Course Summary

Anupam Datta

CMU

Spring 2018

Recent successes of deep learning

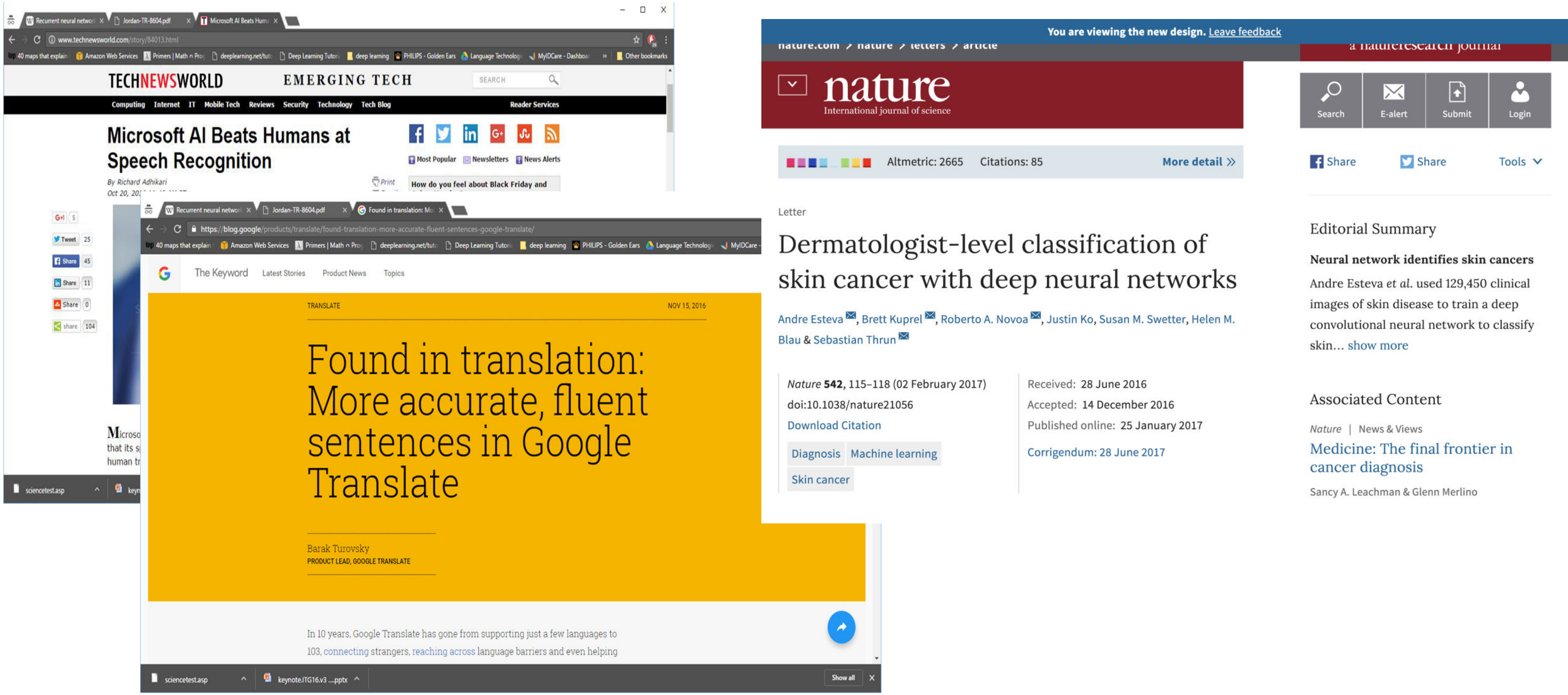
