Advanced Topics in Machine Learning

Alejo Nevado-Holgado

Lecture 15 (NLP 7) - Question answering, conference resolution and CNNs

V 0.3 (26 Feb 2020 - final version)

Course structure

- ➢ **Introduction:** What is NLP. Why it is hard. Why NNs work well ← Lecture 9 (NLP 1)
- ➢ **Word representation:** How to represent the meaning of individual words
  - Old technology: One-hot representations, synsets ← Lecture 9 (NLP 1)
  - Embeddings: First trick that boosted the performance of NNs in NLP ← Lecture 9 (NLP 1)
    - Word2vec: Single layer NN. CBOW and skip-gram ← Lecture 10 (NLP 2)
    - Co-occurrence matrices: Basic counts and SVD improvement ← Lecture 10 (NLP 2)
    - Glove: Combining word2vec and co-occurrence matrices idea ← Lecture 10 (NLP 2)
    - Evaluating performance of embeddings ← Lecture 10 (NLP 2)
- ➢ **Named Entity Recognition (NER):** How to find words of specific meaning within text
  - Multilayer NNs: Margin loss. Forward- and back-propagation ← Lecture 11 (NLP 3)
  - Better loss functions: margin loss, regularisation ← Lecture 11 (NLP 3)
  - Better initializations: uniform, xavier ← Lecture 11 (NLP 3)
  - Better optimizers: Adagrad, RMSprop, Adam... ← Lecture 11 (NLP 3)

Course structure

- **Language modelling:** How to represent the meaning of full pieces of text
  - Old technology: N-grams ← Lecture 12 (NLP 4)
  - Recursive NNs language models (RNNs) ← Lecture 12 (NLP 4)
  - Evaluating performance of language models ← Lecture 12 (NLP 4)
  - Vanishing gradients: Problem. Gradient clipping ← Lecture 13 (NLP 5)
  - Improved RNNs: LSTM, GRU, Bidirectional... ← Lecture 13 (NLP 5)
- **Machine translation:** How to translate text
  - Old technology: Georgetown–IBM experiment and ALPAC report ← Lecture 14 (NLP 6)
  - Seq2seq: Greedy decoding, encoder-decoder, beam search ← Lecture 14 (NLP 6)
  - Attention: Simple attention, transformers, reformers ← Lecture 14 (NLP 6)
  - Evaluating performance: BLEU ← Lecture 14 (NLP 6)

Course structure

➢ **Question Answering:** X
- Task definition, datasets, cloze-style tasks, Attentive Reader ← Lecture 15 (NLP 7)

➢ **Conference Resolution:** X
- Task definition, pairs method, clustering method, language models ← Lecture 15 (NLP 7)

➢ **Convolutional Neural Networks:** X
- CNNs in vision, CNNs in language, example ← Lecture 15 (NLP 7)

➢ **Transformers:** X
- Architecture: encoder, self-attention, encoding position, decoder ← Lecture 16 (NLP 8)
- Existing systems. Ranking ← Lecture 16 (NLP 8)

Questions answering: The problem

Question
When were the first pyramids built?
Jean-Claude Juncker
How old is Keir Starmer?
What is the current price for AAPL?
What's the weather like in London?
Whom did Juncker meet with?
When did you get to this lecture?
Why do we yawn?

Questions answering: The problem

Question	Answer
When were the first pyramids built?	*2630 BC*
Jean-Claude Juncker	*Jean-Claude Juncker is a Luxembourgish politician. Since 2014, Juncker has been President of the European Commission.*
How old is Keir Starmer?	*54 years*
What is the current price for AAPL?	*136.50 USD*
What's the weather like in London?	*7 degrees Celsius. Clear with some clouds.*
Whom did Juncker meet with?	*The European Commission president was speaking after meeting with Irish Taoiseach Enda Kenny in Brussels.*
When did you get to this lecture?	*Five minutes after it started.*
Why do we yawn?	*When we're bored or tired we don't breathe as deeply as we normally do. This causes a drop in our blood-oxygen levels and yawning helps us counter-balance that.*

Questions answering: Why deal with it?

- **Because QA is awesome**
  - That’s it
- **Because QA is AI-complete**
  - Theoretically, if we solve QA, we can solve everything
- **Because QA has many immediate applications**
  - Fine grain search, dialogue, information extraction, summarisation...
- **Some very good results already**
  - IBM Watson and Jeopardy, Siri, Google Search...
- **Many other improvements and applications left to do**
  - Both challenging problems and low hanging fruit

Questions answering: Deployed systems

Questions answering: Deployed systems

- **Before 2015**
  - MCTest (Richardson et al 2013): 2.6K questions
  - ProcessBank (Berant et al 2014): 500 questions

- **After 2015**
  - CNN/Daily Mail, Mage-GRU[arXiv:1703.02620v1]
  - Children Book Test, GPT2
  - WikiQA, TANDA[arXiv:1911.04118v2]
  - TriviaQA, MemoReader[10.18653/v1/D18-1237]
  - SQuAD 2.0, alBERT[arXiv:2001.09694v1]
  - SQuAD 1.1, XLNet[arXiv:1906.08237v2]
  - News QA, BERT[arXiv:1907.10529v3]
  - MS MARCO, Masque[arXiv:1901.02262v2]
  - More than 100K questions!

http://nlpprogress.com/english/question_answering.html

Questions answering: Deployed systems

- **Before 2015**
  - Lexical matching
  - Logistic regression

- **After 2015**
  - Attentive reader
  - Memory networks
  - ReasoNet
  - Match-LSTM
  - Attention sum reader
  - Attention-over-attention reader
  - Iterative attention reader
  - Dynamic coattention networks
  - Bi-directional Attention Flow Network
  - Multi-perspective Context Matching

Questions answering: SQuAD 2.0 (2019)

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	**85.082**	**87.615**
2 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967
3 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	83.536	86.096
4 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043

Questions answering: SQuAD 2.0 (2020)

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	**90.115**	**92.580**
2 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.002	92.425
3 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
4 Jan 23, 2020	albert+transform+verify (ensemble) *qianxin*	89.528	92.059

Cloze-style task: CNN/Daily Mail

Close task: A test where words or pieces of text have been removed, in order for the test-taker to infer them.

In ML we replace nouns by anonymous @tokens to prevent the NN from using world knowledge, and force it to use logical inference based solely on what is written in the text of the task, not what is written in other datasets used for pre-training.

P

(@entity4) if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel " @entity11 " will feature a capable but flawed @entity13 official named @entity14 who " also happens to be a lesbian . " the character is the first gay figure in the official @entity6 -- the movies , television shows , comics and books approved by @entity6 franchise owner @entity22 -- according to @entity24 , editor of " @entity6 "

Q

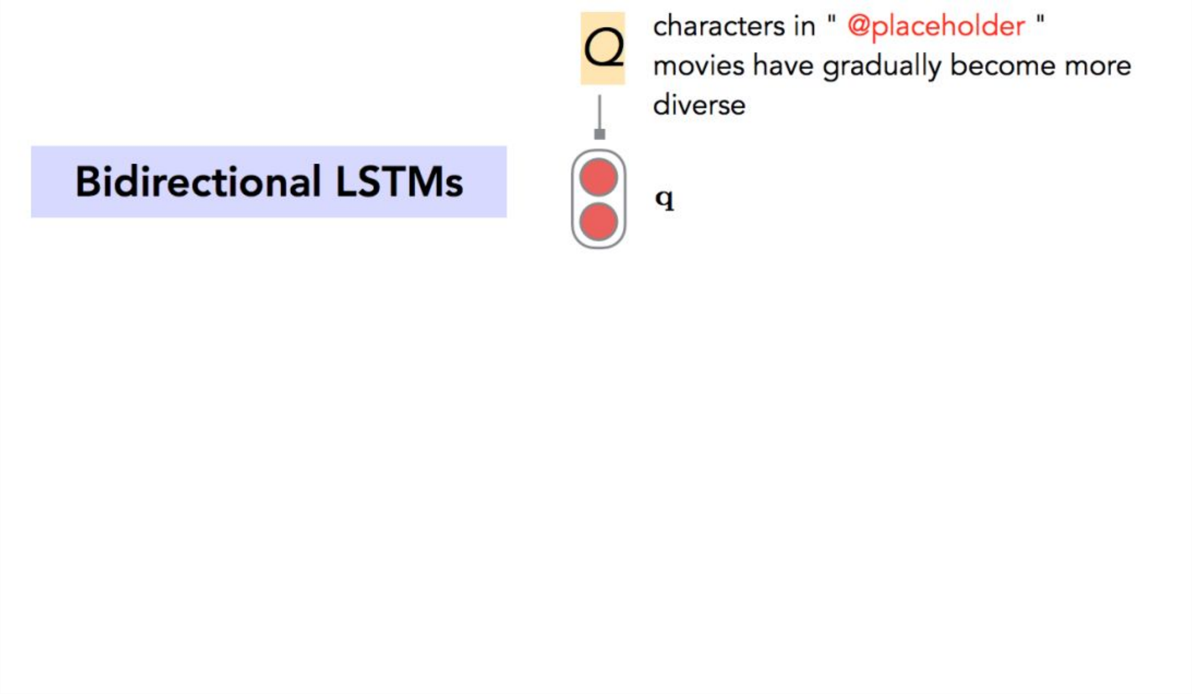
characters in " @placeholder " movies have gradually become more diverse

A

@entity6

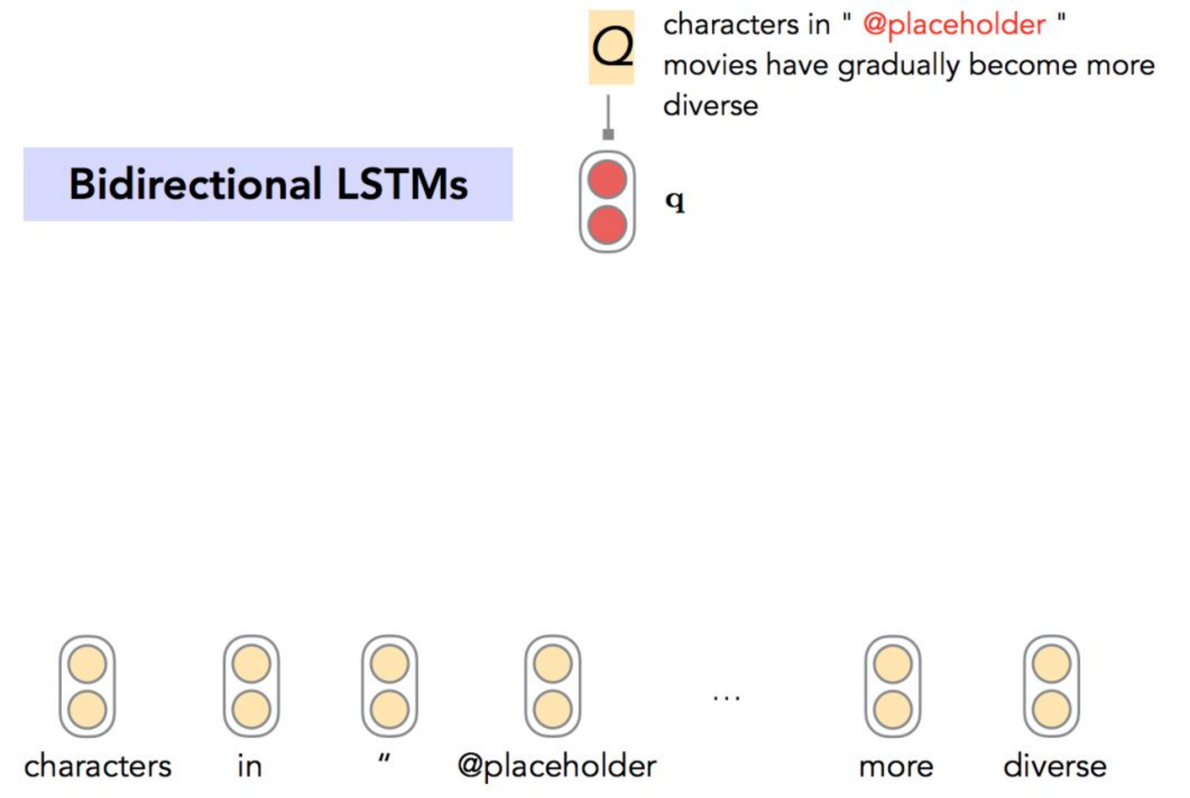
https://www.tensorflow.org/datasets/catalog/cnn_dailymail

CNN/Daily Mail: Stanford Attentive Reader



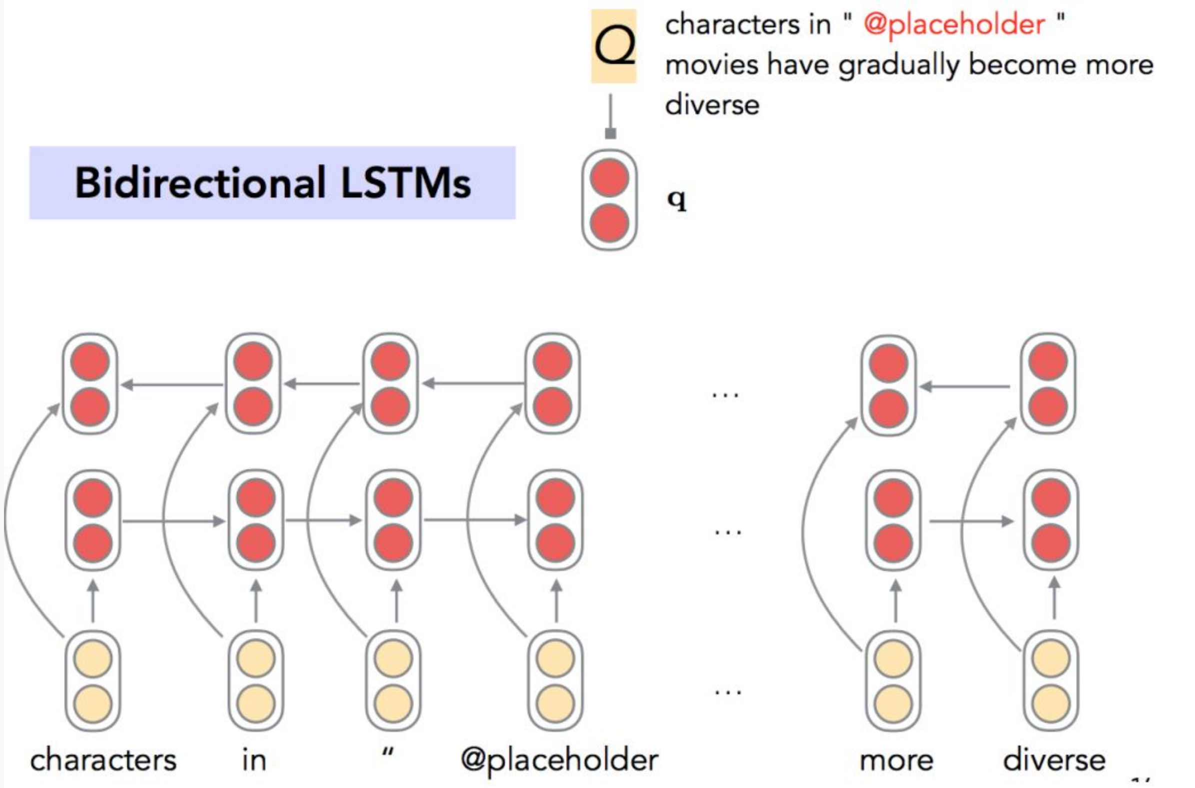
A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task
D Chen, J Bolton & CD Manning, Stanford University [arXiv:1606.02858v2]

