CS11-747 Neural Networks for NLP

Convolutional Networks for Text

Pengfei Liu

Carnegie Mellon University
Language Technologies Institute

Site
https://phontron.com/class/nn4nlp2020/

With some slides by Graham Neubig

Outline

1. Feature Combinations
2. CNNs and Key Concepts
3. Case Study on Sentiment Classification
4. CNN Variants and Applications
5. Structured CNNs
6. Summary

An Example Prediction Problem: Sentiment Classification

I hate this movie ?

very good
good
neutral
bad
very bad

I love this movie ?

very good
good
neutral
bad
very bad

An Example Prediction Problem: Sentiment Classification

very good
good
neutral
bad
very bad

I **hate** this movie

very good
good
neutral
bad
very bad

I **love** this movie

An Example Prediction Problem: Sentiment Classification

I **hate** this movie → very bad

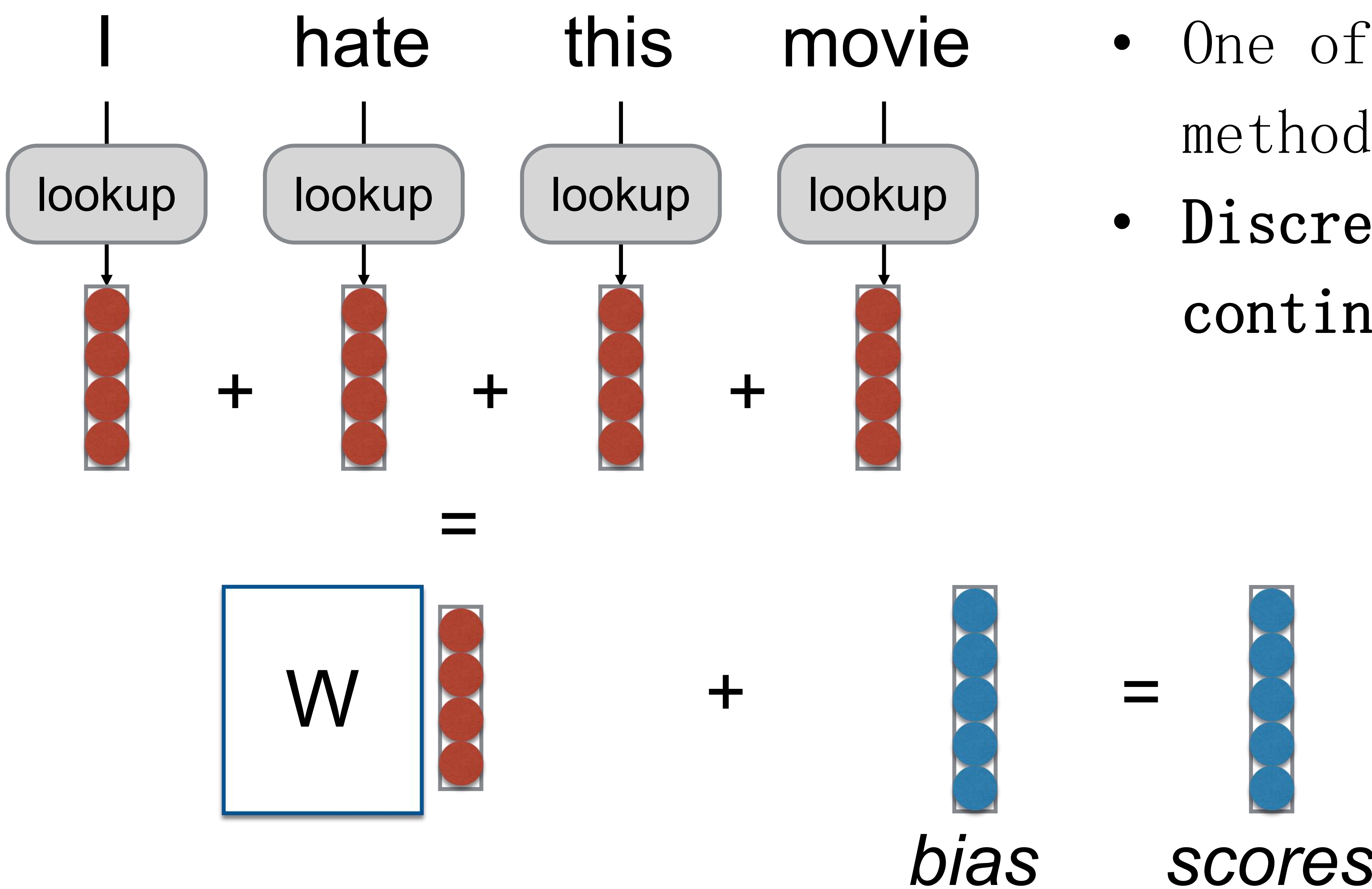
very good
good
neutral
bad
very bad

I **love** this movie → very good

very good
good
neutral
bad
very bad

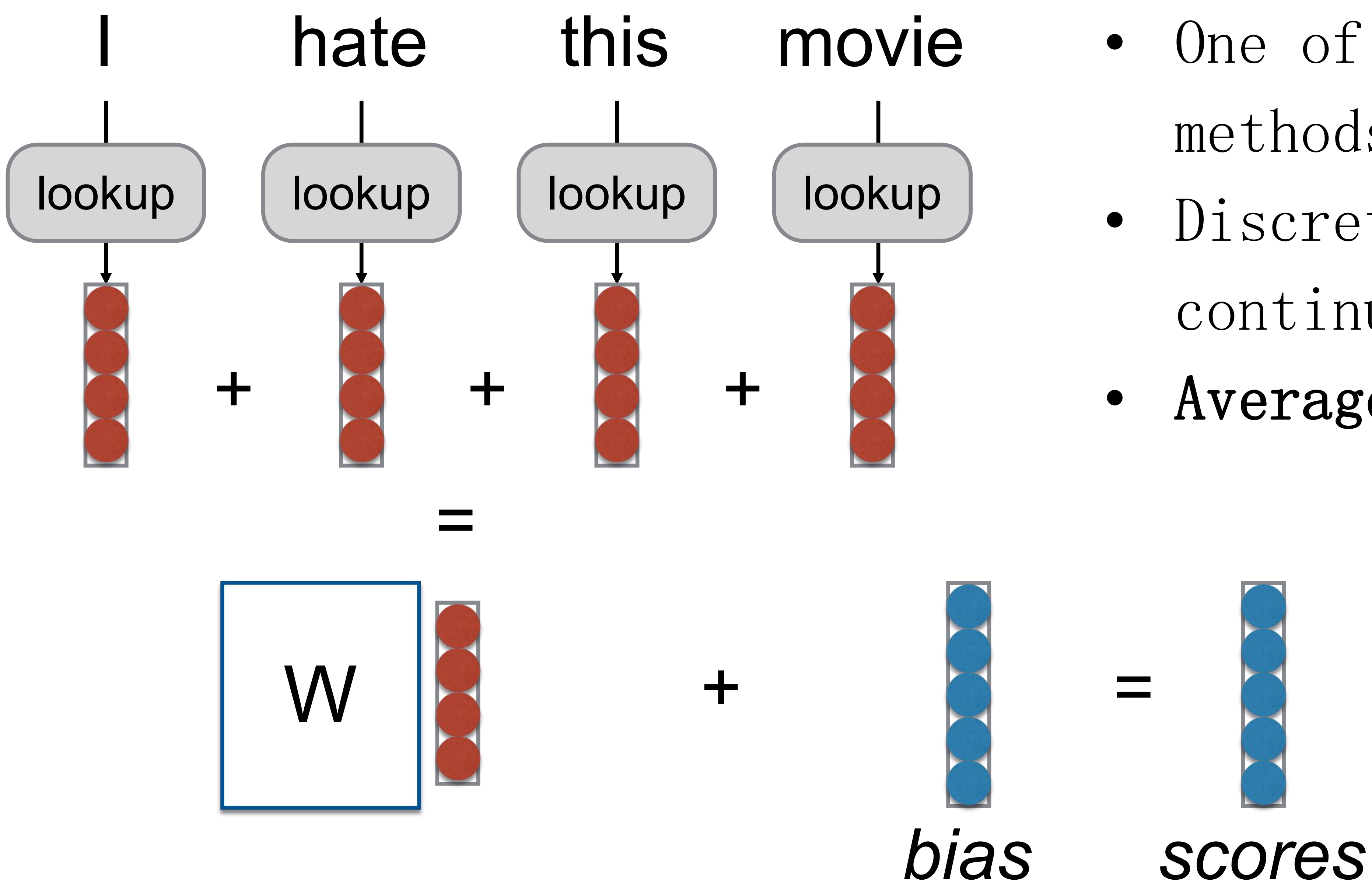
how does our machine to do this task?