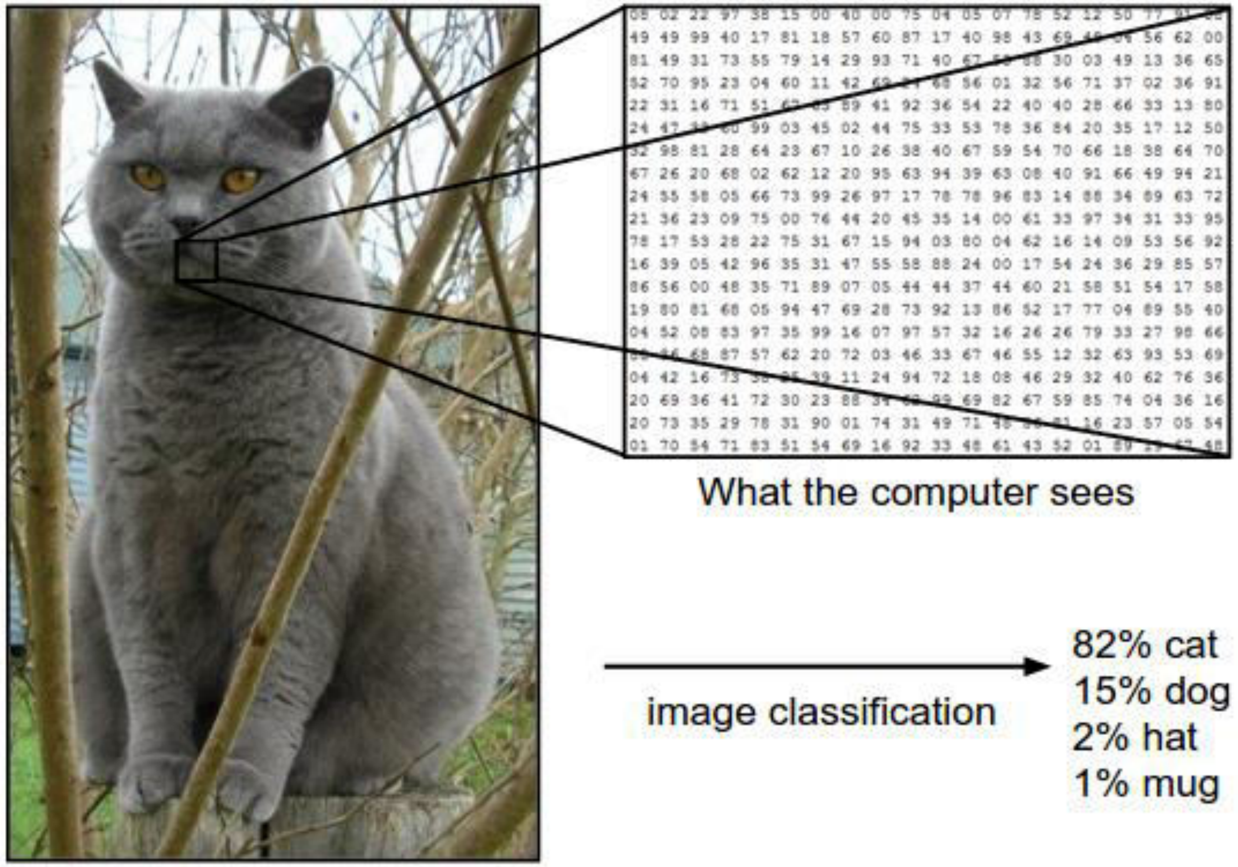
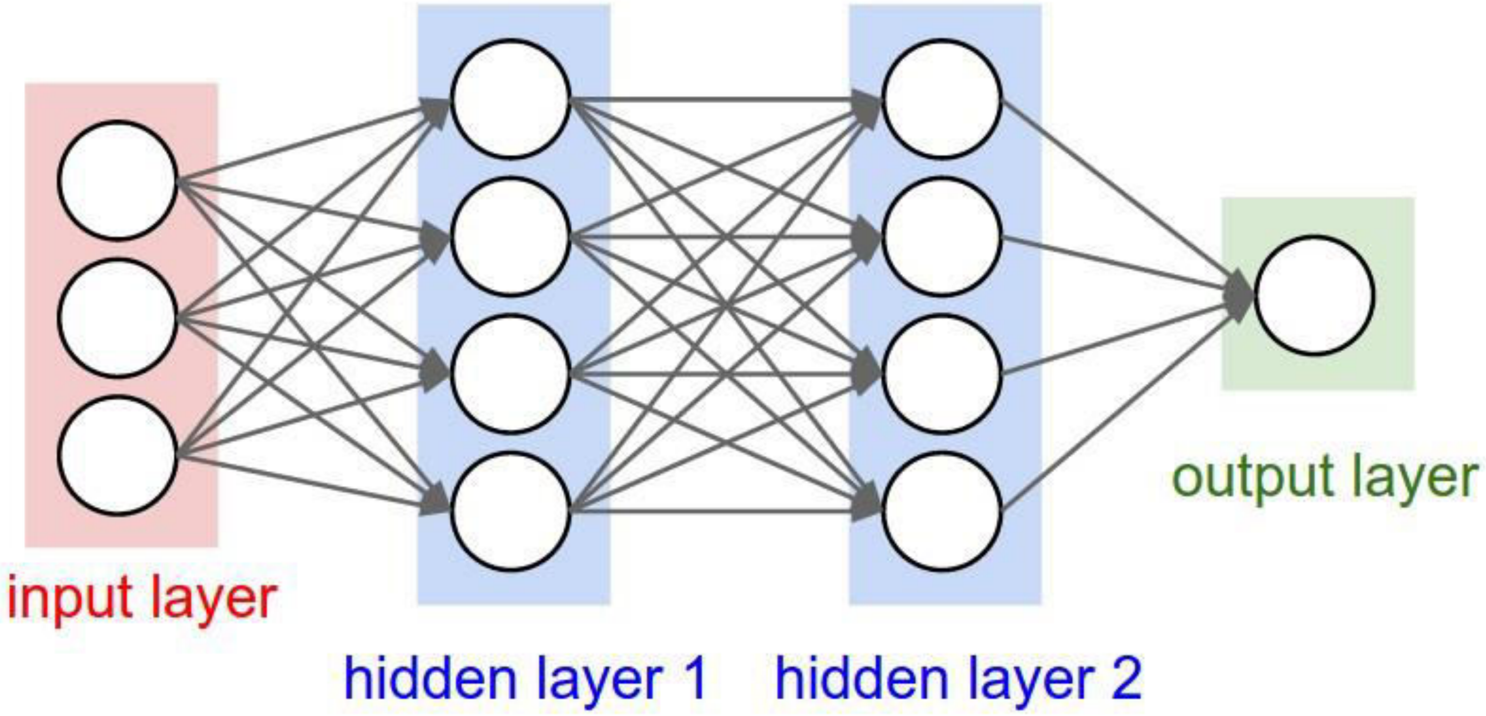


