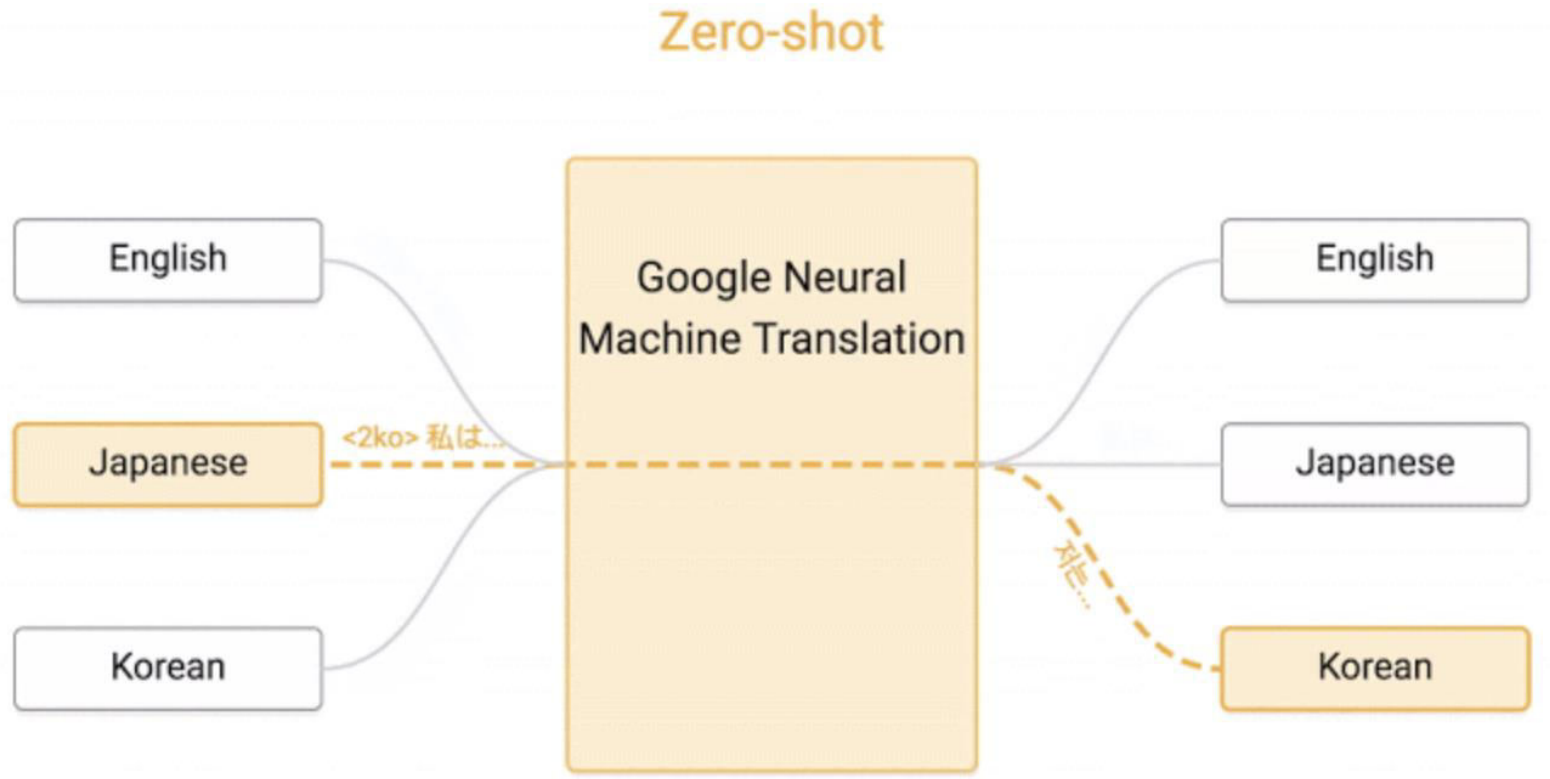
Model	Single	Multi	Diff
WMT English→German (oversampling)	24.67	24.49	-0.18
WMT English→French (oversampling)	38.95	36.23	-2.72
WMT German→English (oversampling)	30.43	29.84	-0.59
WMT French→English (oversampling)	35.50	34.89	-0.61
WMT English→German (no oversampling)	24.67	21.92	-2.75
WMT English→French (no oversampling)	38.95	37.45	-1.50
WMT German→English (no oversampling)	30.43	29.22	-1.21
WMT French→English (no oversampling)	35.50	35.93	+0.43
Prod English→Japanese	23.66	23.12	-0.54
Prod English→Korean	19.75	19.73	-0.02
Prod Japanese→English	23.41	22.86	-0.55
Prod Korean→English	25.42	24.76	-0.66
Prod English→Spanish	34.50	34.69	+0.19
Prod English→Portuguese	38.40	37.25	-1.15
Prod Spanish→English	38.00	37.65	-0.35
Prod Portuguese→English	44.40	44.02	-0.38

Large-scale experiments

Table 4: Large-scale experiments: BLEU scores for single language pair and multilingual models.