Continuous Bag of Words (CBOW)



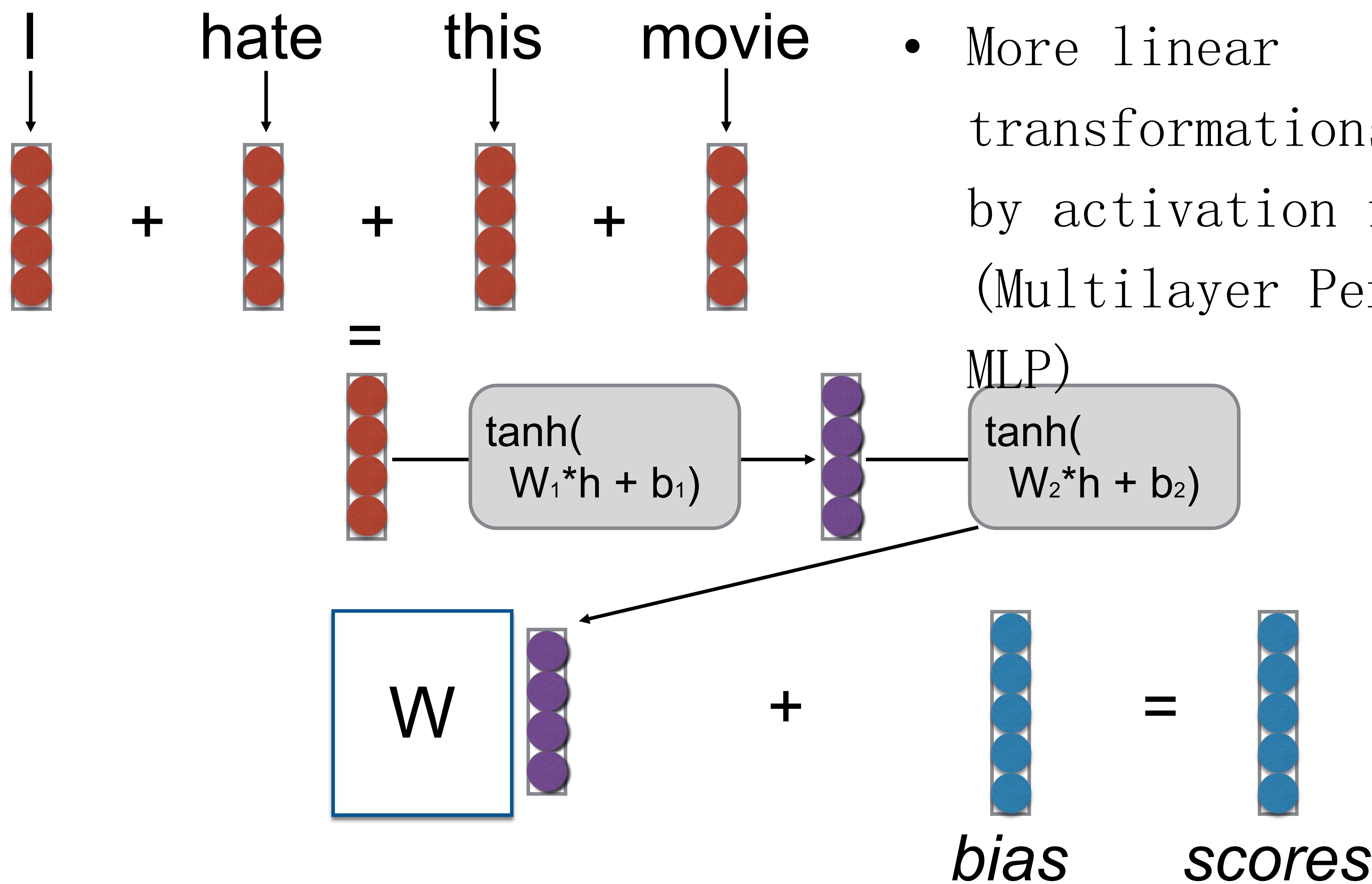
- One of the simplest methods
- **Discrete symbols to continuous vectors**

Continuous Bag of Words (CBOW)



- One of the simplest methods
- Discrete symbols to continuous vectors
- **Average all vectors**

Deep CBOW



- More linear transformations followed by activation functions (Multilayer Perceptron, MLP)

What’s the Use of the “Deep”