Image classification



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

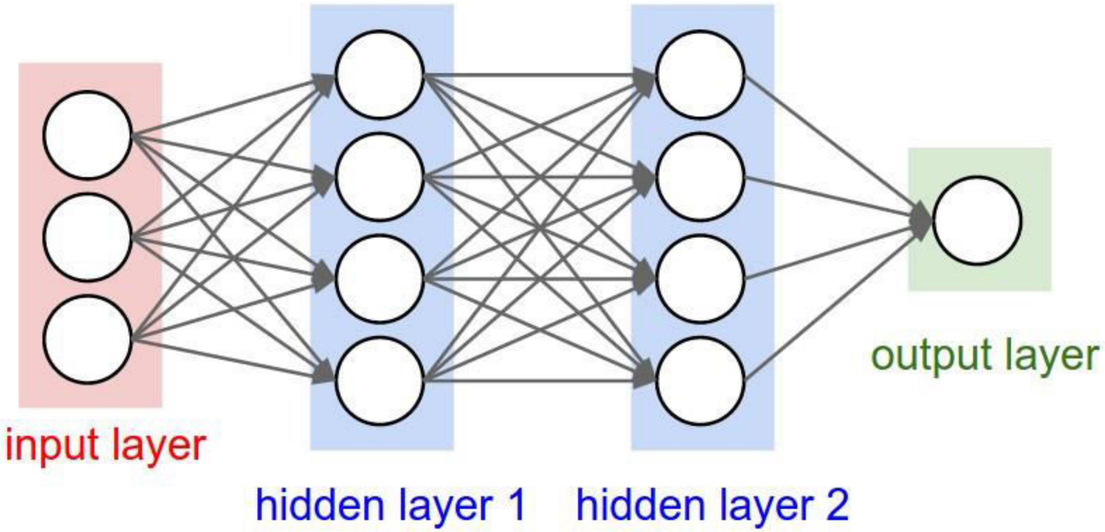
1. Fundamentals of deep networks
2. Unlocking the black box
3. Security of deep learning models
4. Fairness of deep learning