CNN/Daily Mail: Stanford Attentive Reader

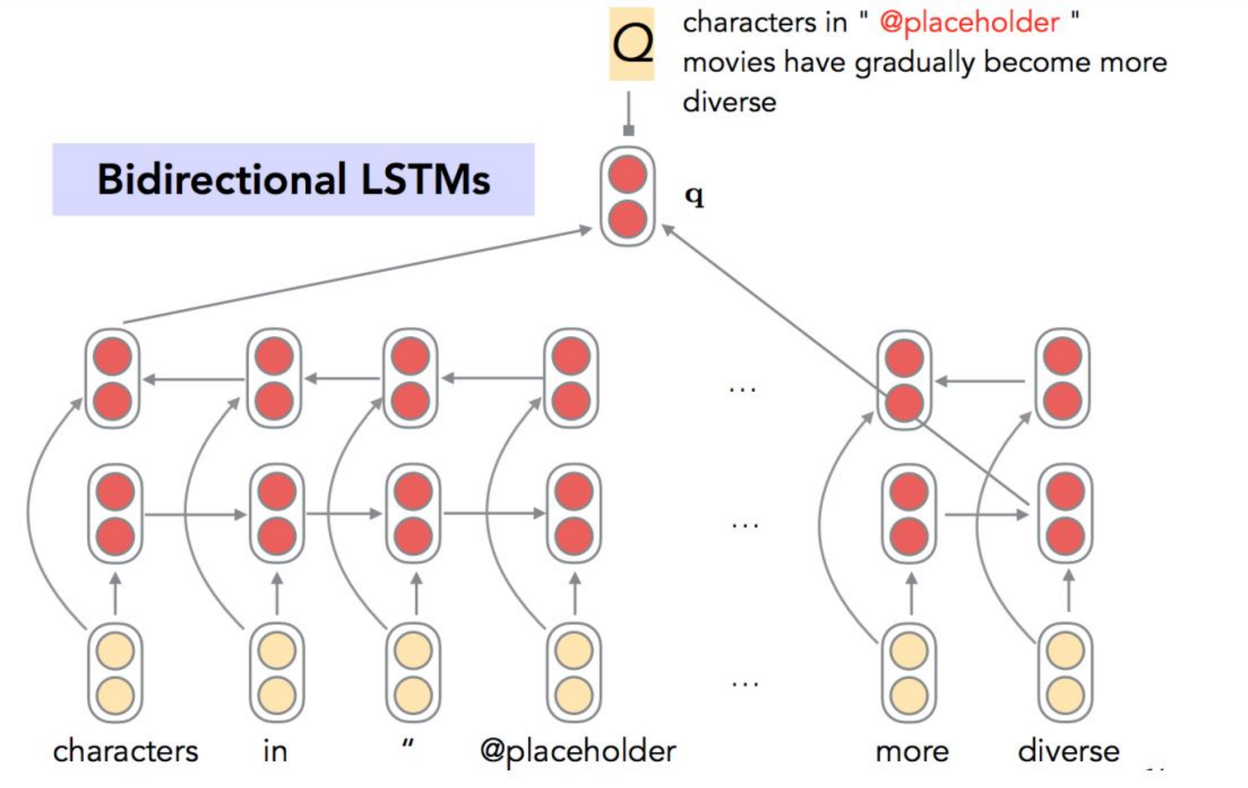


← pre-trained GloVe embeddings

CNN/Daily Mail: Stanford Attentive Reader

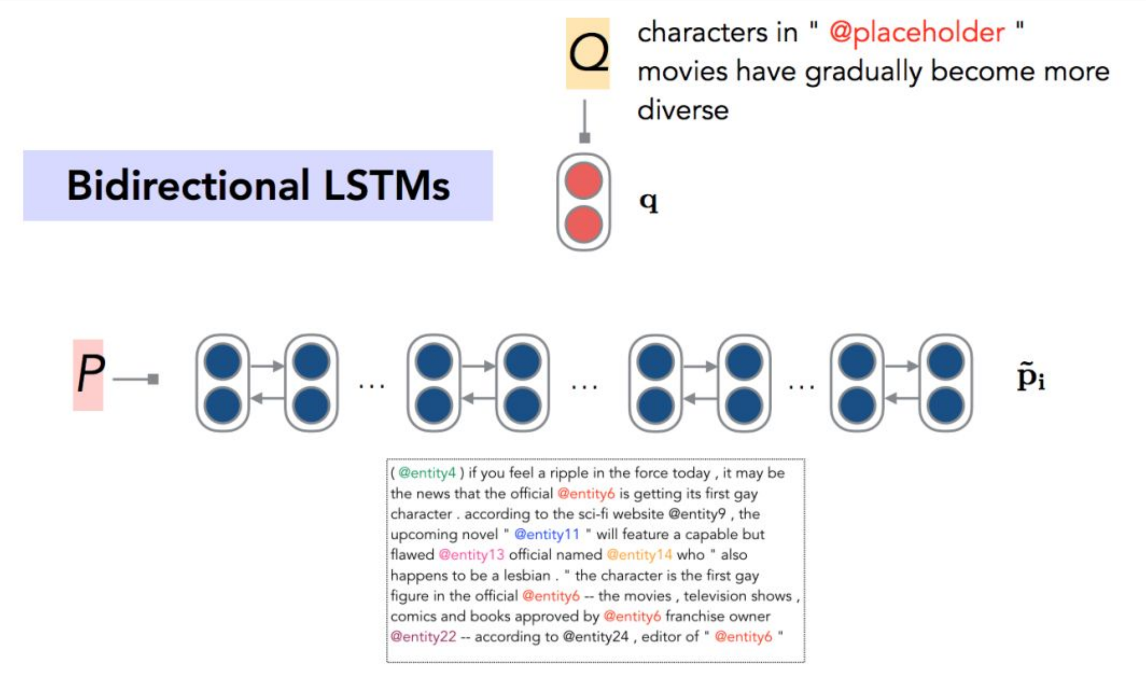


CNN/Daily Mail: Stanford Attentive Reader



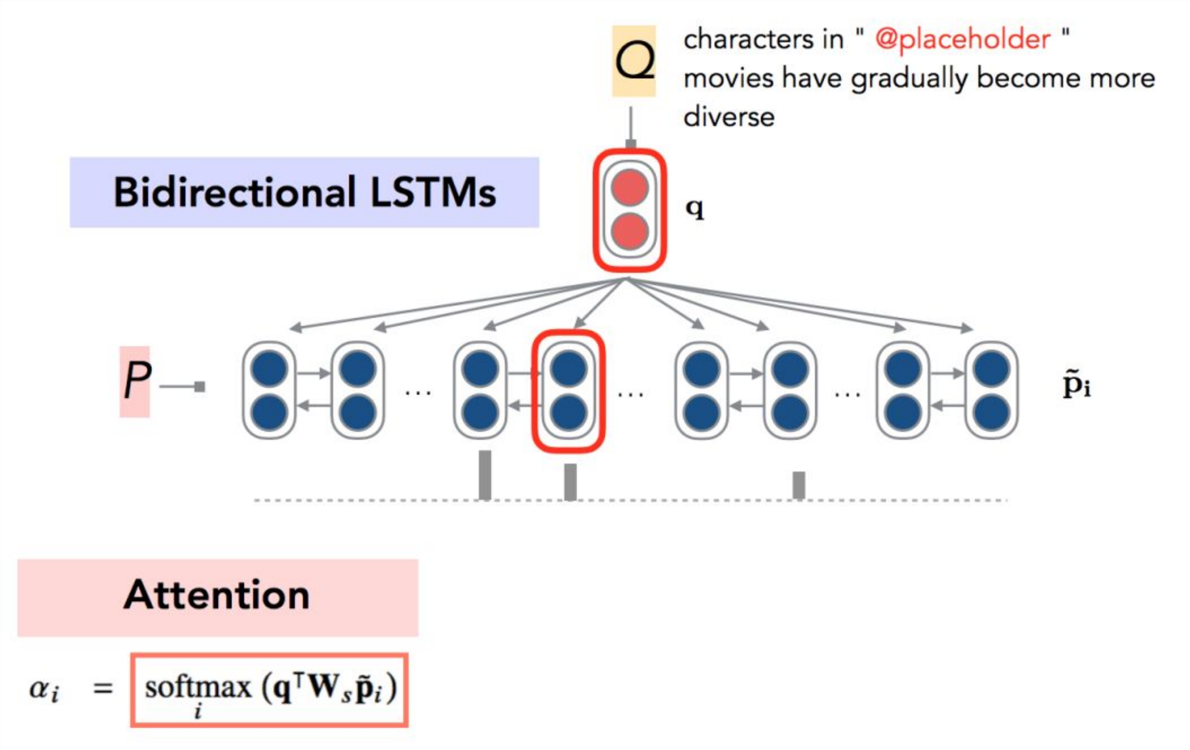
← Take only last LSTM states and concatenate them

CNN/Daily Mail: Stanford Attentive Reader



← This is also a 1-layer bidirectional LSTM (or GRU) on top of GloVe embeddings

CNN/Daily Mail: Stanford Attentive Reader



← Multiplicative attention!

Translation: Attention

- There are several typical versions of attention:
- **Dot-product attention:** values and query are dot-multiplied to obtain attention score

$$\mathbf{e}^{(t)} = [\, \mathbf{s}^{(t)} \cdot \mathbf{h}^{(1)}, \mathbf{s}^{(t)} \cdot \mathbf{h}^{(2)}, \ldots, \mathbf{s}^{(t)} \cdot \mathbf{h}^{(T)} \,] \in \mathbb{R}^T$$

- **Multiplicative attention:** the query is linearly transformed with **W**

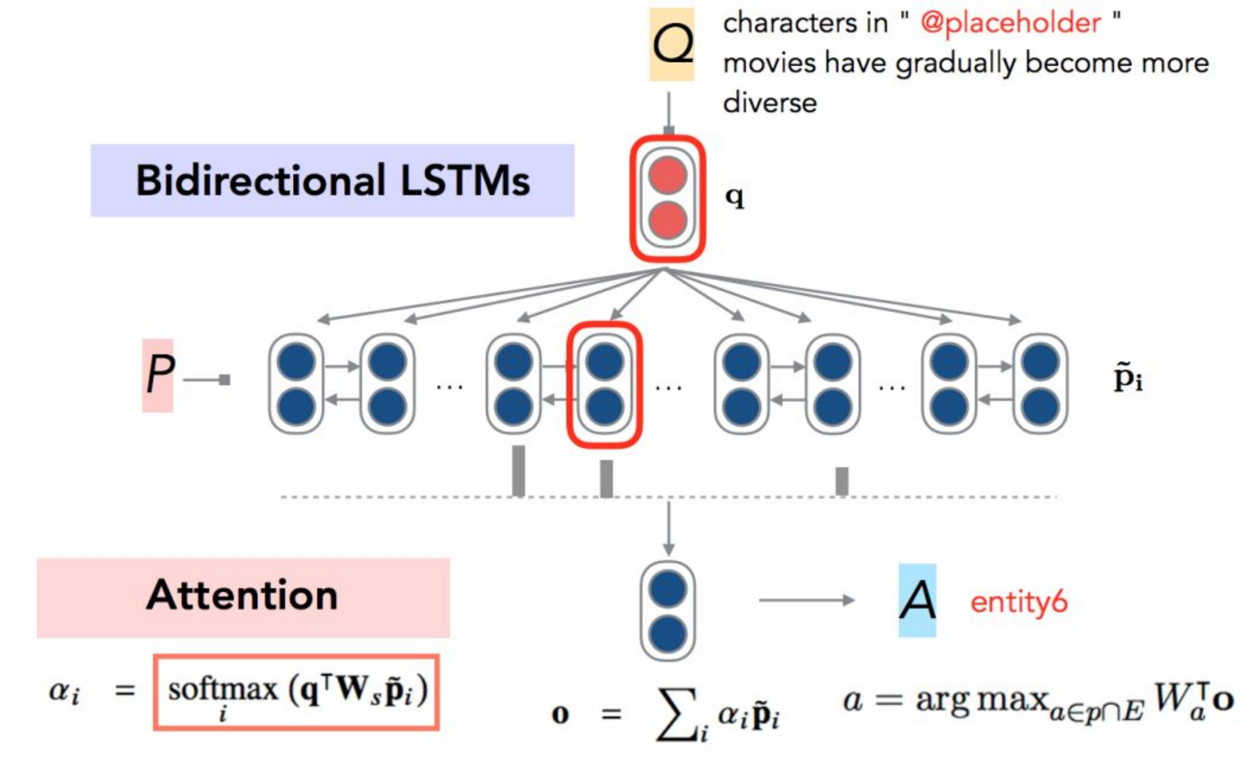
$$\mathbf{e}^{(t)} = [\, \mathbf{s}^{(t)}\, \mathbf{W}\, \mathbf{h}^{(1)}, \mathbf{s}^{(t)}\, \mathbf{W}\, \mathbf{h}^{(2)}, \ldots, \mathbf{s}^{(t)}\, \mathbf{W}\, \mathbf{h}^{(T)} \,] \in \mathbb{R}^T$$

- **Additive attention:** values and query are both linearly transformed by $\mathbf{W}_h$ and $\mathbf{W}_s$, respectively. The result, is averaged with average-weights **v**

$$\mathbf{e}^{(t)} = [\, \mathbf{v} \cdot \tanh(\, \mathbf{W}_s\, \mathbf{s}^{(t)} + \mathbf{W}_h\, \mathbf{h}^{(1)}), \mathbf{v} \cdot \tanh(\, \mathbf{W}_s\, \mathbf{s}^{(t)} + \mathbf{W}_h\, \mathbf{h}^{(2)}), \ldots \,] \in \mathbb{R}^T$$

Deep learning for NLP best practices", Ruder, 2017. http://ruder.io/deep-learning-nlp-best-practices/index.html#attention
Massive exploration of neural machine translation architectures", Britz et al., 2017. https://arxiv.org/pdf/1703.03906.pdf

CNN/Daily Mail: Stanford Attentive Reader



← p ∩ E = all abstract @entities present in the paragraph $\{p_i\}$

Conference resolution: The problem

The problem: Similar to NER. In NER we wanted to identify mentions that refer to any entity of a given class. In conference resolution, in addition to this, we want to identify which mentions refer to exactly the same particular entity.

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

Conference resolution: The problem

The problem: Similar to NER. In NER we wanted to identify all mentions that refer to any entity of a given class. In conference resolution, in addition to this, we want to identify which mentions refer to exactly the same particular entity.

> Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

Conference resolution: The problem

The problem: Similar to NER. In NER we wanted to identify all mentions that refer to any entity of a given class. In conference resolution, in addition to this, we want to identify which mentions refer to exactly the same particular entity.

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

Conference resolution: The problem

The problem: Similar to NER. In NER we wanted to identify all mentions that refer to any entity of a given class. In conference resolution, in addition to this, we want to identify which mentions refer to exactly the same particular entity.

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience

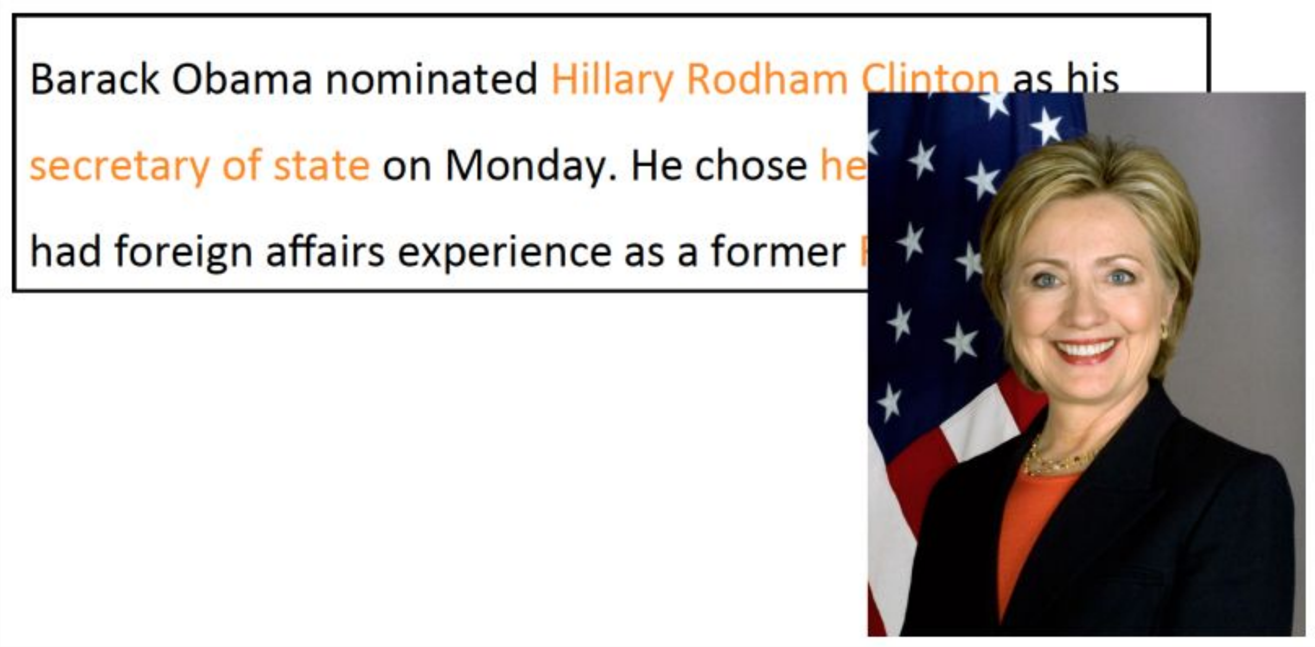
Conference resolution: The problem

The problem: Similar to NER. In NER we wanted to identify all mentions that refer to any entity of a given class. In conference resolution, in addition to this, we want to identify which mentions refer to exactly the same particular entity.

> Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

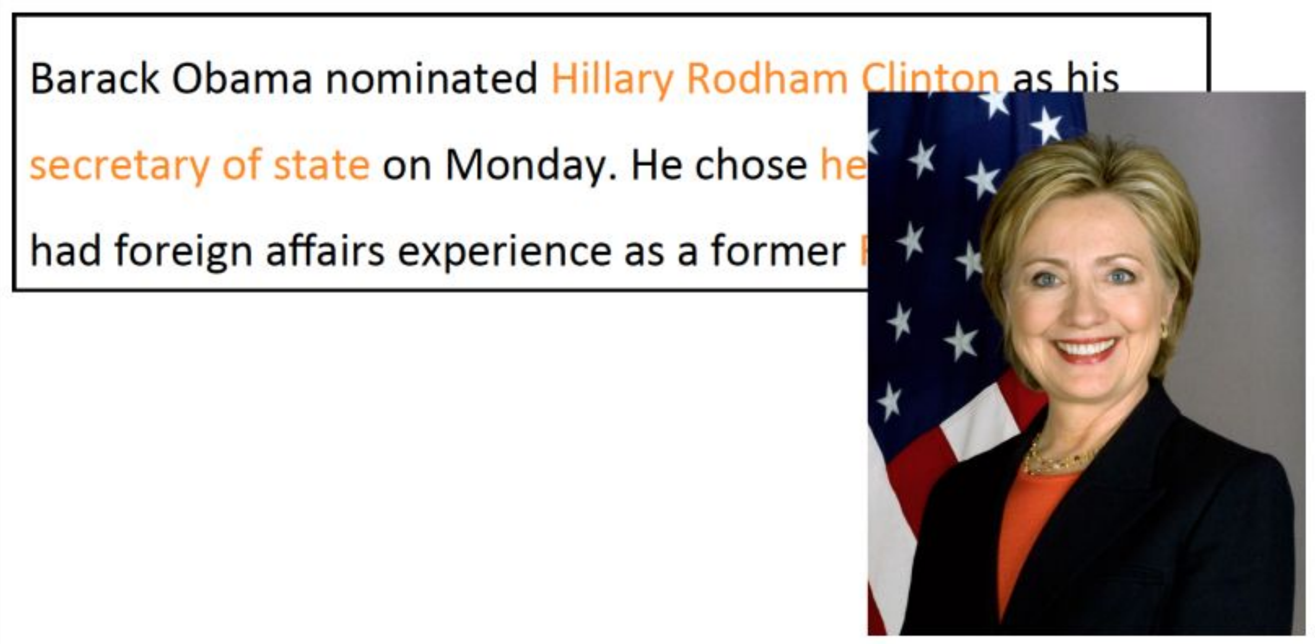
Conference resolution: The problem

The problem: Similar to NER. In NER we wanted to identify all mentions that refer to any entity of a given class. In conference resolution, in addition to this, we want to identify which mentions refer to exactly the same particular entity.



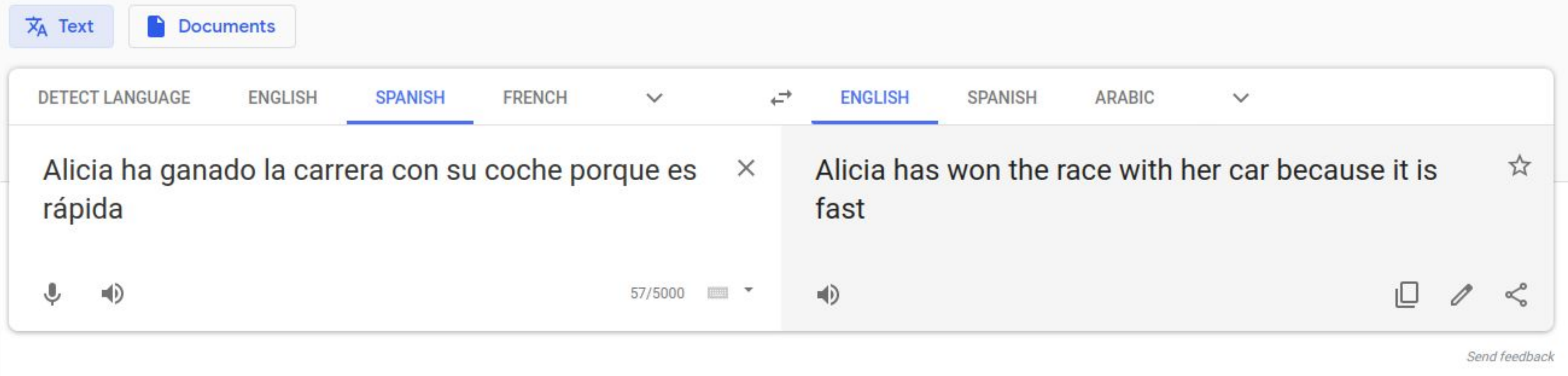
Conference resolution: The problem

The problem: Similar to NER. In NER we wanted to identify all mentions that refer to any entity of a given class. In conference resolution, in addition to this, we want to identify which mentions refer to exactly the same particular entity.



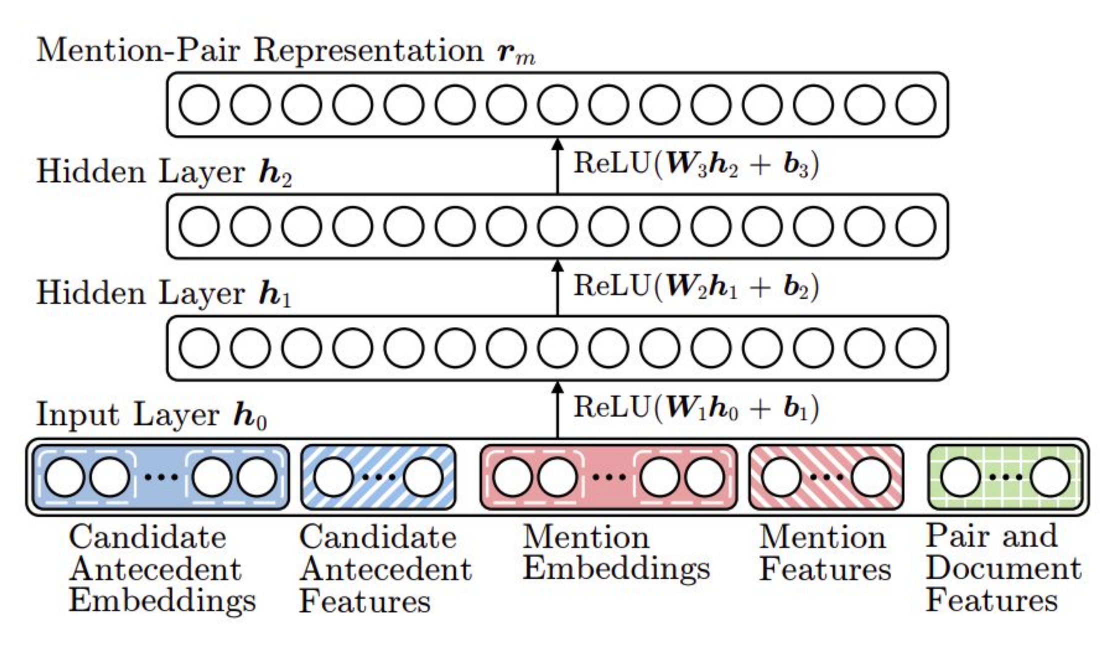
Conference resolution: Why deal with it?

- ➢ **Because Conf Res is awesome**
  - But not as much as QA
- ➢ **Because Conf Res can help many downstream NLP tasks**
  - Information extraction, summarisation, question answering, full text understanding...
- ➢ **Many other improvements and applications left to do**
  - Both challenging problems and low hanging fruit



Conference resolution: The solution

➢ **Solution 1:** Use a NN to model the probability that any pair of words in the text refer to the same entity



➢ **Mention features:**

- Distance
- Part of Speech
- Document class
- Speaker information
- ...

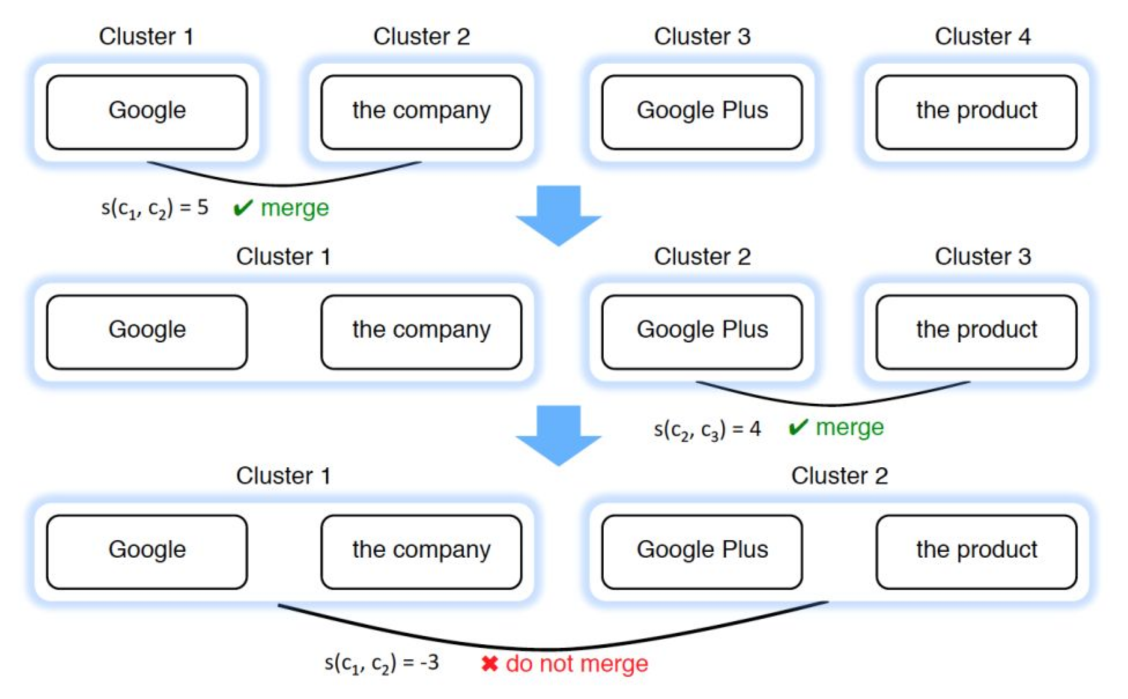
“Improving coreference resolution by learning entity-level distributed representations", Clark et al., 2016. https://arxiv.org/pdf/1606.01323 58/10

Conference resolution: The solution

Solution 2: As in agglomerative clustering, gradually merge clusters of words that refer to the same entity, starting with one cluster per word

“Improving coreference resolution by learning entity-level distributed representations", Clark et al., 2016. https://arxiv.org/pdf/1606.01323

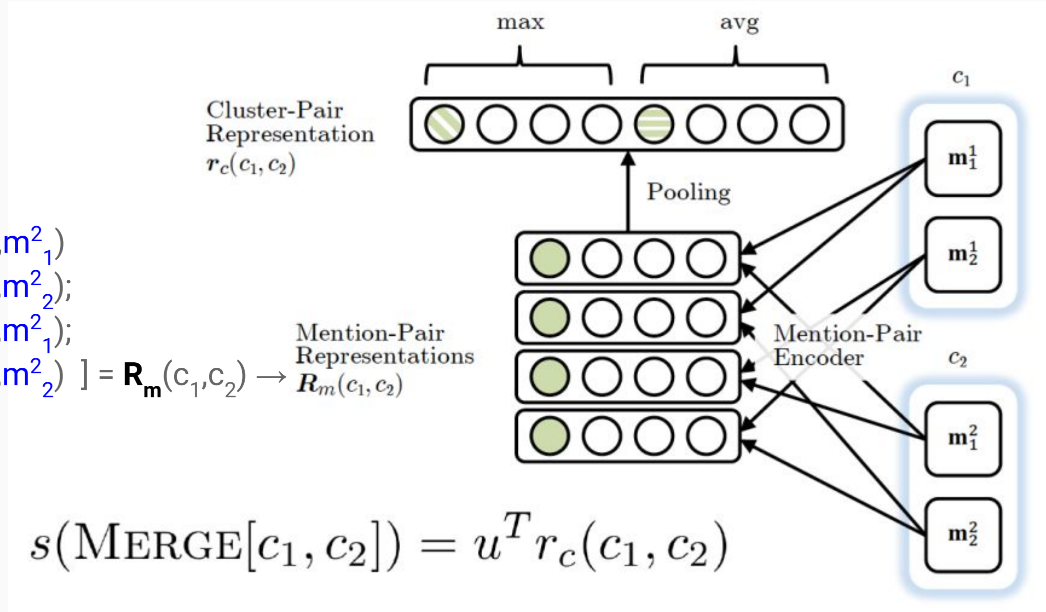
***Google** recently … **the company** announced **Google Plus** … **the product** features*



Conference resolution: The solution

➢ **Solution 2:** As in agglomerative clustering, gradually merge clusters of words that refer to the same entity, starting with one cluster per word

[$\mathbf{r}_m(m^1{}_1, m^2{}_1)$
$\mathbf{r}_m(m^1{}_1, m^2{}_2)$;
$\mathbf{r}_m(m^1{}_2, m^2{}_1)$;
$\mathbf{r}_m(m^1{}_2, m^2{}_2)$] = $\mathbf{R}_m(c_1, c_2) \rightarrow$



$\leftarrow$ ($m^1{}_1$: Google, $m^2{}_1$: Google Plus)

$\leftarrow$ ($m^1{}_1$: Google, $m^2{}_2$: the product)

$\leftarrow$ ($m^1{}_2$: the company, $m^2{}_1$: Google Plus)

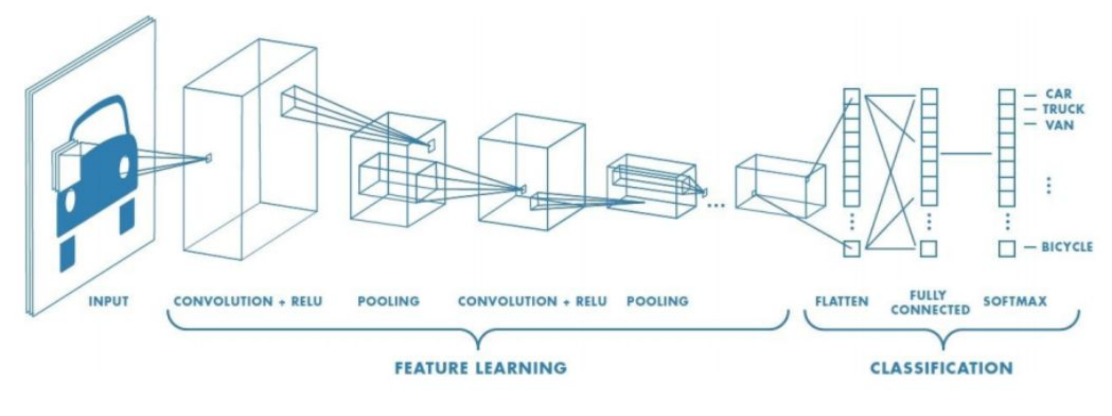
$\leftarrow$ ($m^1{}_2$: the company, $m^2{}_2$: the product)