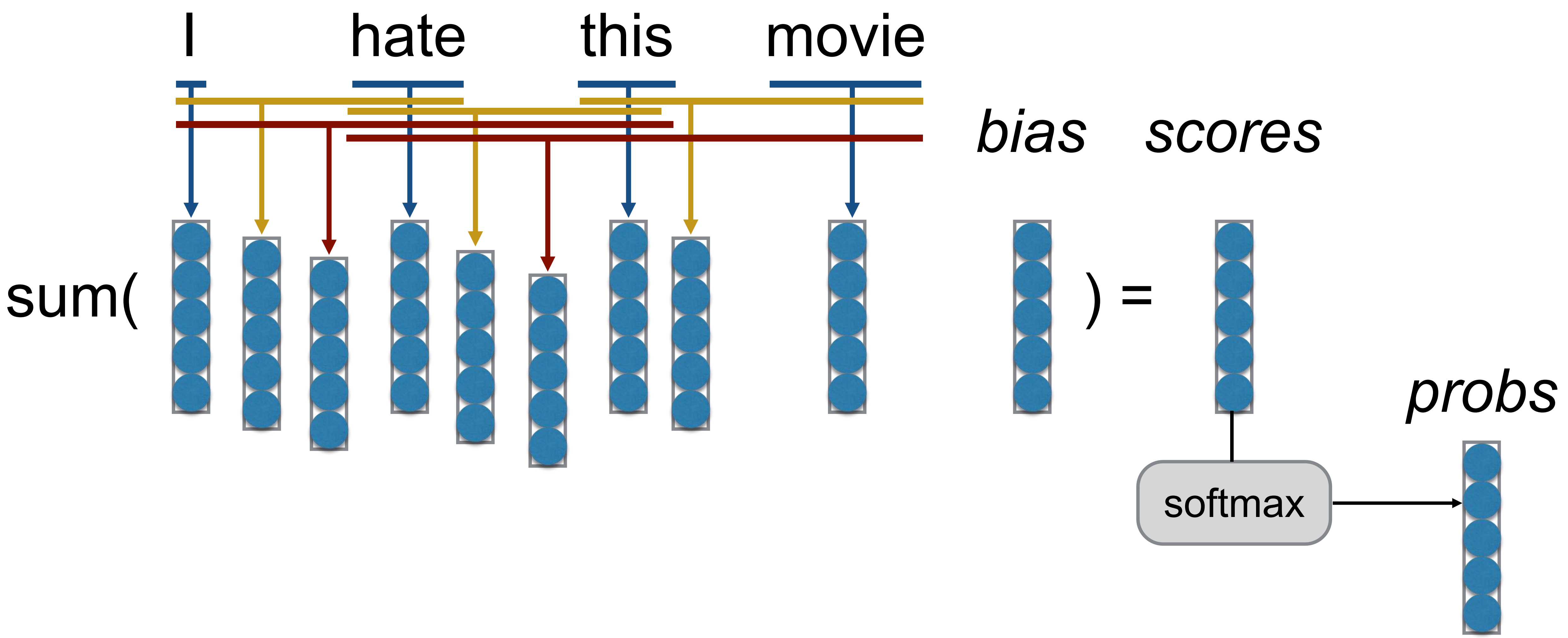
- Multiple MLP layers allow us easily to learn feature combinations (a node in the second layer might be “feature 1 AND feature 5 are active”)

- e.g. capture things such as “not” AND “hate”

- BUT! Cannot handle “not hate”

Handling Combinations

Bag of n-grams

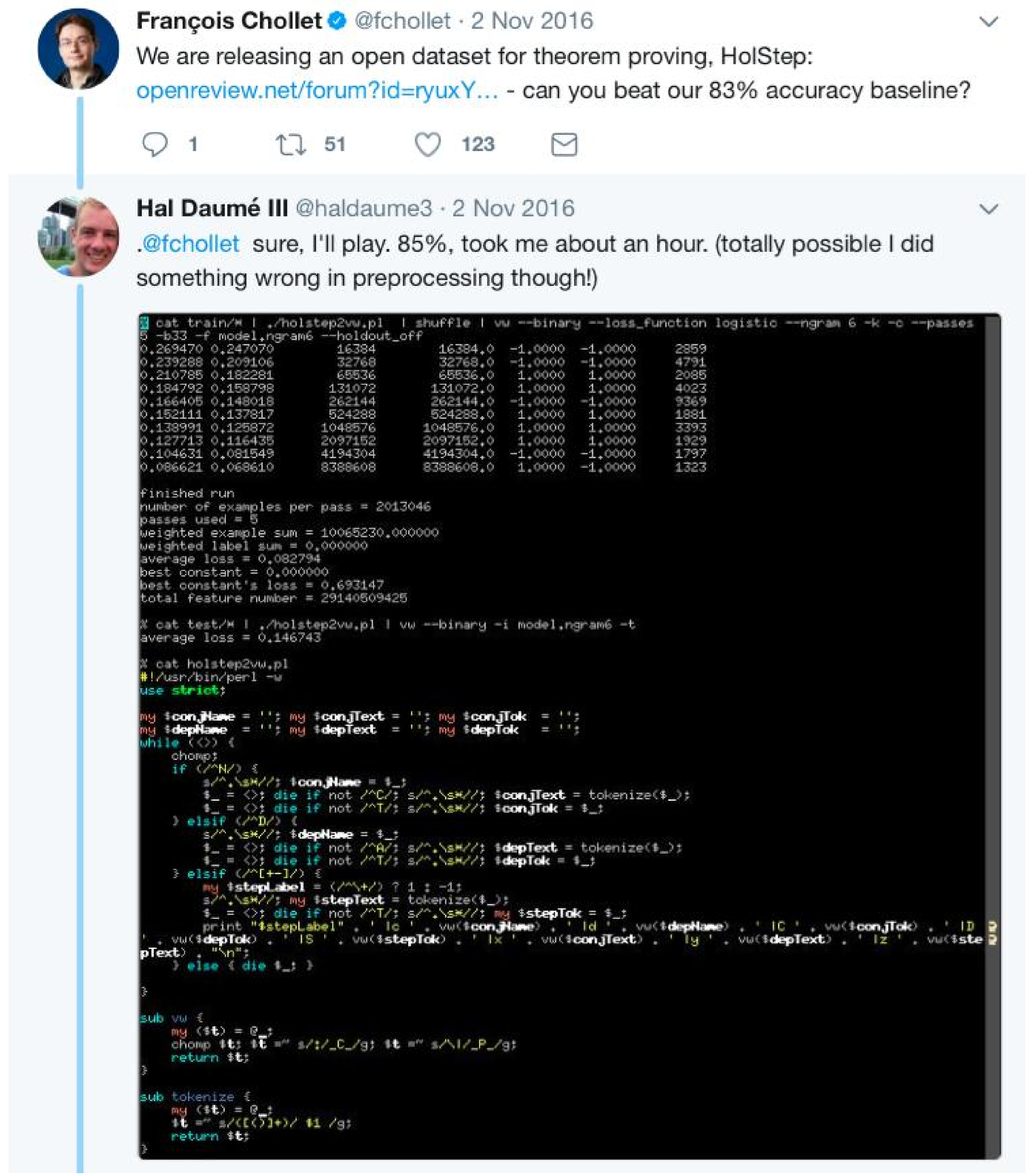


- A contiguous sequence of words
- Concatenate word vectors

Why Bag of n-grams?

- Allow us to capture combination features in a simple way "don't love", "not the best"

- Decent baseline and works pretty well



What Problems w/ Bag of n-grams?