Organized around

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

Course modules

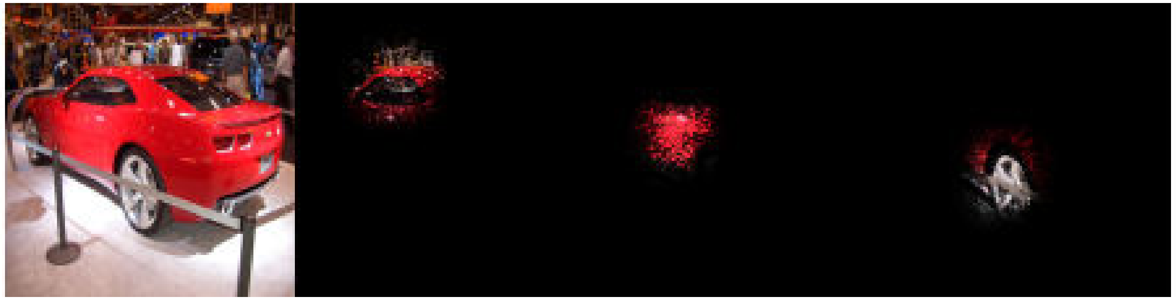
1. Fundamentals of deep networks
    - Background on machine learning
    - Architectures, training, platforms
    - Focus on convolutional and recurrent neural networks