Conference resolution: The solution

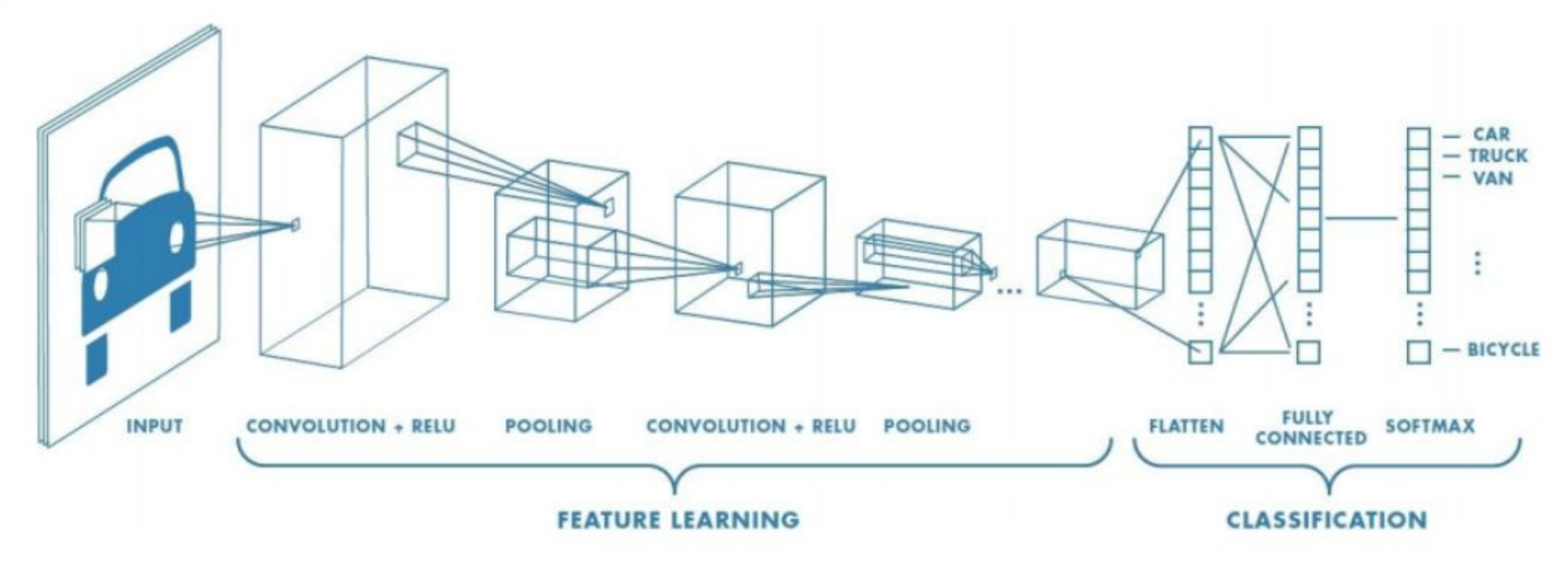
➢ **Solution 3:** Same as before… but with language models based on Transformers. We will see in next lecture

Convolutional NNs: Why a new arch.?

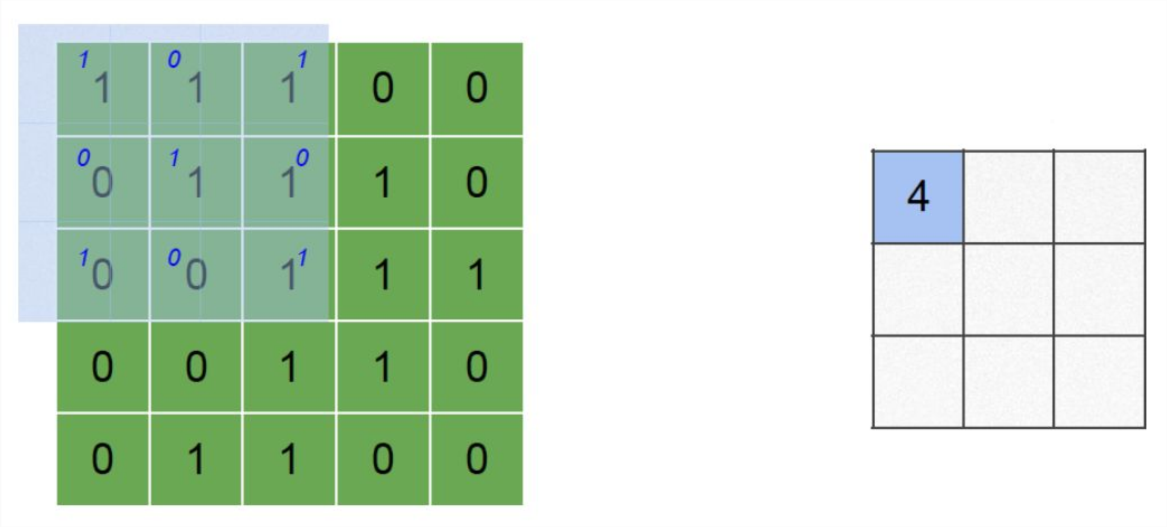
- **The problem:** Traditional NN solutions to NLP problems mostly used some type of RNN (e.g. LSTM, GRU...). Recurrences are however very slow to train
- **The solution:** Convolutional NNs (CNNs) are the standard architecture in vision, where the Convolutions allow them to integrate information from all pixels in an image. Convolutions are much faster to train than Recurrences. We can apply Convolutions in language to integrate information from all words in a document.