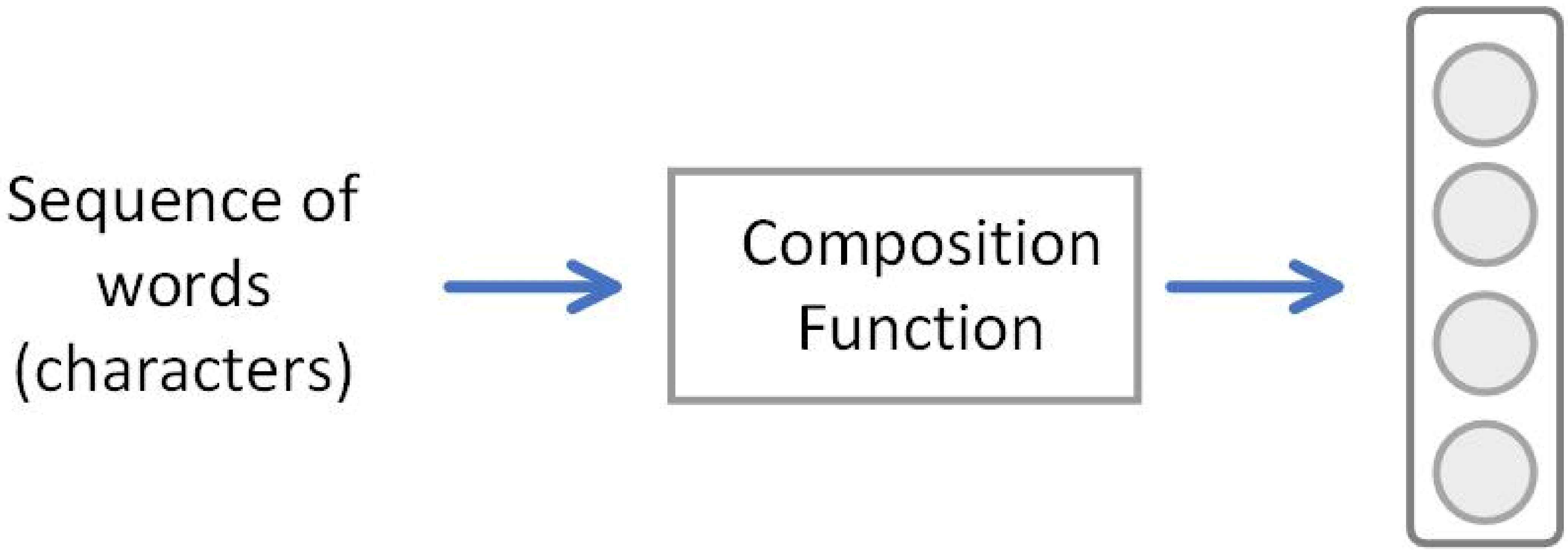
- Same as before: parameter explosion
- No sharing between similar words/n-grams
- Lose the global sequence order

What Problems w/ Bag of n-grams?

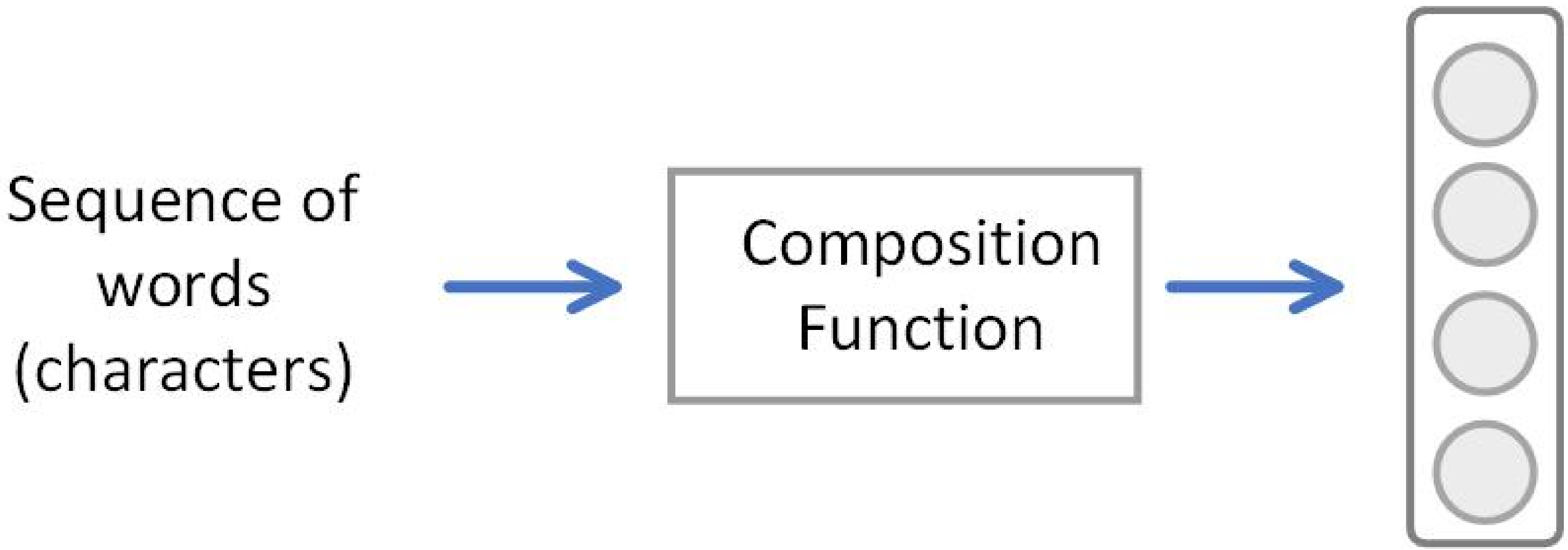
- Same as before: parameter explosion
- No sharing between similar words/n-grams
- Lose the global sequence order

Other solutions?