- 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



Keras

theano

- HW3: Explanations for CNNs

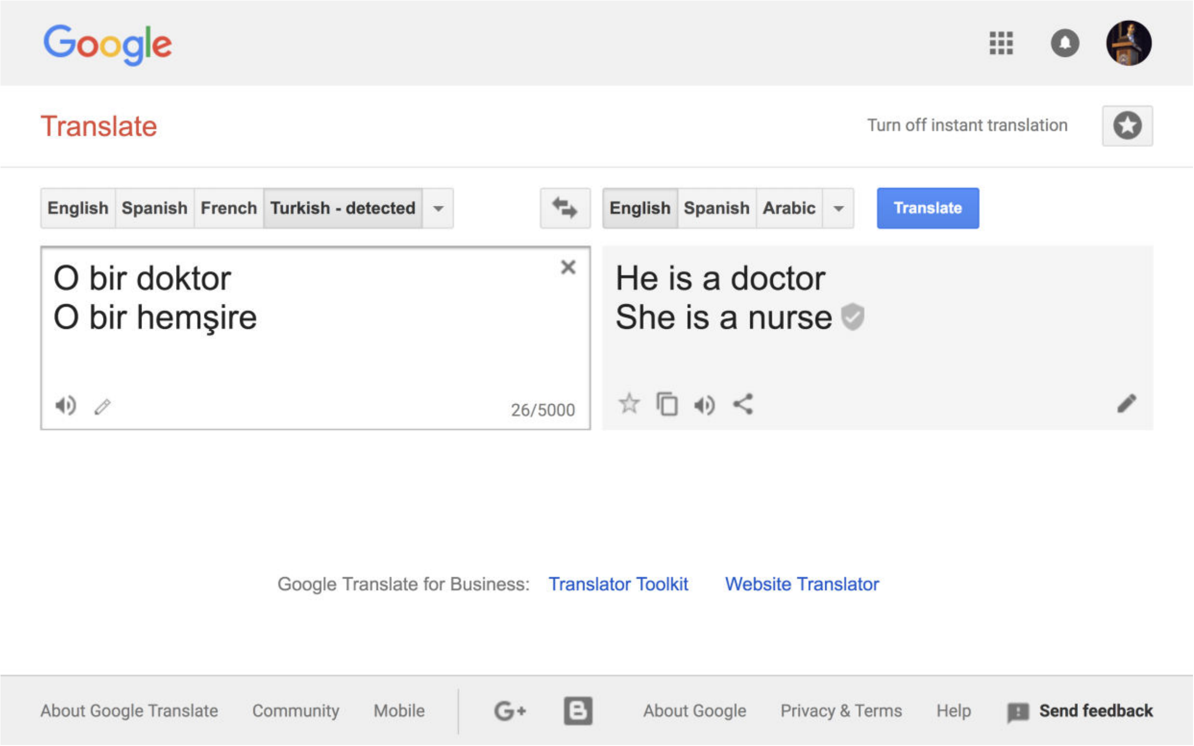
Course modules

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



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

Multilingual Neural Machine Translation

Anupam Datta