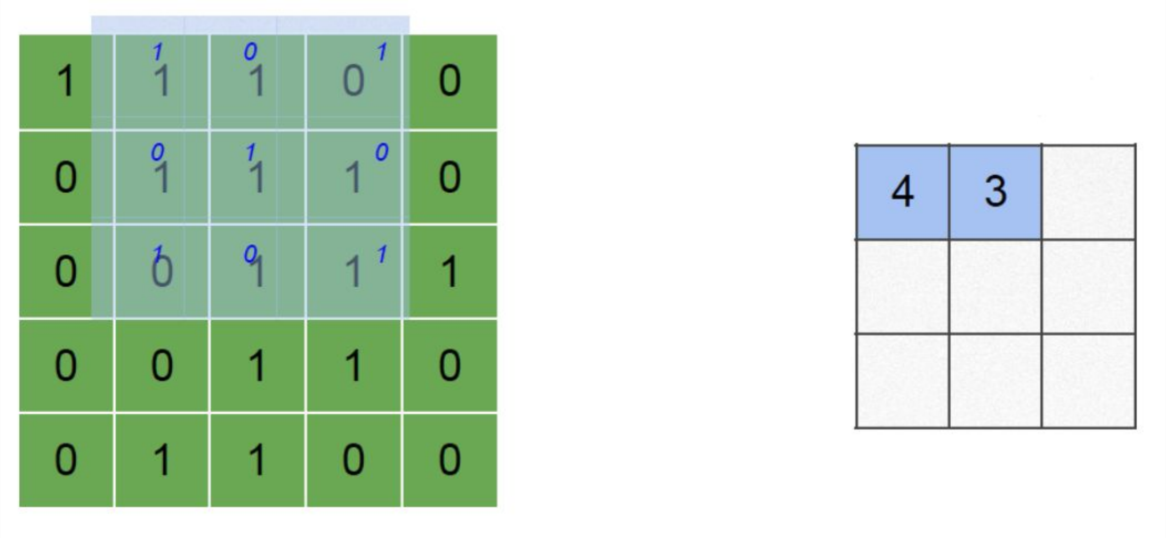
Convolutional NNs: Vision



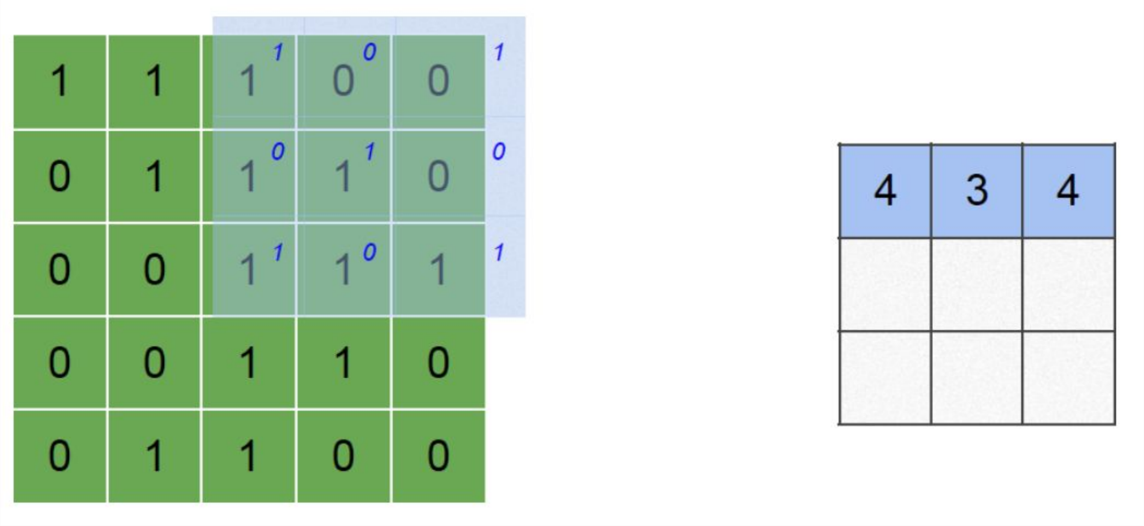
Convolutional NNs: Vision



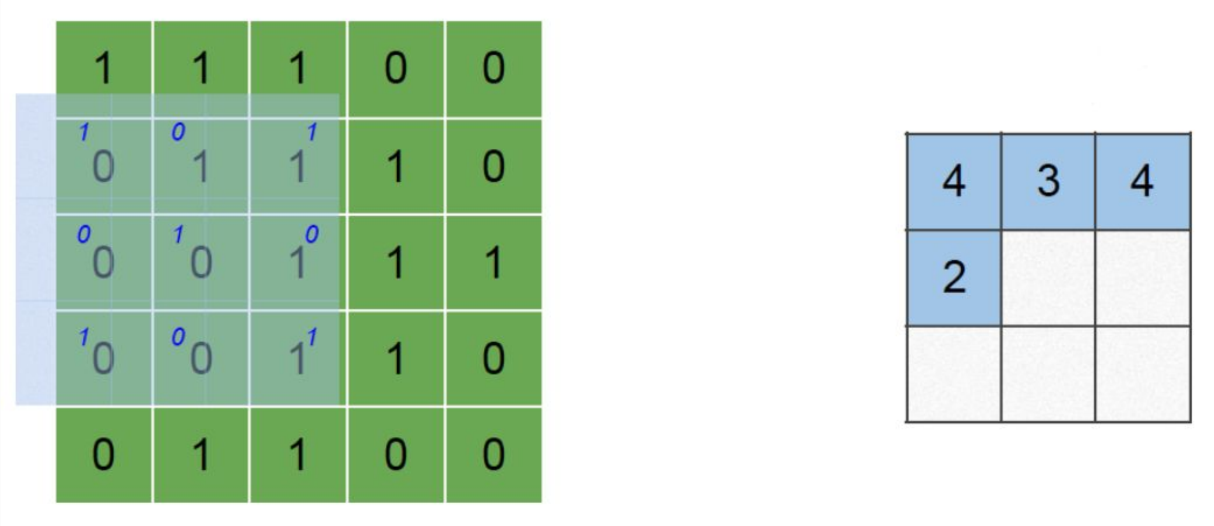
Convolutional NNs: Vision



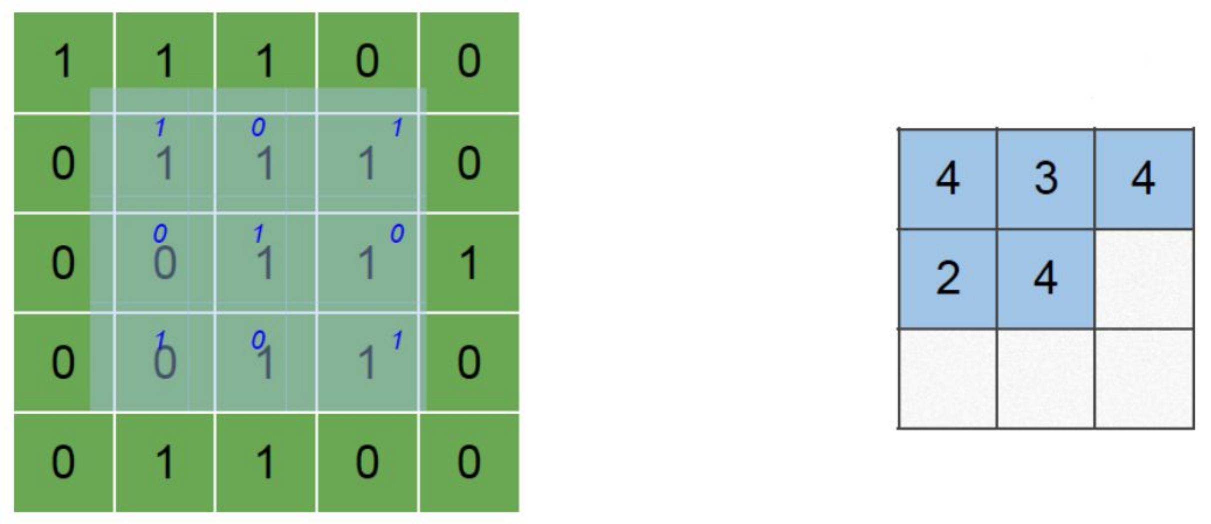
Convolutional NNs: Vision



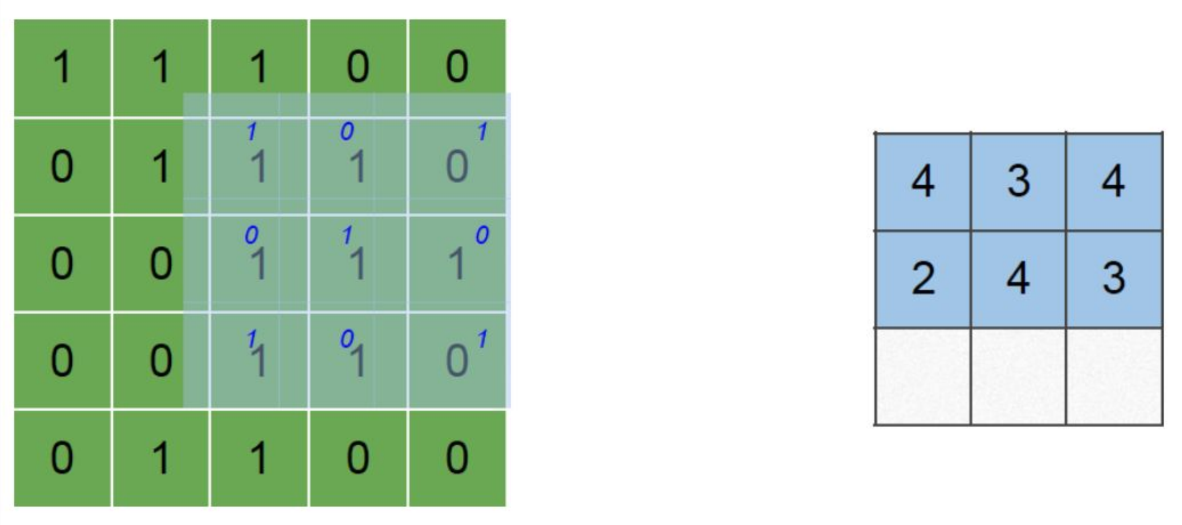
Convolutional NNs: Vision



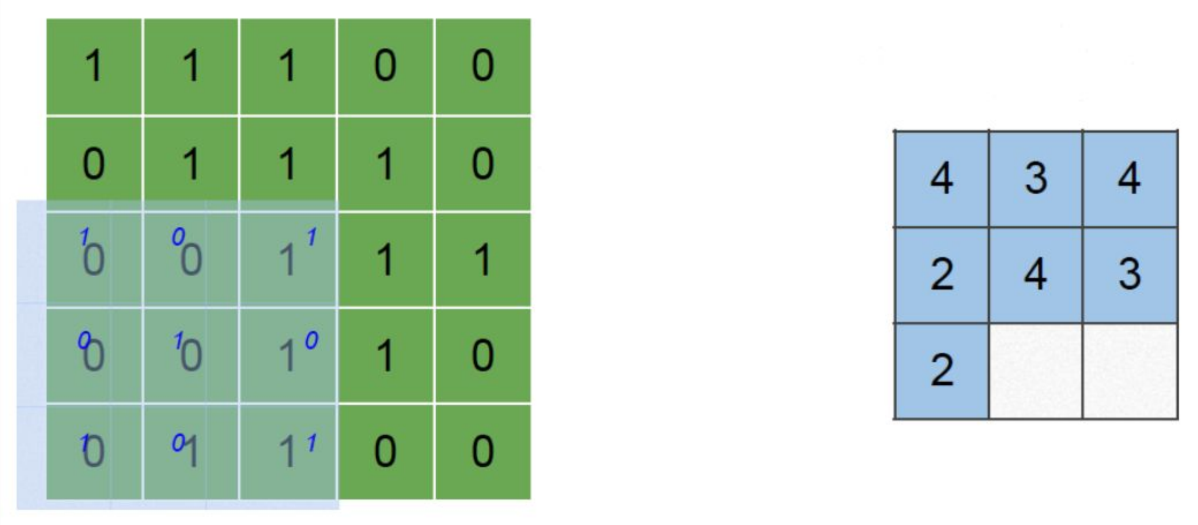
Convolutional NNs: Vision



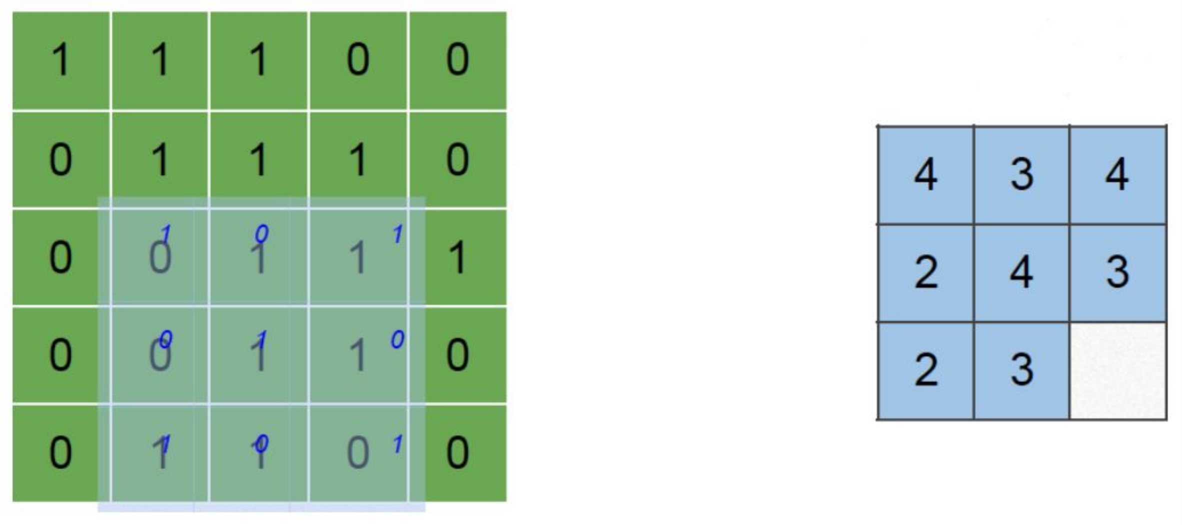
Convolutional NNs: Vision



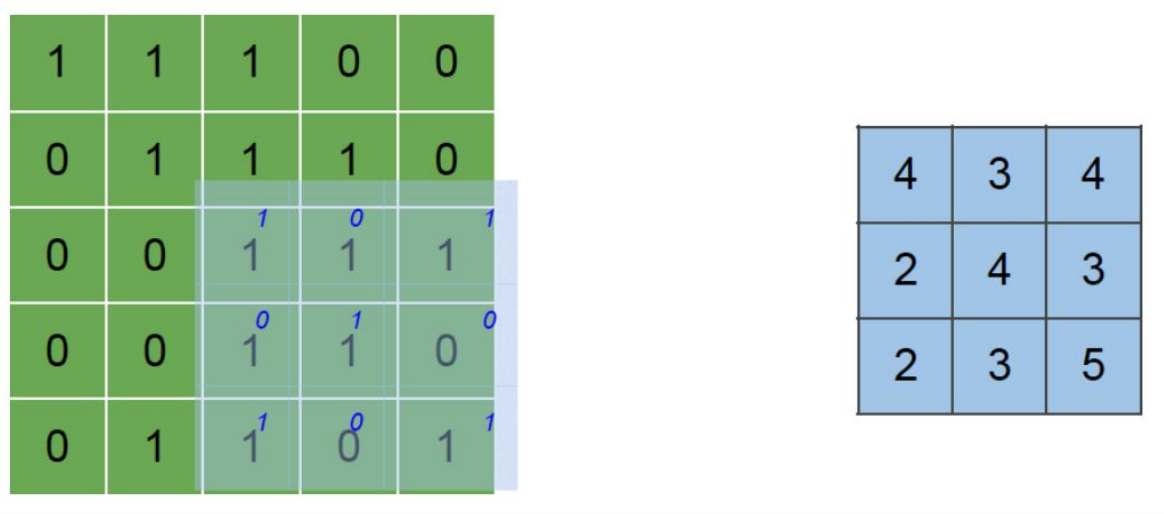
Convolutional NNs: Vision



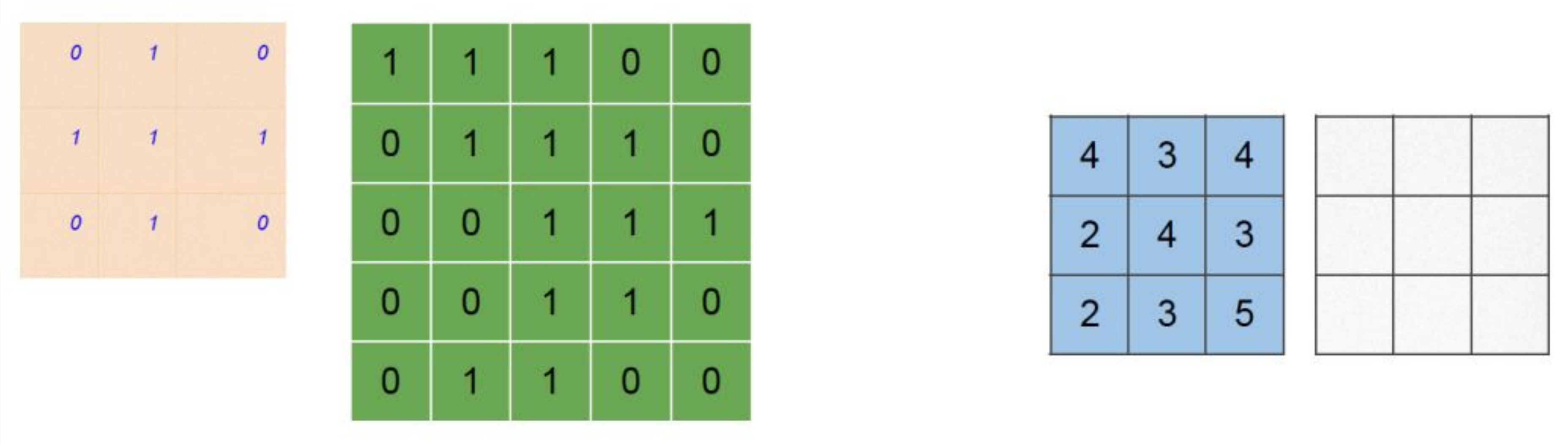
Convolutional NNs: Vision



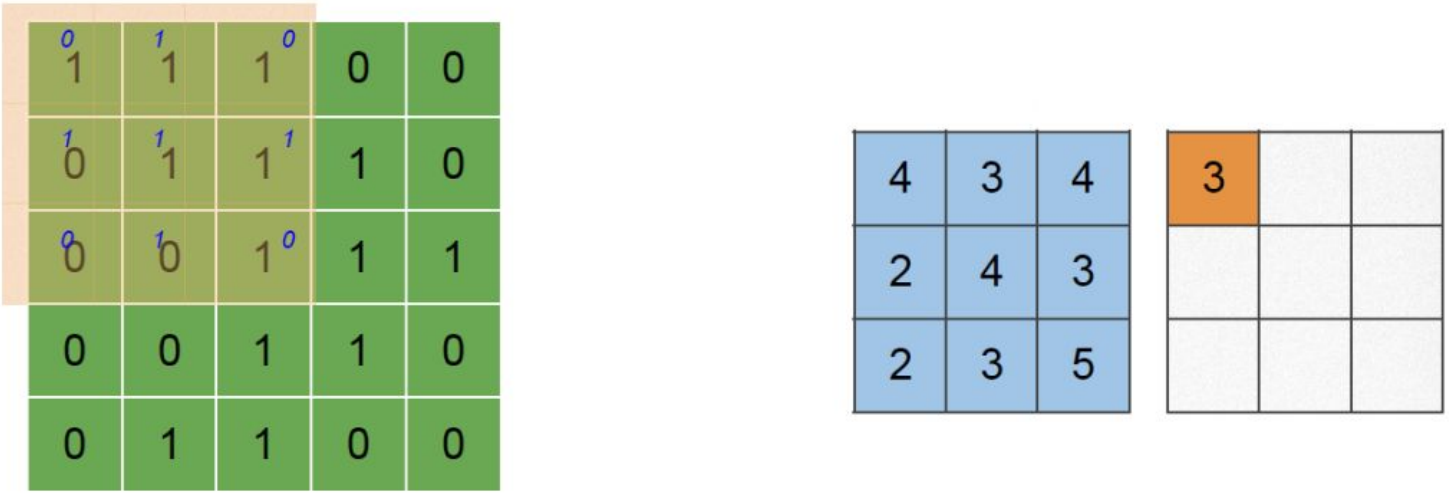
Convolutional NNs: Vision



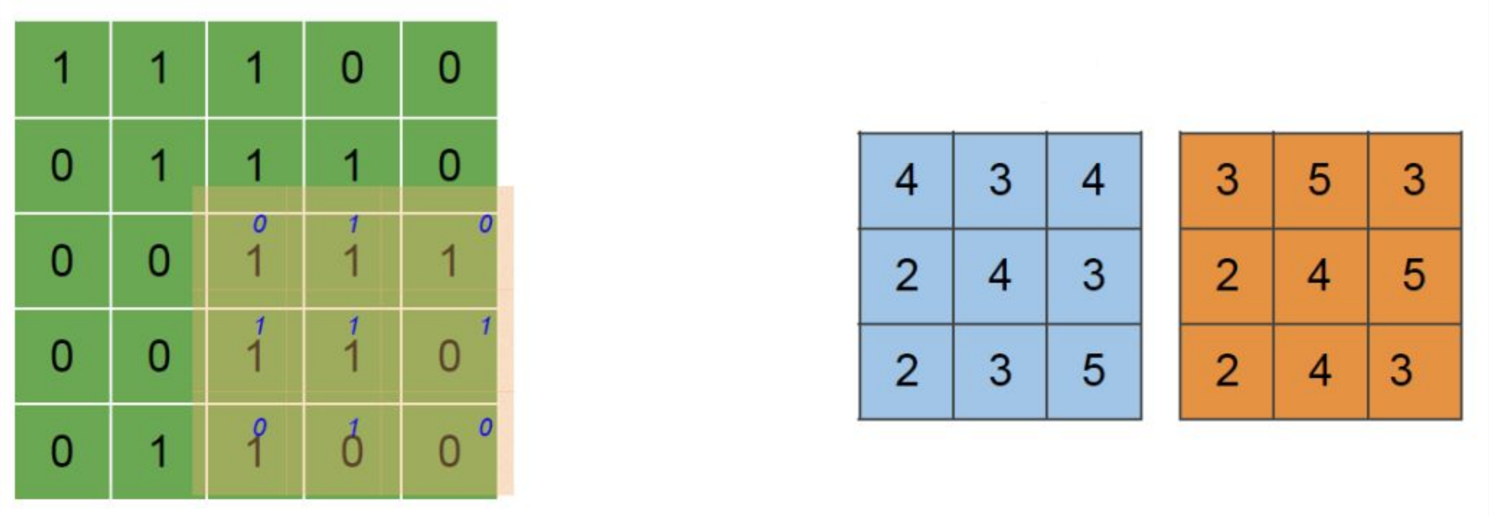
Convolutional NNs: Vision



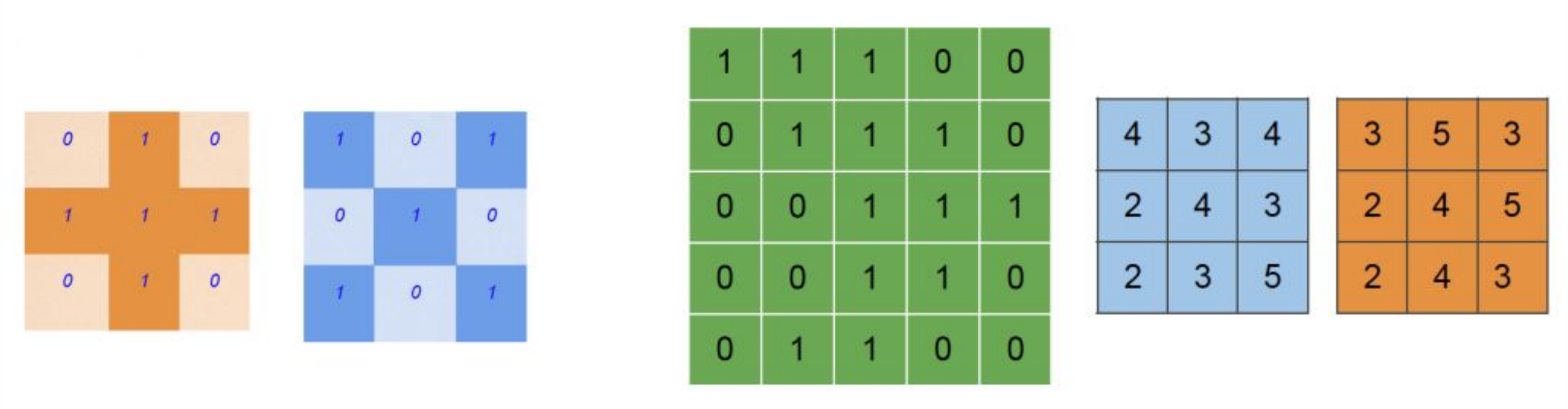
Convolutional NNs: Vision



Convolutional NNs: Vision



Convolutional NNs: Vision



Convolutional NNs: Vision

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

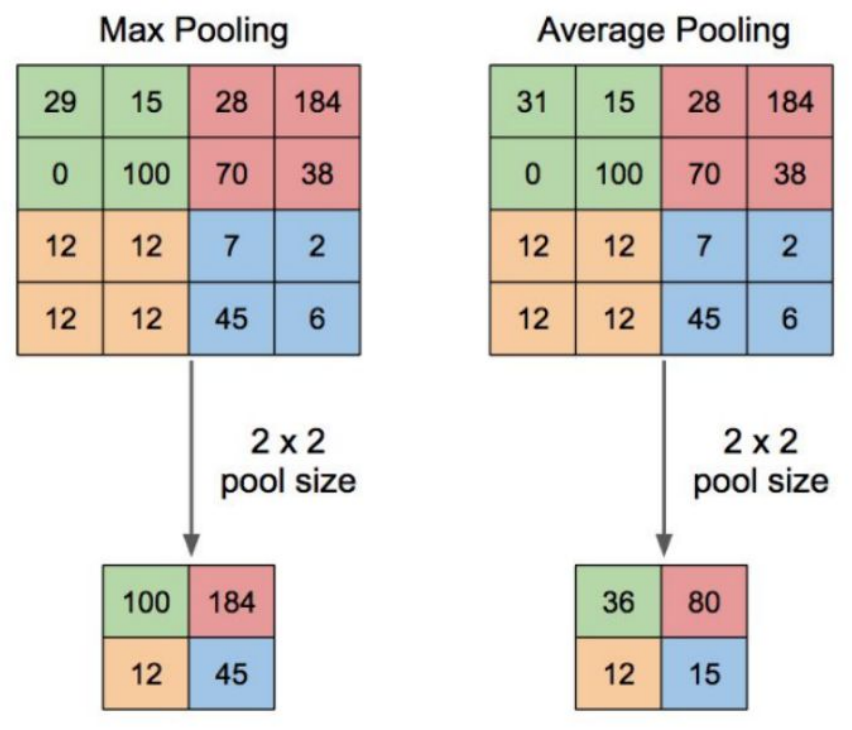
4	3	4
2	4	3
2	3	5

3	5	3
2	4	5
2	4	3

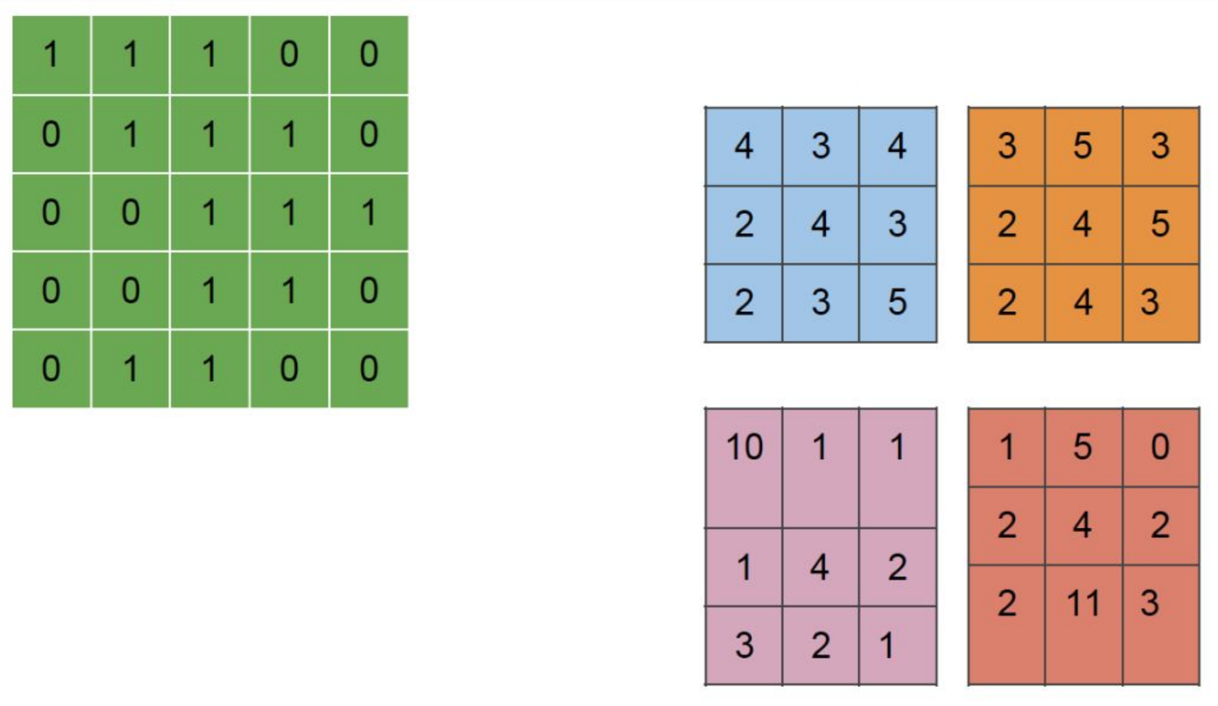
10	1	1
1	4	2
3	2	1

1	5	0
2	4	2
2	11	3

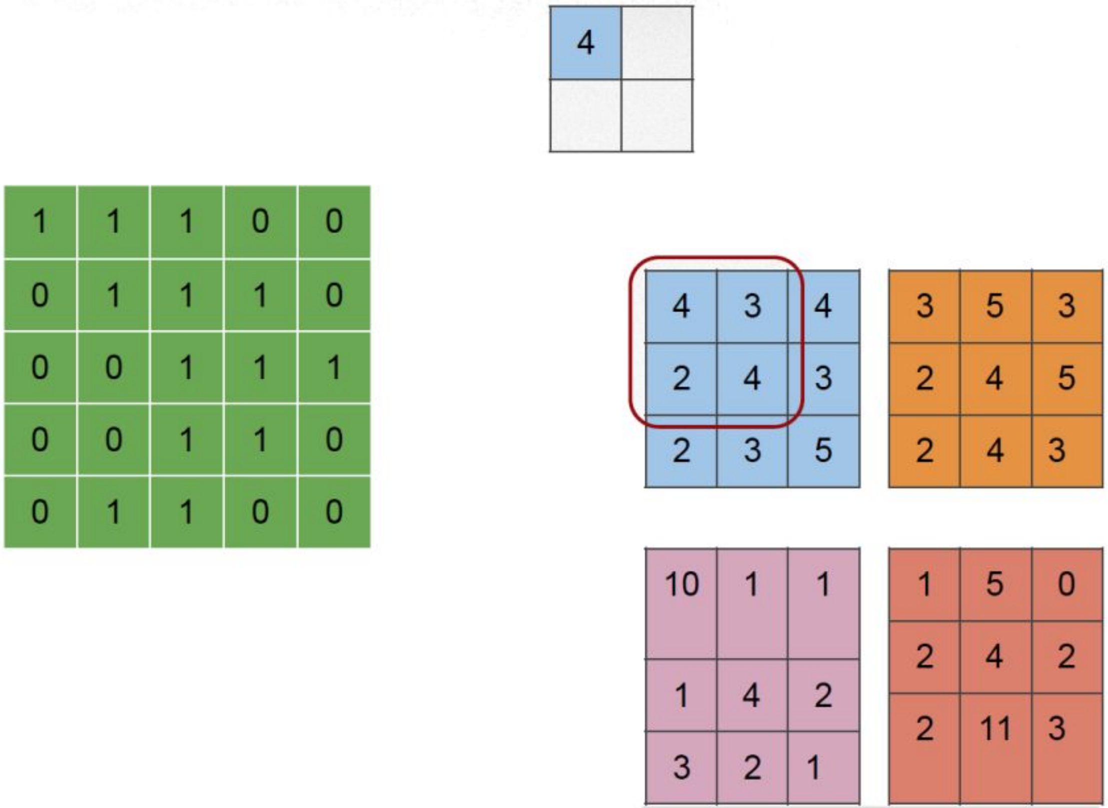
Convolutional NNs: Vision



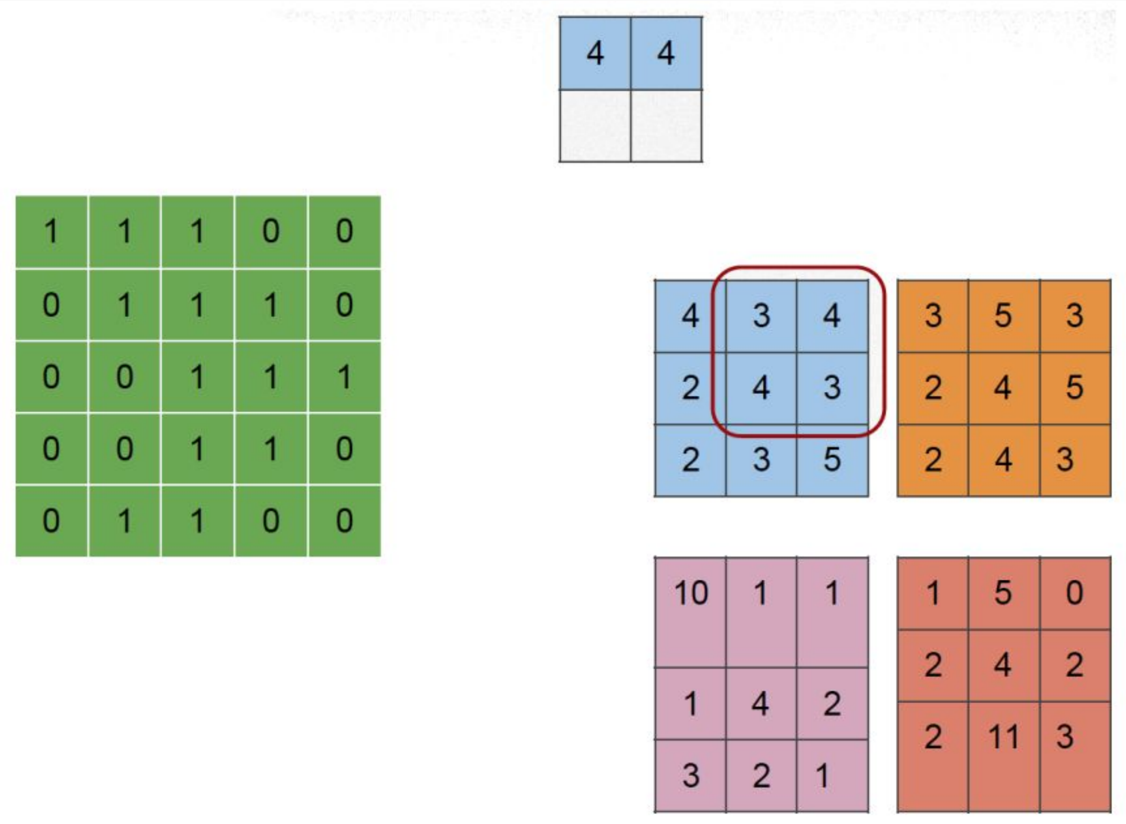
Convolutional NNs: Vision



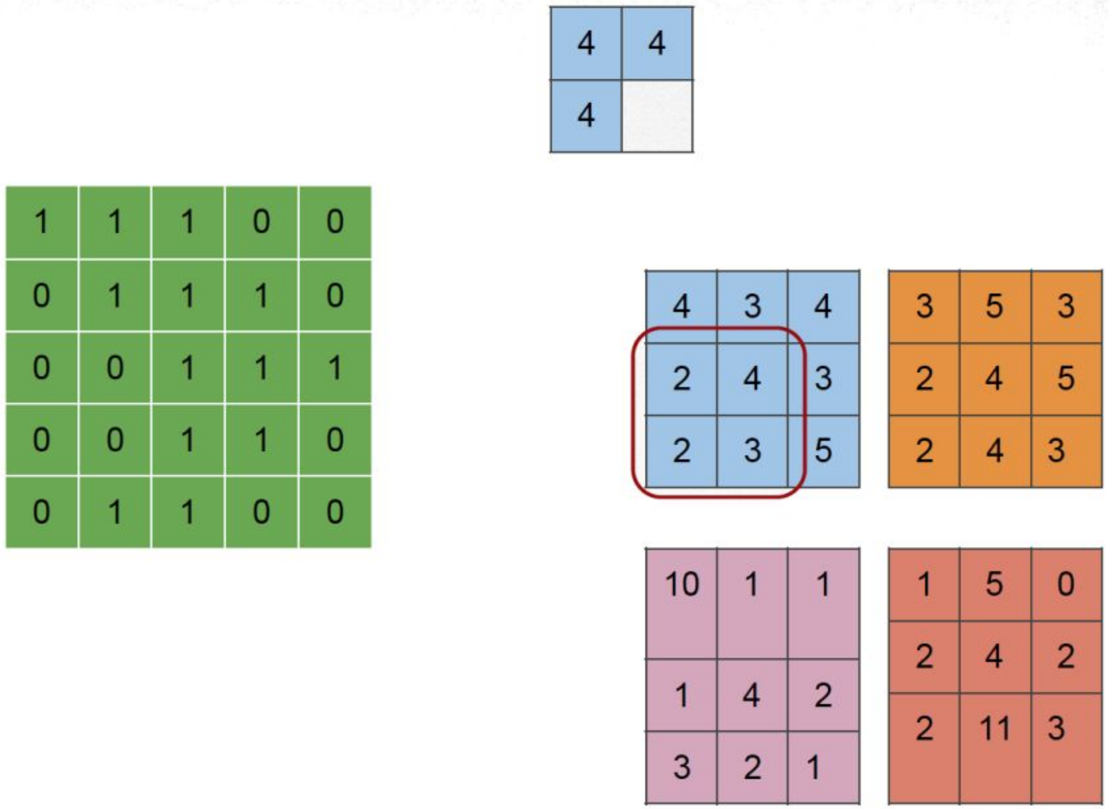
Convolutional NNs: Vision



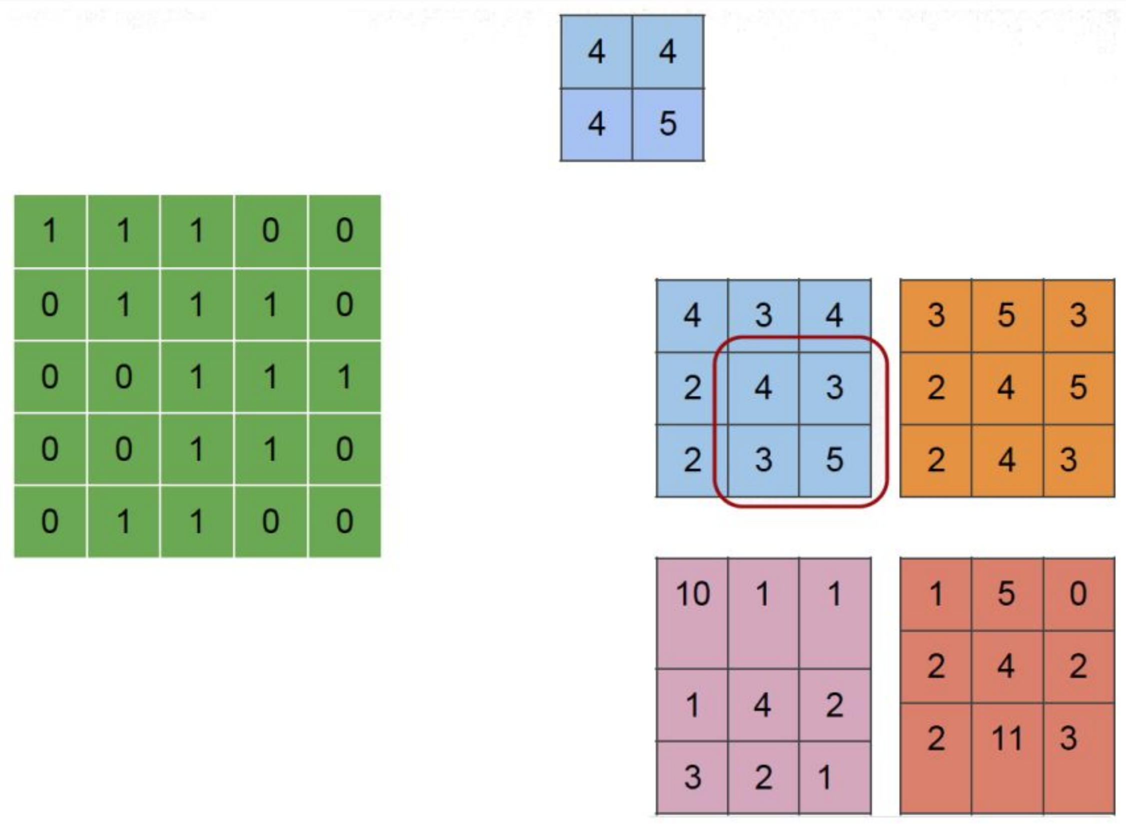
Convolutional NNs: Vision



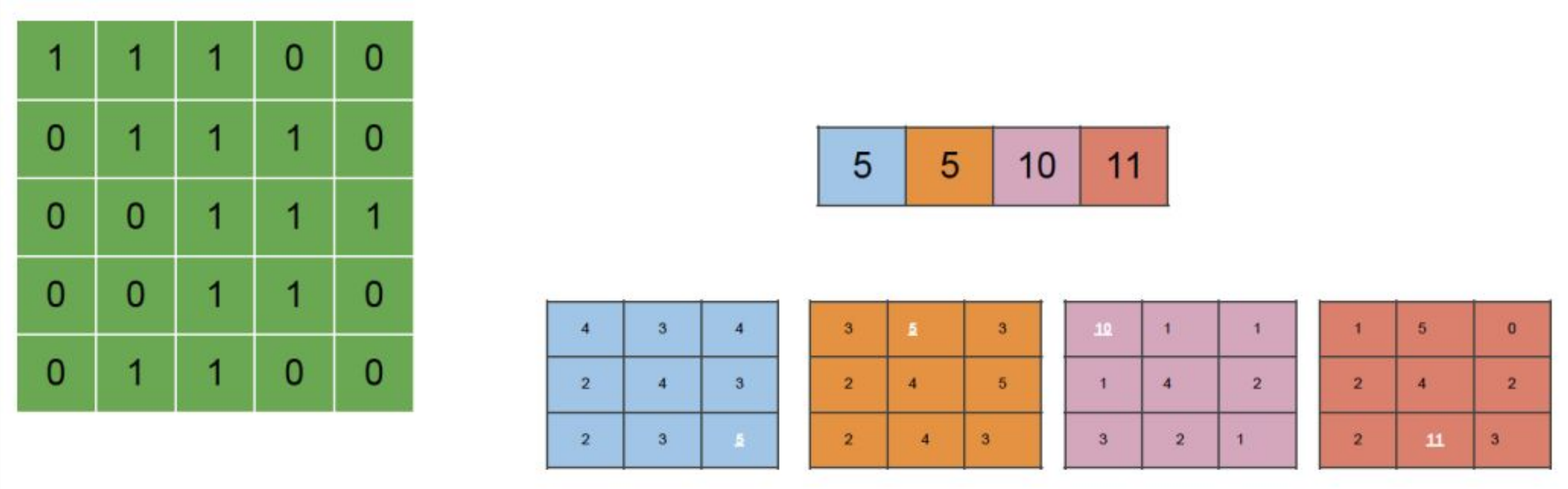