Model	Single	Multi	Multi	Multi	Multi
#nodes	1024	1024	1280	1536	1792
#params	3B	255M	367M	499M	650M
Prod English→Japanese	23.66	21.10	21.17	21.72	21.70
Prod English→Korean	19.75	18.41	18.36	18.30	18.28
Prod Japanese→English	23.41	21.62	22.03	22.51	23.18
Prod Korean→English	25.42	22.87	23.46	24.00	24.67
Prod English→Spanish	34.50	34.25	34.40	34.77	34.70
Prod English→Portuguese	38.40	37.35	37.42	37.80	37.92
Prod Spanish→English	38.00	36.04	36.50	37.26	37.45
Prod Portuguese→English	44.40	42.53	42.82	43.64	43.87
Prod English→German	26.43	23.15	23.77	23.63	24.01
Prod English→French	35.37	34.00	34.19	34.91	34.81
Prod German→English	31.77	31.17	31.65	32.24	32.32
Prod French→English	36.47	34.40	34.56	35.35	35.52
ave diff	-	-1.72	-1.43	-0.95	-0.76
vs single	-	-5.6%	-4.7%	-3.1%	-2.5%

Zero-shot translation



Zero-shot translation

Table 5: Portuguese→Spanish BLEU scores using various models.

	Model	Zero-shot	BLEU
(a)	PBMT bridged	no	28.99
(b)	NMT bridged	no	30.91
(c)	NMT Pt→Es	no	31.50
(d)	Model 1 (Pt→En, En→Es)	yes	21.62
(e)	Model 2 (En↔{Es, Pt})	yes	24.75
(f)	Model 2 + incremental training	no	31.77

Universal interlingua representation

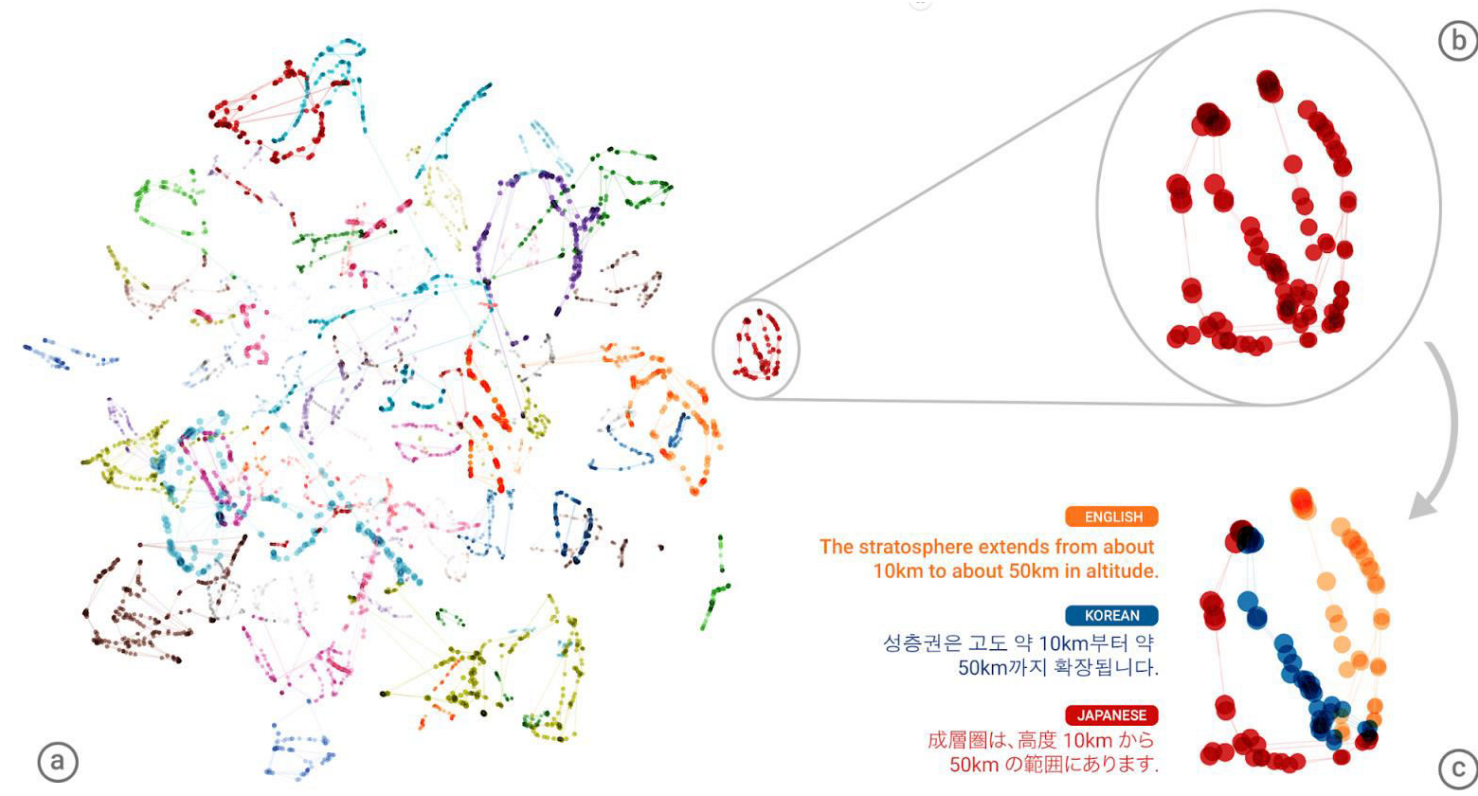
- Is the network learning some sort of shared representation, in which sentences with the same meaning are represented in similar ways regardless of language?

Universal interlingua representation

Examine sequence of context vectors generated during translation (i.e., the sum of internal encoder states weighted by their attention probabilities per step)

- Do sentences cluster together depending on the source or target language?
- Or instead do sentences with similar meanings cluster, regardless of language?

Universal interlingua representation



Shared wordpiece

- **Word:** Jet makers feud over seat width with big orders at stake
- **wordpieces:** _J et _makers _fe ud _over _seat _width _with _big _orders _at _stake