Neural Sequence Models

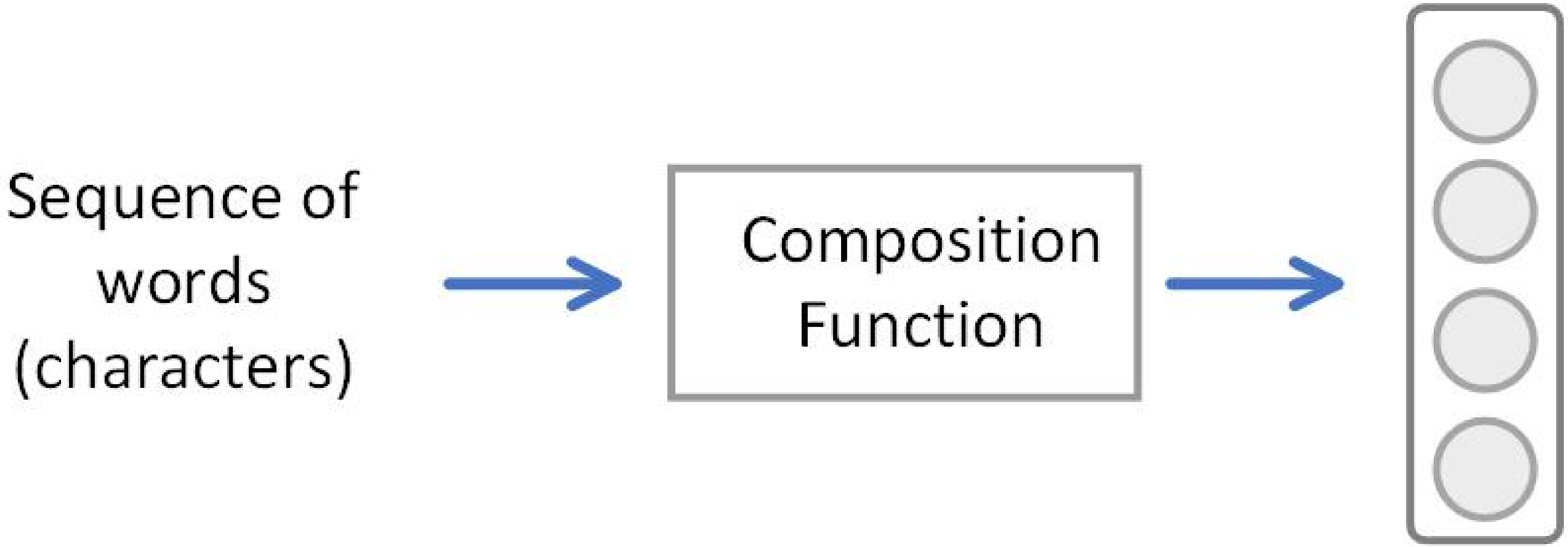


Neural Sequence Models



Most of NLP tasks → Sequence representation learning problem

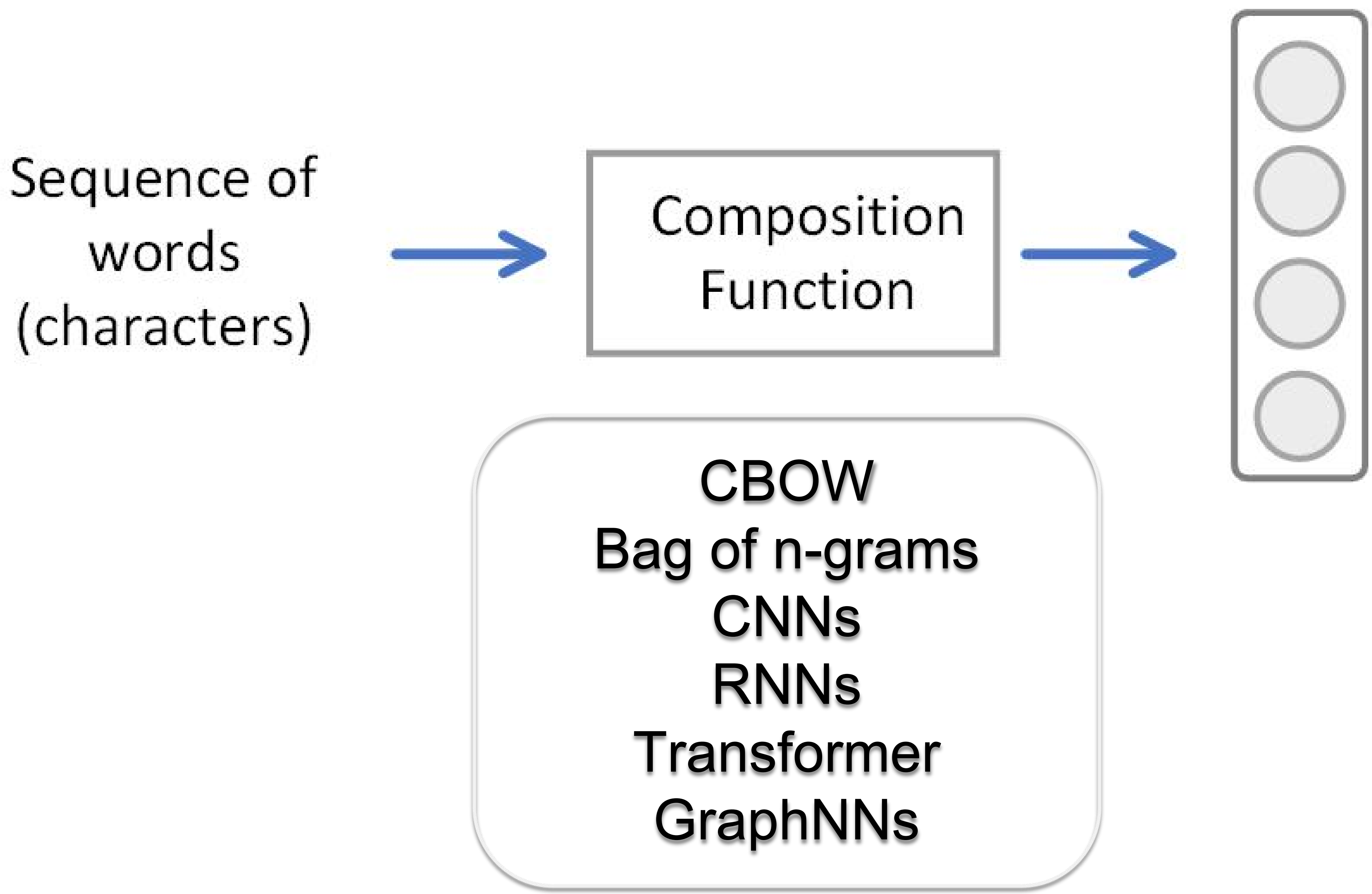
Neural Sequence Models



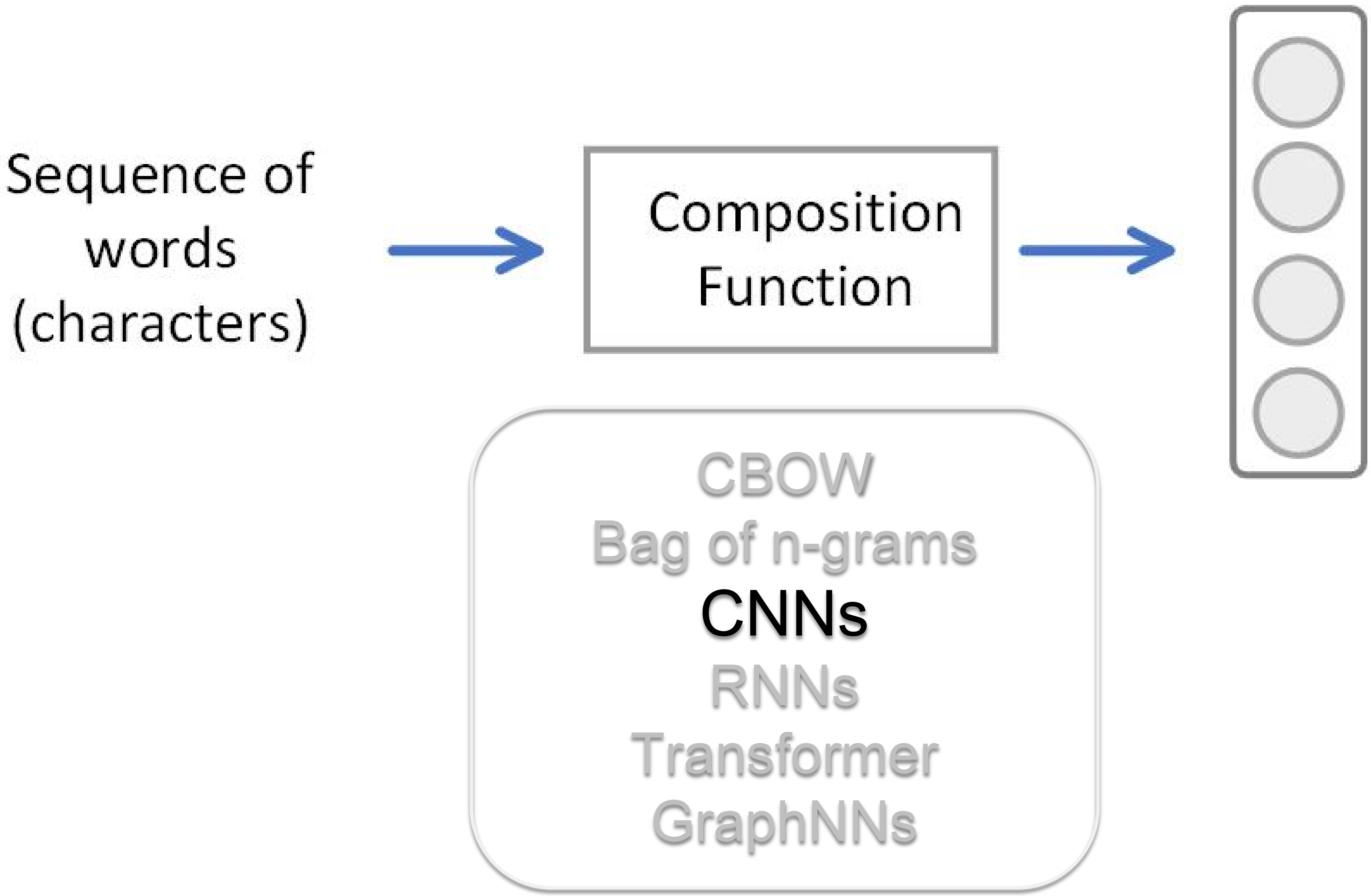
char: i-m-p-o-s-s-i-b-l-e

word: I-love-this-movie

Neural Sequence Models



Neural Sequence Models



Convolutional Neural Networks