Convolutional NNs: Vision



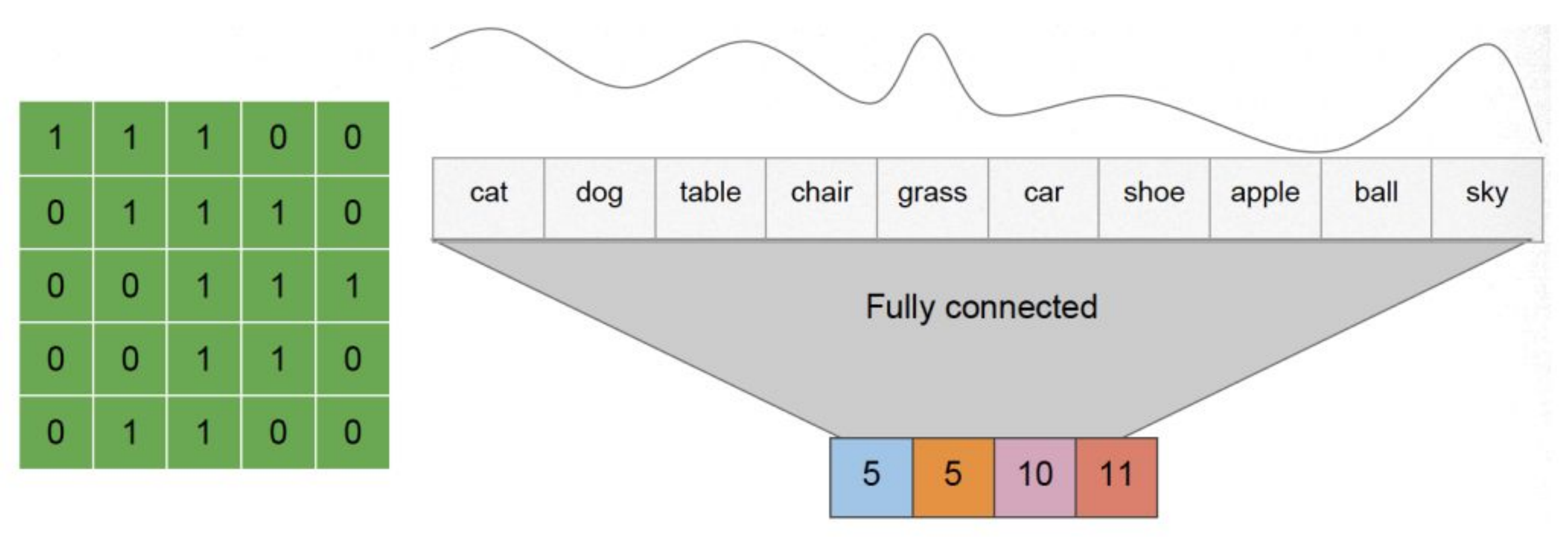
Convolutional NNs: Vision



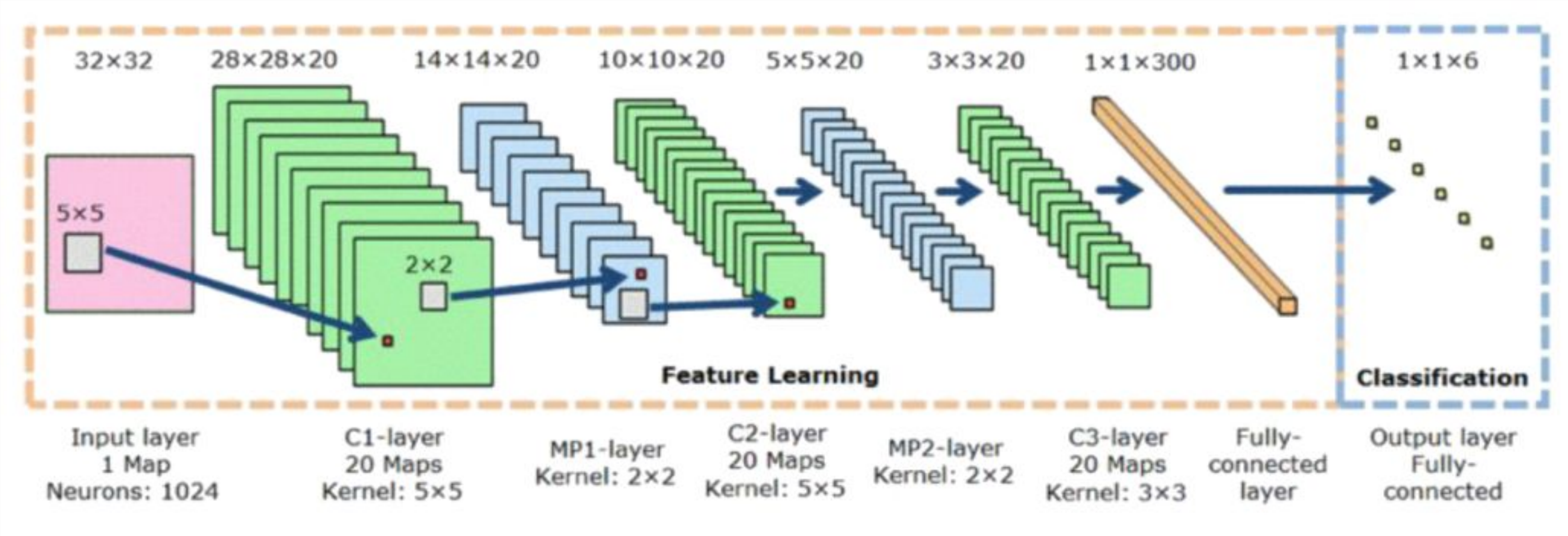
Convolutional NNs: Vision



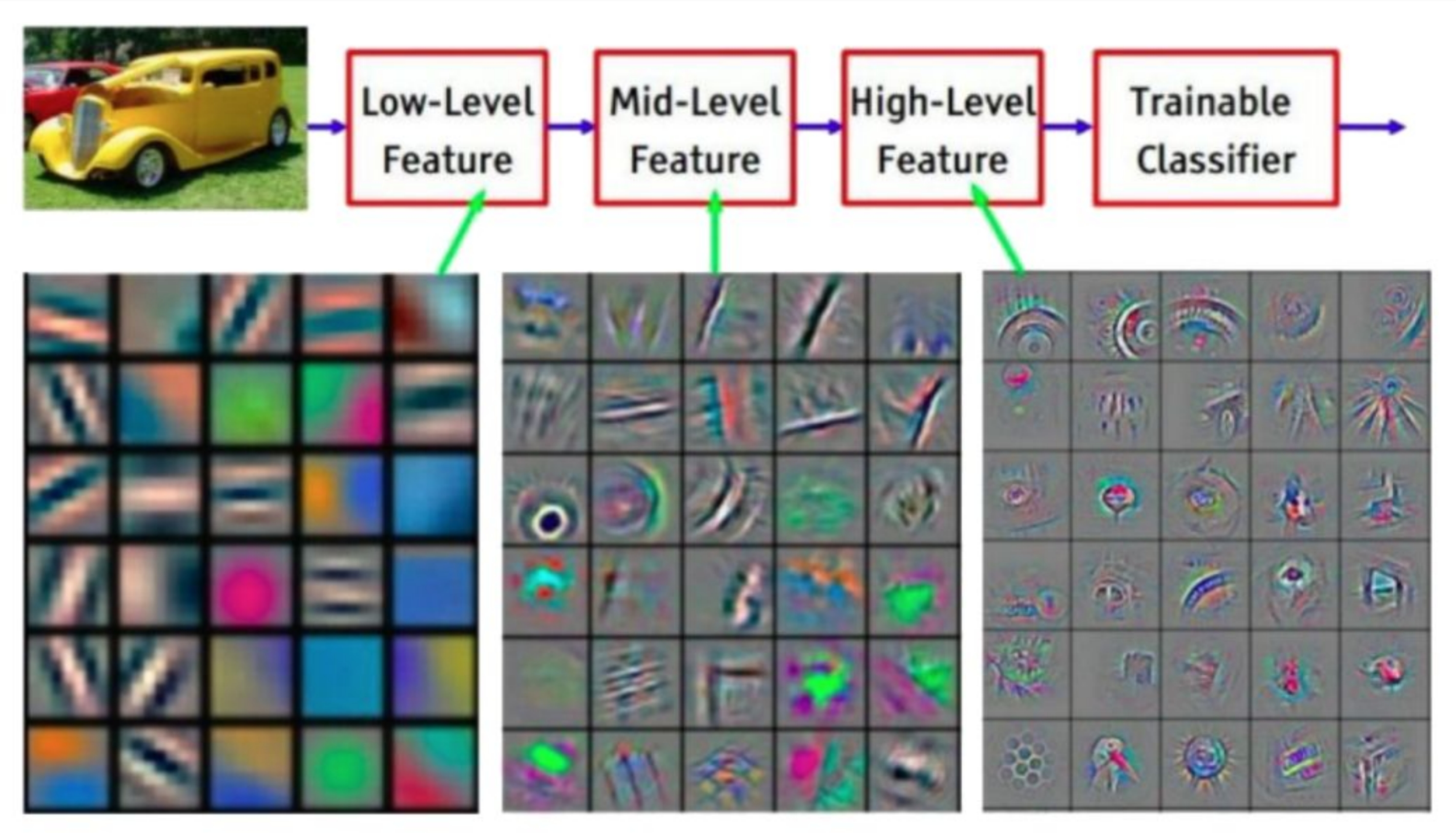
Convolutional NNs: Vision



Convolutional NNs: Vision



Convolutional NNs: Vision



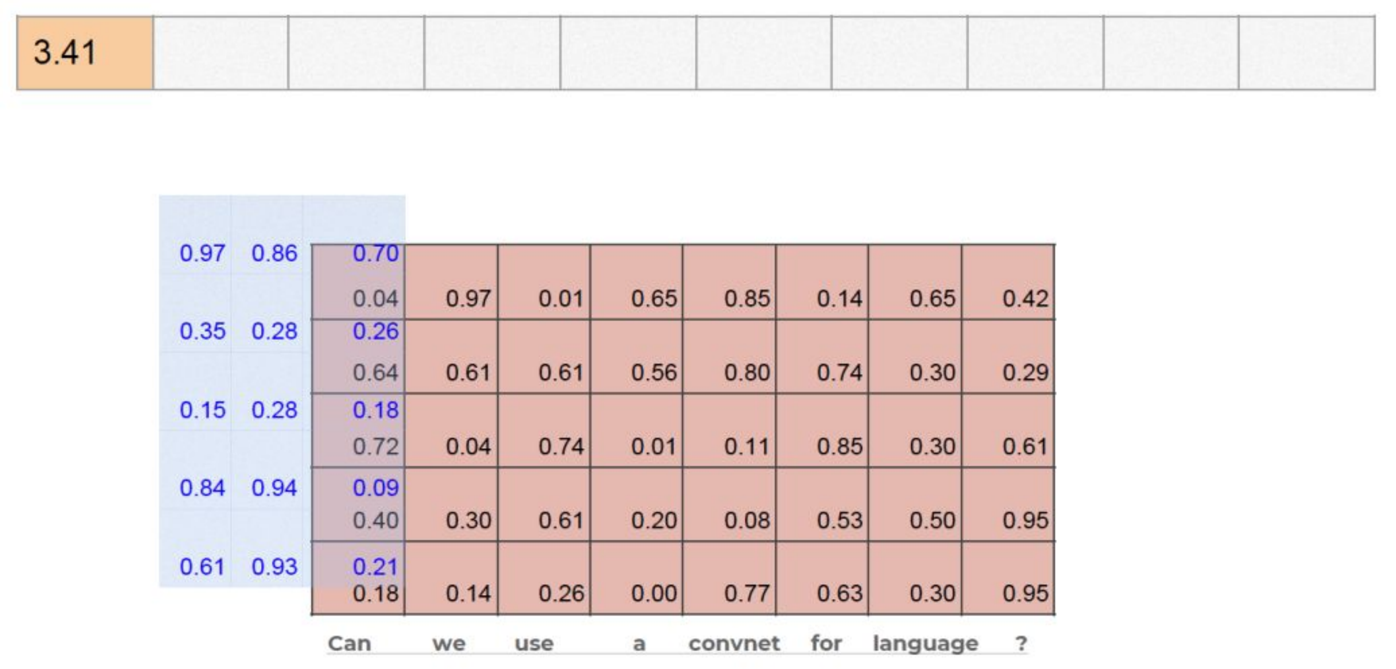
Convolutional NNs: NLP

filter

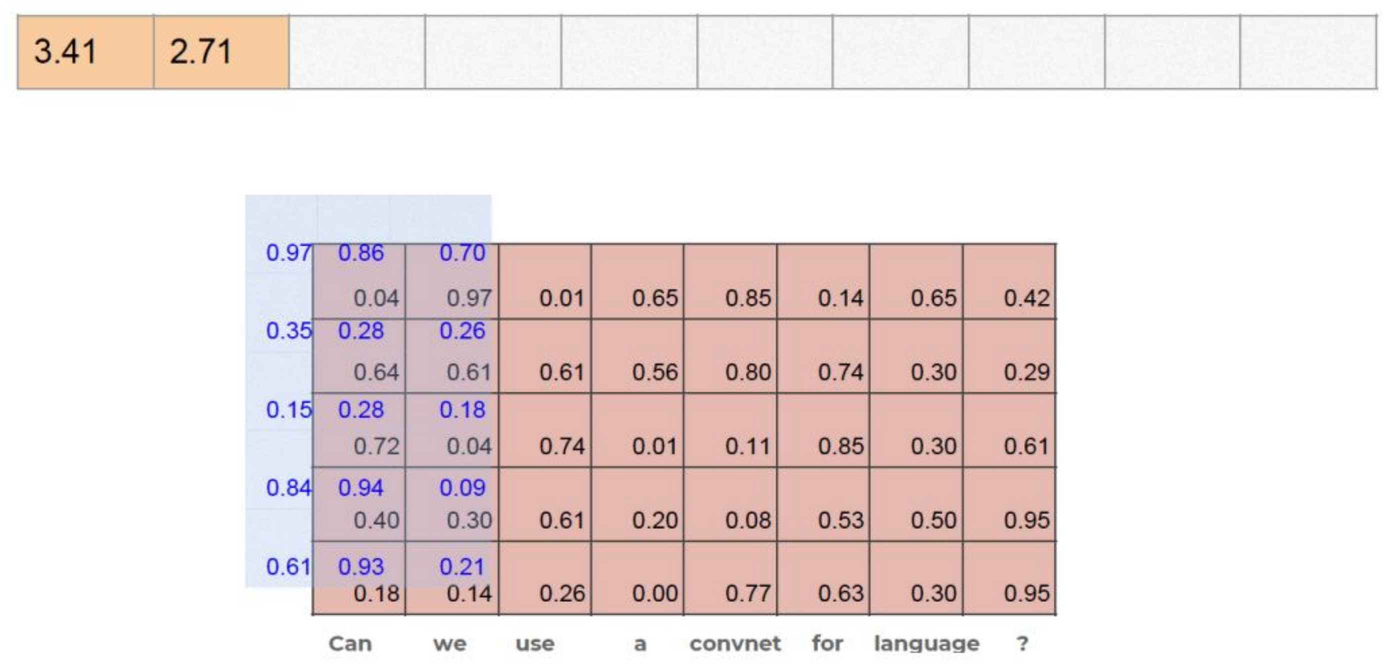
0.97	0.86	0.70
0.35	0.28	0.26
0.15	0.28	0.18
0.84	0.94	0.09
0.61	0.93	0.21

0.04	0.97	0.01	0.65	0.85	0.14	0.65	0.42
0.64	0.61	0.61	0.56	0.80	0.74	0.30	0.29
0.72	0.04	0.74	0.01	0.11	0.85	0.30	0.61
0.40	0.30	0.61	0.20	0.08	0.53	0.50	0.95
0.18	0.14	0.26	0.00	0.77	0.63	0.30	0.95
Can	we	use	a	convnet	for	language	?

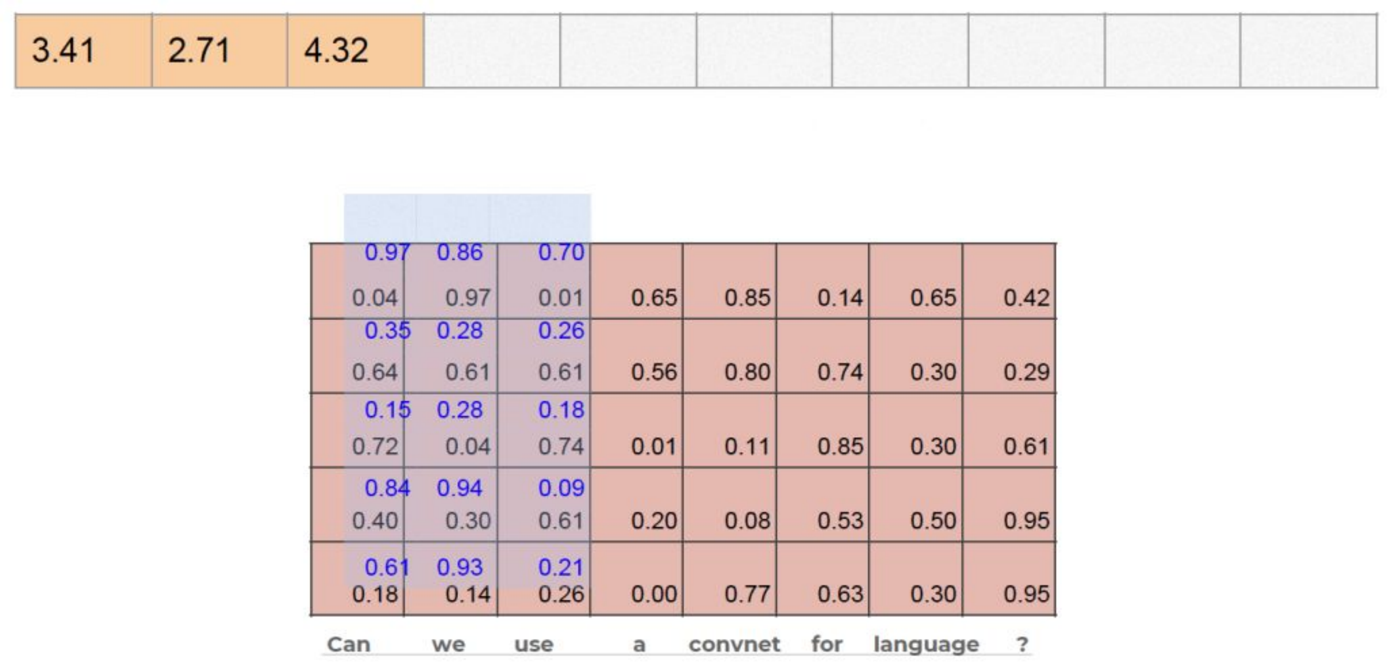
Convolutional NNs: NLP



Convolutional NNs: NLP



Convolutional NNs: NLP



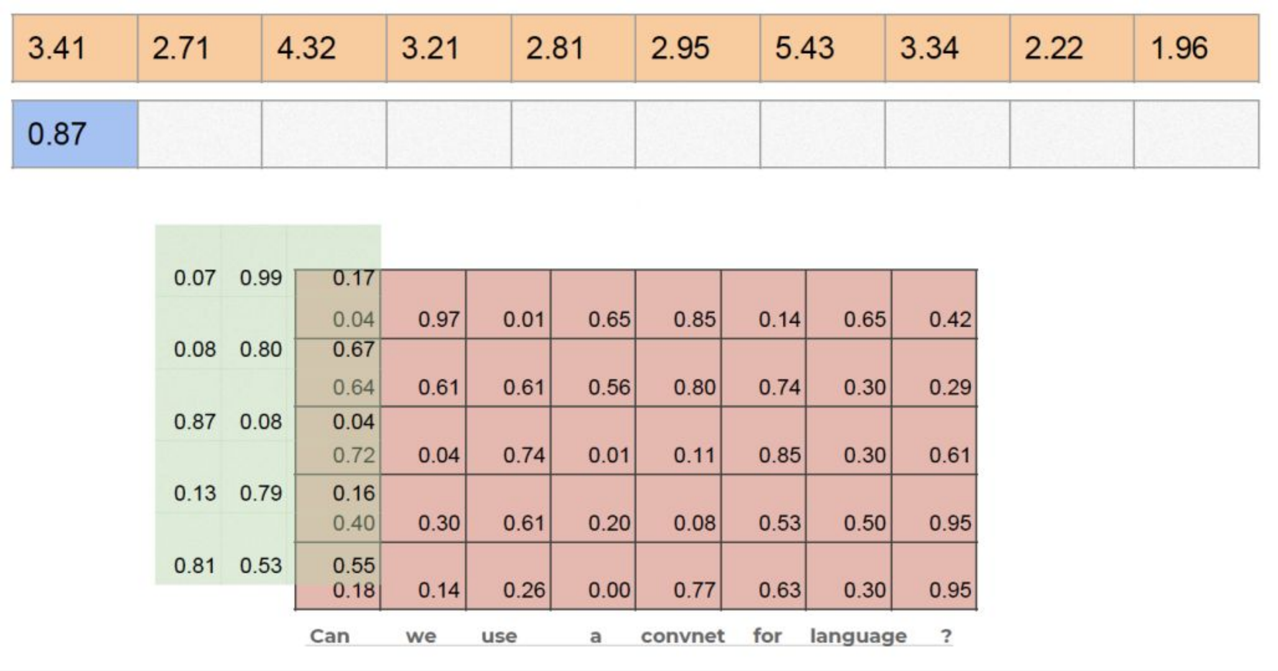
Convolutional NNs: NLP

3.41	2.71	4.32	3.21	2.81	2.95	5.43	3.34	2.22	1.96

0.04	0.97	0.01	0.65	0.85	0.14	0.65	0.42
0.64	0.61	0.61	0.56	0.80	0.74	0.30	0.29
0.72	0.04	0.74	0.01	0.11	0.85	0.30	0.61
0.40	0.30	0.61	0.20	0.08	0.53	0.50	0.95
0.18	0.14	0.26	0.00	0.77	0.63	0.30	0.95
Can	we	use	a	convnet	for	language	?

0.97	0.86	0.70
0.35	0.28	0.26
0.15	0.28	0.18
0.84	0.94	0.09
0.61	0.93	0.21

Convolutional NNs: NLP



Convolutional NNs: NLP

3.41	2.71	4.32	3.21	2.81	2.95	5.43	3.34	2.22	1.96

0.87	0.28	0.64	4.30	3.66	2.71	4.90	2.55	0.30	0.80

0.04	0.97	0.01	0.65	0.85	0.14	0.65	0.42
0.64	0.61	0.61	0.56	0.80	0.74	0.30	0.29