CMU

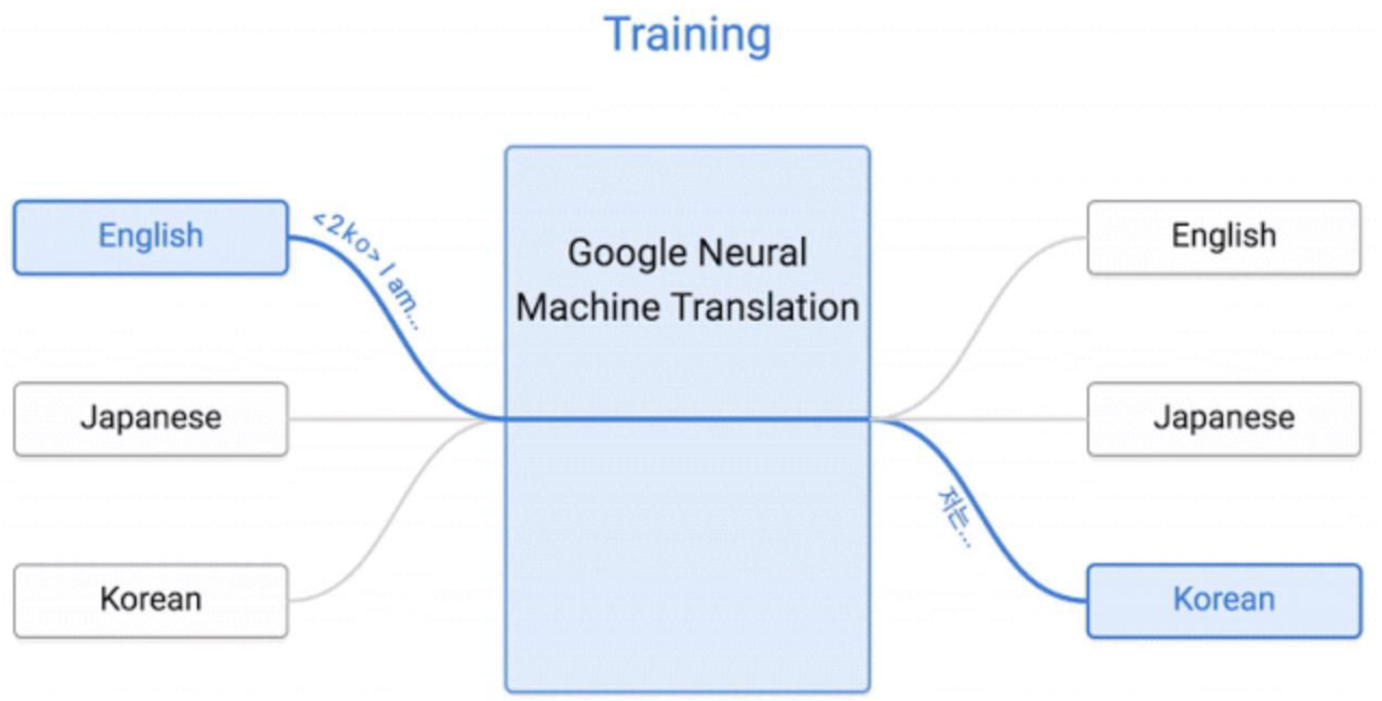
Spring 2018

Google's Multilingual Neural Machine Translation System

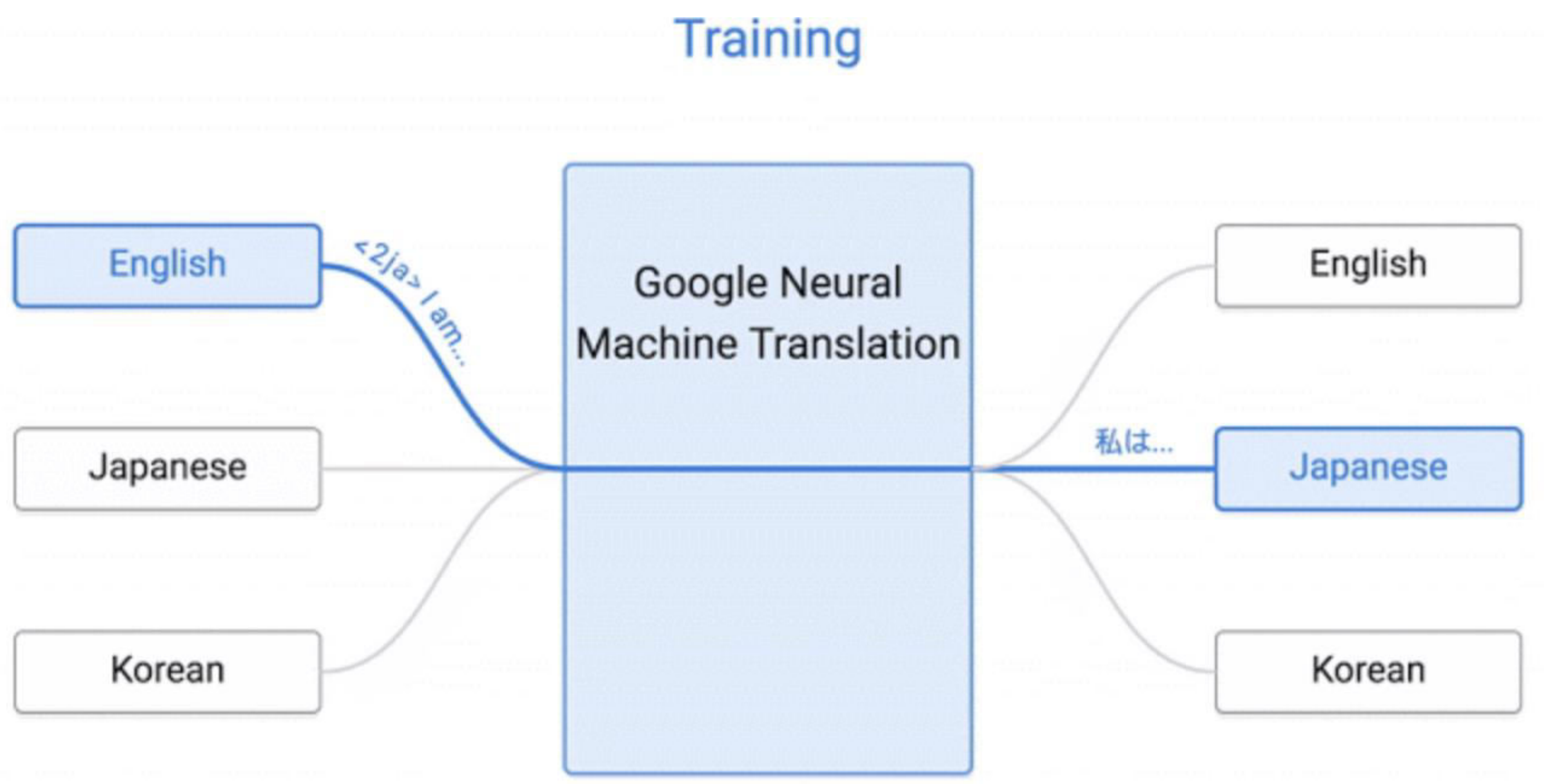
- Single neural machine translation model to translate between multiple languages
- Used in production

- 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.