Definition of Convolution

Convolution -- > mathematical operation

- Continuous

$$(f * g)(t) = \int_{-\infty}^{\infty} f(t-\tau)g(\tau)d\tau$$

- Discrete

$$(f * g)[n] = \sum_{n=-M}^{M} f[n-m]g[m]$$

Definition of Convolution

Convolution -- > mathematical operation

- Continuous

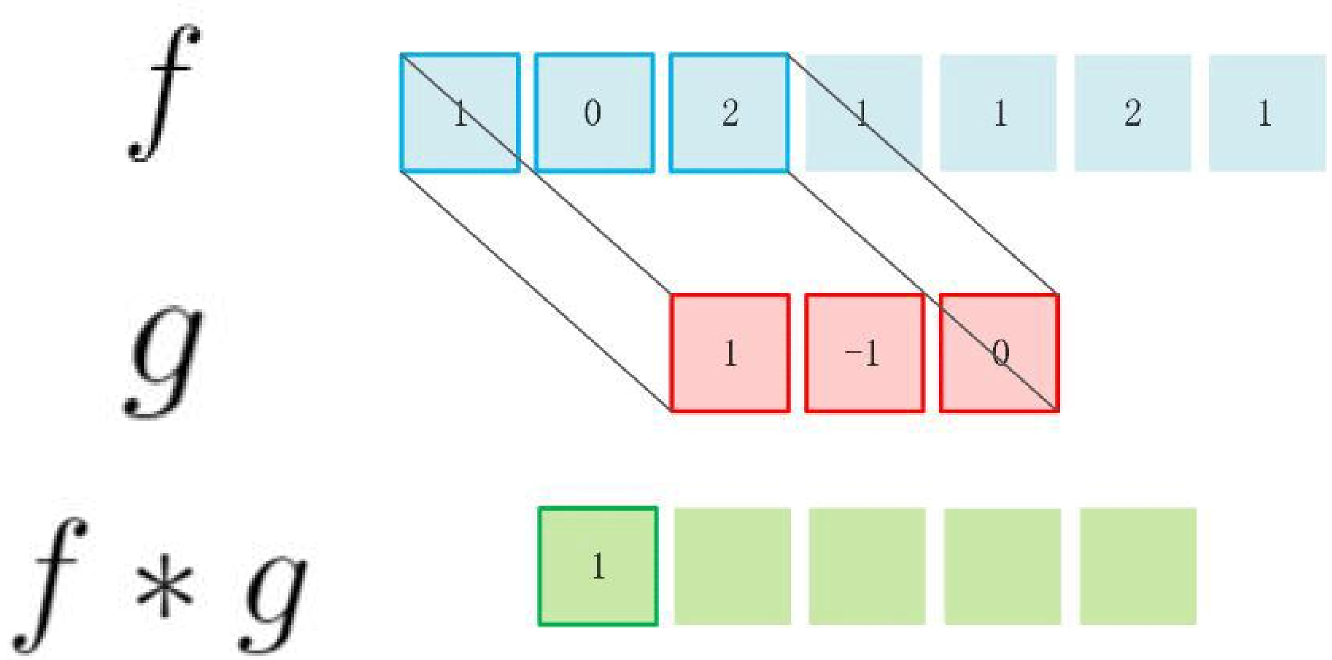
$$(f * g)(t) = \int_{-\infty}^{\infty} f(t - \tau) g(\tau) d\tau$$