0.72	0.04	0.74	0.01	0.11	0.85	0.30	0.61
0.40	0.30	0.61	0.20	0.08	0.53	0.50	0.95
0.18	0.14	0.26	0.00	0.77	0.63	0.30	0.95
Can	we	use	a	convnet	for	language	?

0.07	0.99	0.17
0.08	0.80	0.67
0.87	0.08	0.04
0.13	0.79	0.16
0.81	0.53	0.55

Convolutional NNs: NLP

3.41	2.71	4.32	3.21	2.81	2.95	5.43	3.34	2.22	1.96
0.87	0.28	0.64	4.30	3.66	2.71	4.90	2.55	0.30	0.80
4.51	3.84	1.63	1.71	1.67	3.51	4.69	4.01	3.55	4.68
0.68	2.43	4.51	4.30	1.69	0.26	3.52	1.67	3.27	2.96
2.68	2.43	4.51	0.30	3.69	0.26	3.52	2.67	4.27	2.96

Convolutional NNs: NLP

4.32	
4.30	
4.51	
4.51	
3.69	

3.41	2.71	4.32	3.21	2.81	2.95	5.43	3.34	2.22	1.96
0.87	0.28	0.64	4.30	3.66	2.71	4.90	2.55	0.30	0.80
4.51	3.84	1.63	1.71	1.67	3.51	4.69	4.01	3.55	4.68
0.68	2.43	4.51	4.30	1.69	0.26	3.52	1.67	3.27	2.96
2.68	2.43	3.51	0.30	3.69	0.26	3.52	2.67	4.27	2.96

Convolutional NNs: An NLP example

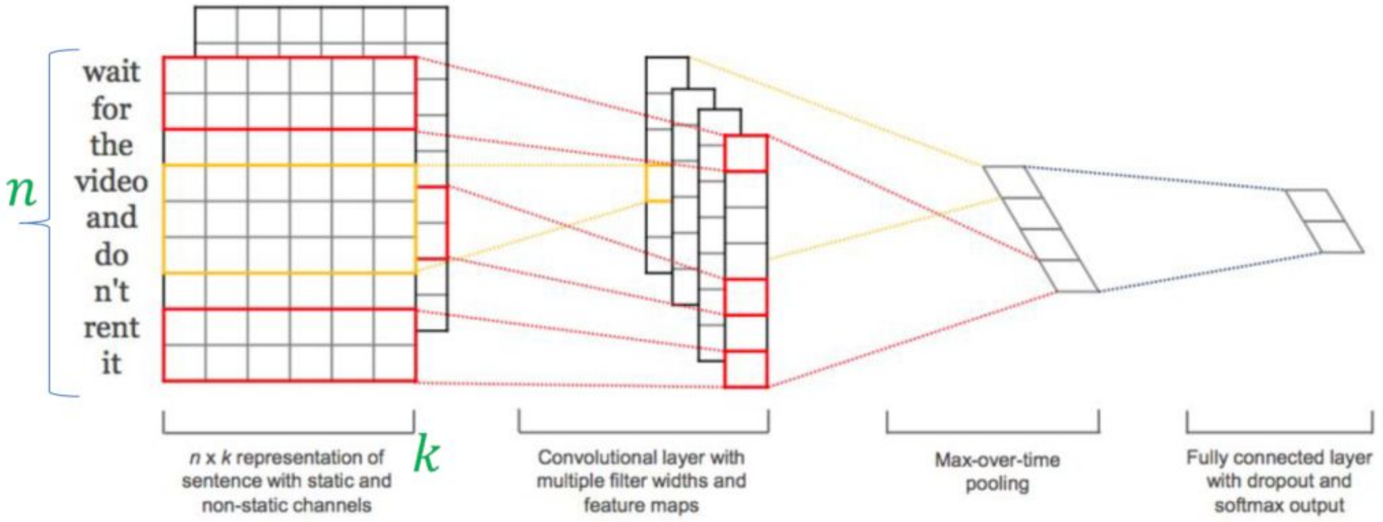


Figure 1: Model architecture with two channels for an example sentence.

n-words (possibly zero padded) and each word vector has k-dimensions

"Improving coreference resolution by learning entity-level distributed representations", Clark et al., 2016. https://arxiv.org/pdf/1606.01323 58/10

Course structure

- ➢ **Question Answering:** X
  - Task definition, datasets, cloze-style tasks, Attentive Reader ← Lecture 15 (NLP 7)
- ➢ **Conference Resolution:** X
  - Task definition, pairs method, clustering method, language models ← Lecture 15 (NLP 7)