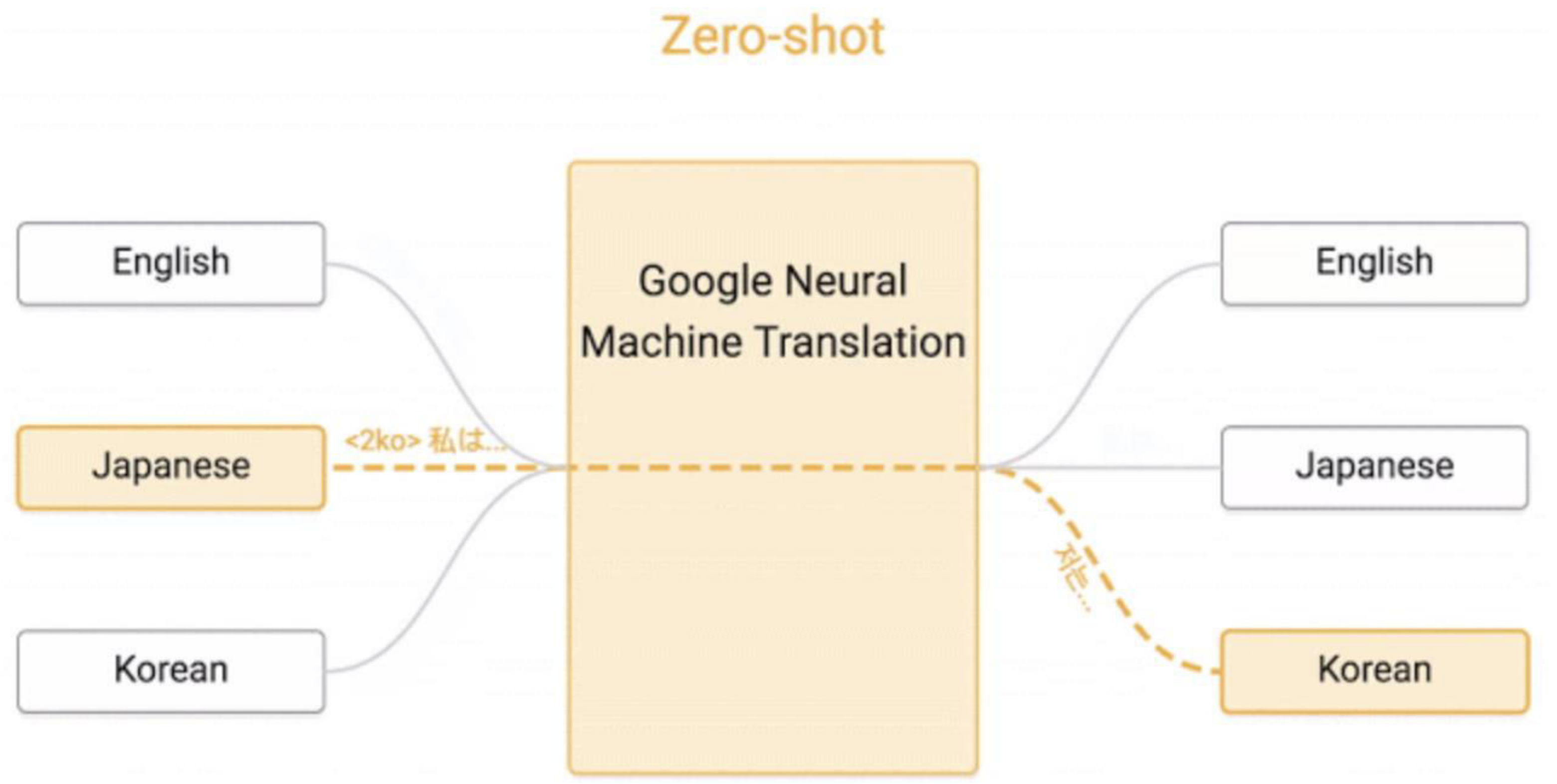
Model	Single	Multi	Diff
WMT English→German (oversampling)	24.67	24.49	-0.18
WMT English→French (oversampling)	38.95	36.23	-2.72
WMT German→English (oversampling)	30.43	29.84	-0.59
WMT French→English (oversampling)	35.50	34.89	-0.61
WMT English→German (no oversampling)	24.67	21.92	-2.75
WMT English→French (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