- Discrete

$$\boxed{(f * g)[n] = \sum_{m=-M}^{M} f[n] g[m]}$$

Intuitive Understanding

$$(f * g)[n] = \sum_{m=-M}^{M} f[n]g[m]$$



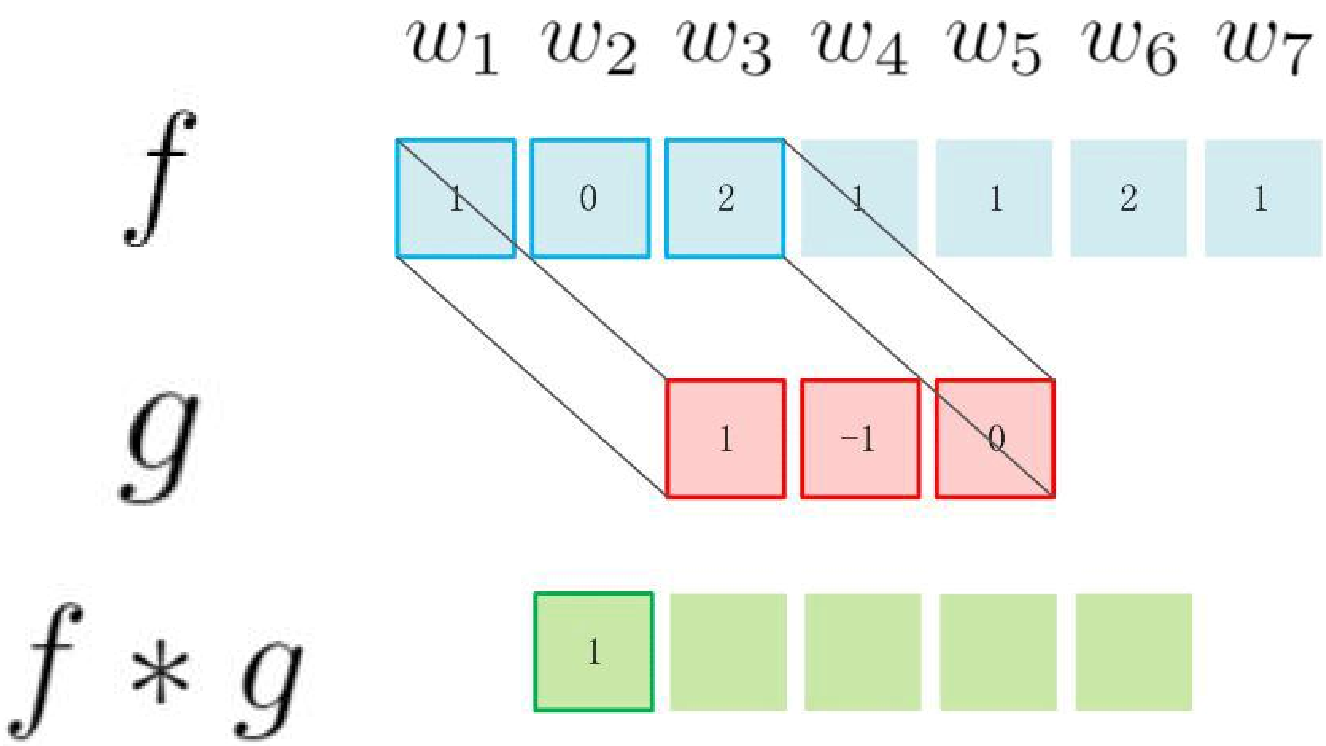
Input: feature vector

Filter: learnable param.