- ➢ **Convolutional Neural Networks:** X
  - CNNs in vision, CNNs in language, example ← Lecture 15 (NLP 7)
- ➢ **Transformers:** X
  - Architecture: encoder, self-attention, encoding position, decoder ← Lecture 16 (NLP 8)
  - Existing systems. Ranking ← Lecture 16 (NLP 8)

Literature

➢ **Papers** =

- “Statistical Machine Translation", Koehn, 2009. http://www.statmt.org/book/
- “BLEU", Papineni et al., 2002. https://www.aclweb.org/anthology/P02-1040.pdf
- “Sequence to sequence learning with neural networks", Sutskever et al., 2014. https://arxiv.org/pdf/1409.3215
- “Sequence transduction with recurrent neural networks", Graves, 2012. https://arxiv.org/pdf/1211.3711
- “Neural machine translation by jointly learning to align and translate", Bahdanau et al., 2016. https://arxiv.org/pdf/1409.0473
- “Attention and augmented recurrent neural networks", Olah et al., 2016. https://distill.pub/2016/augmented-rnns/
- “Massive exploration of neural machine translation architectures", Britz et al., 2017. https://arxiv.org/pdf/1703.03906
- “Has AI surpassed humans at translation? Not even close!" https://www.skynettoday.com/editorials/state_of_nmt
- “Googles Neural Machine Translation System", Wu et al., 2016. https://arxiv.org/pdf/1609.08144
- “Achieving human parity on automatic Chinese to English news translation", Hassan et al., 2018. https://arxiv.org/pdf/1803.05567
- “Findings of the 2018 Conference on MT", Bojar et al., 2018. http://www.statmt.org/wmt18/pdf/WMT028.pdf