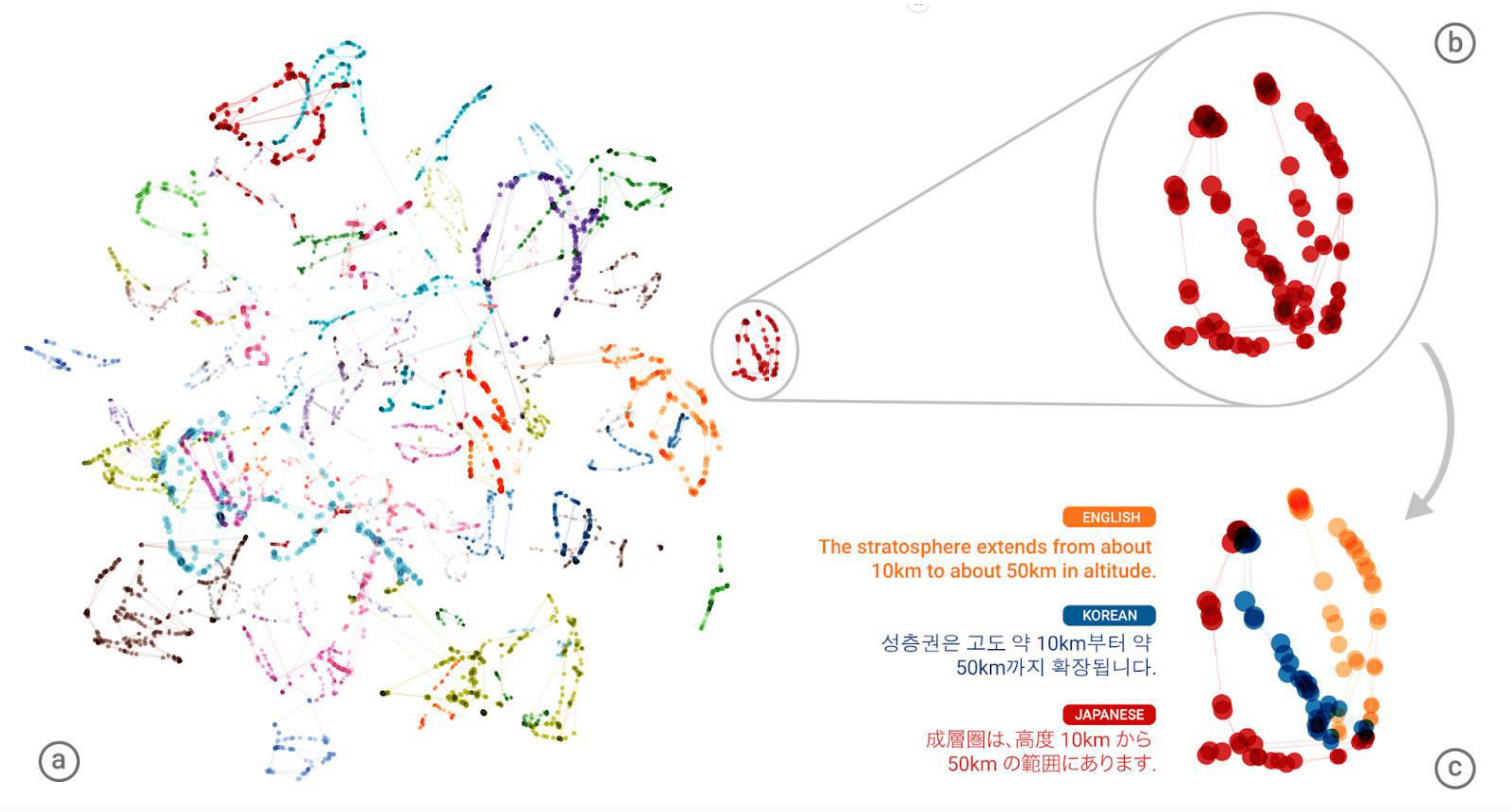
- Is the network learning some sort of shared representation, in which sentences with the same meaning are represented in similar ways regardless of language?

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