Output: hidden vector

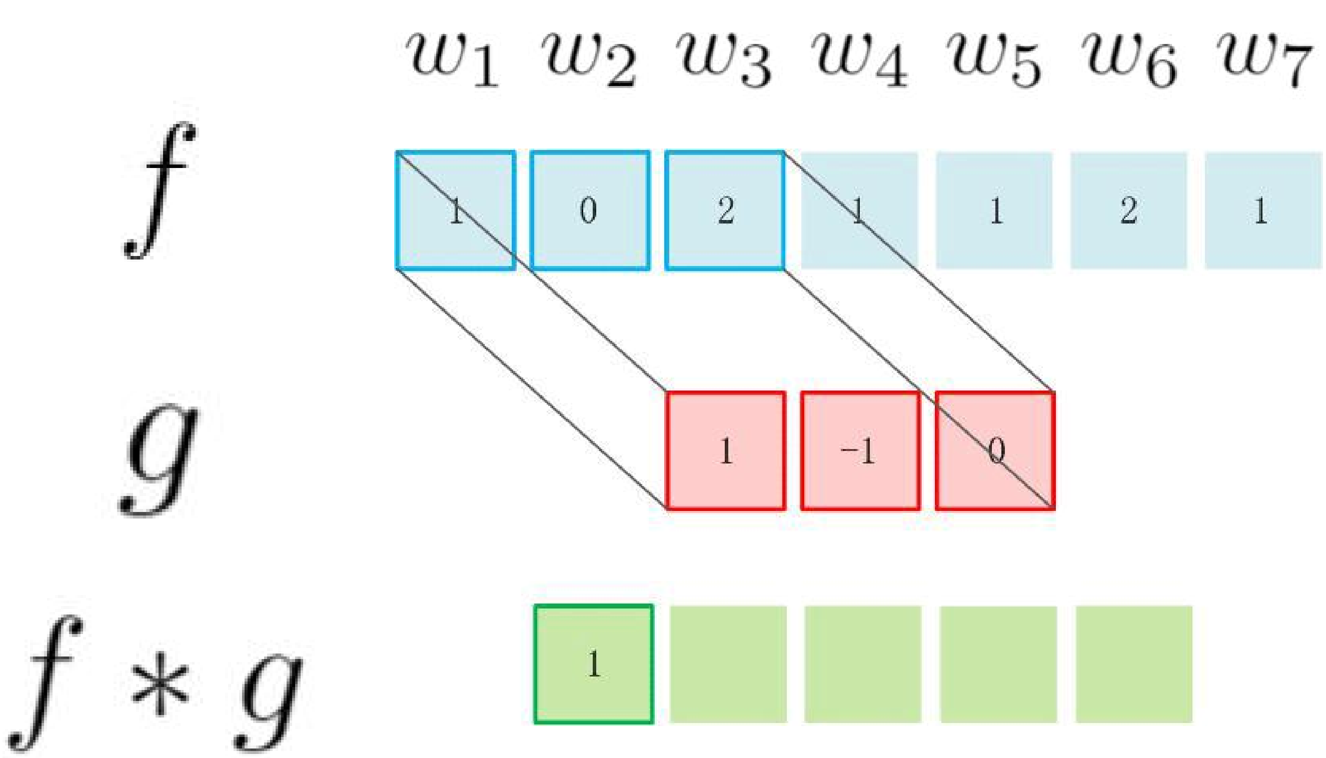
Priori Entailed by CNNs

$$(f * g)[n] = \sum_{m=-M}^{M} f[n]g[m]$$



Priori Entailed by CNNs

$$(f * g)[n] = \sum_{m=-M}^{M} f[n]g[m]$$

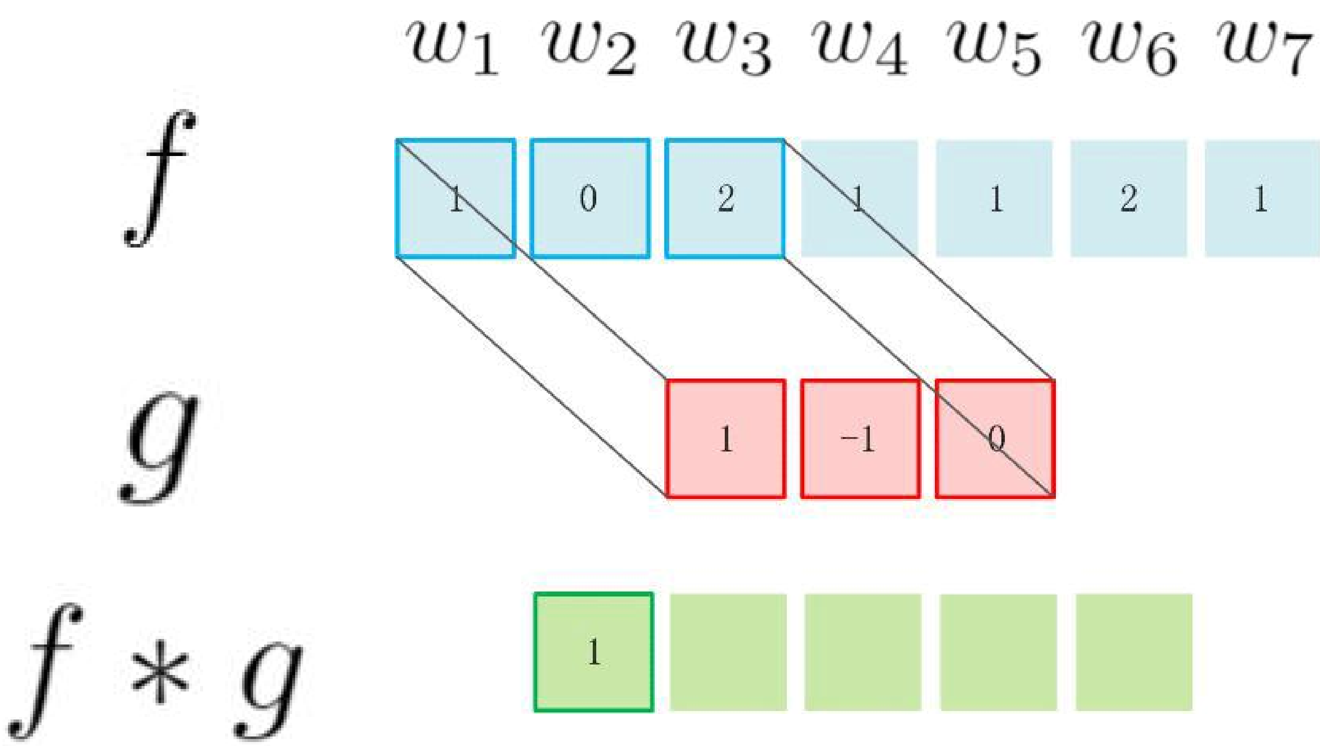


Local bias:

Different words could interact with their neighbors

Priori Entailed by CNNs

$$(f * g)[n] = \sum_{m=-M}^{M} f[n]g[m]$$

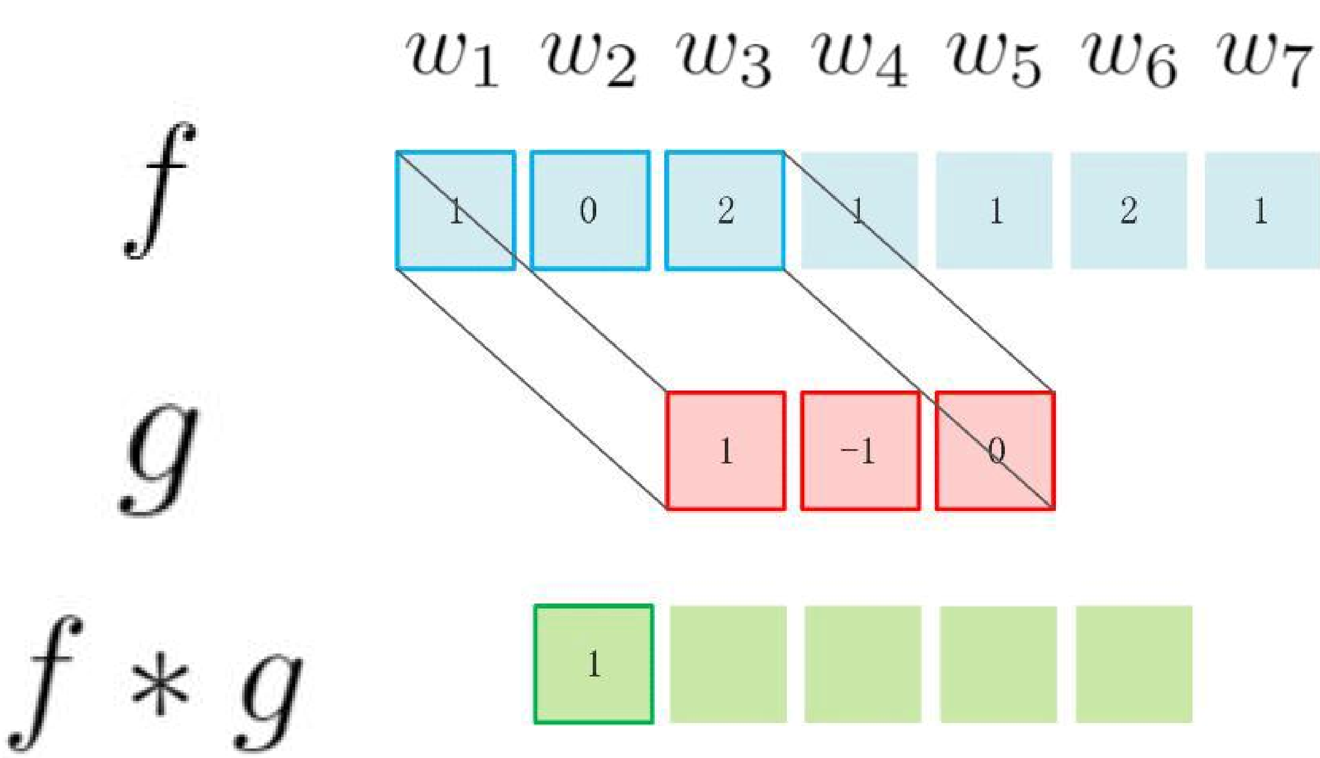


Local bias:

Different words could interact with their neighbors

Priori Entailed by CNNs

$$(f * g)[n] = \sum_{m=-M}^{M} f[n]g[m]$$



Parameter sharing:

The parameters of composition function are the same.

Basics of CNNs

Concept: 2d Convolution

$$(f * g)[n] = \sum_{m=-M}^{M} f[n]g[m]$$

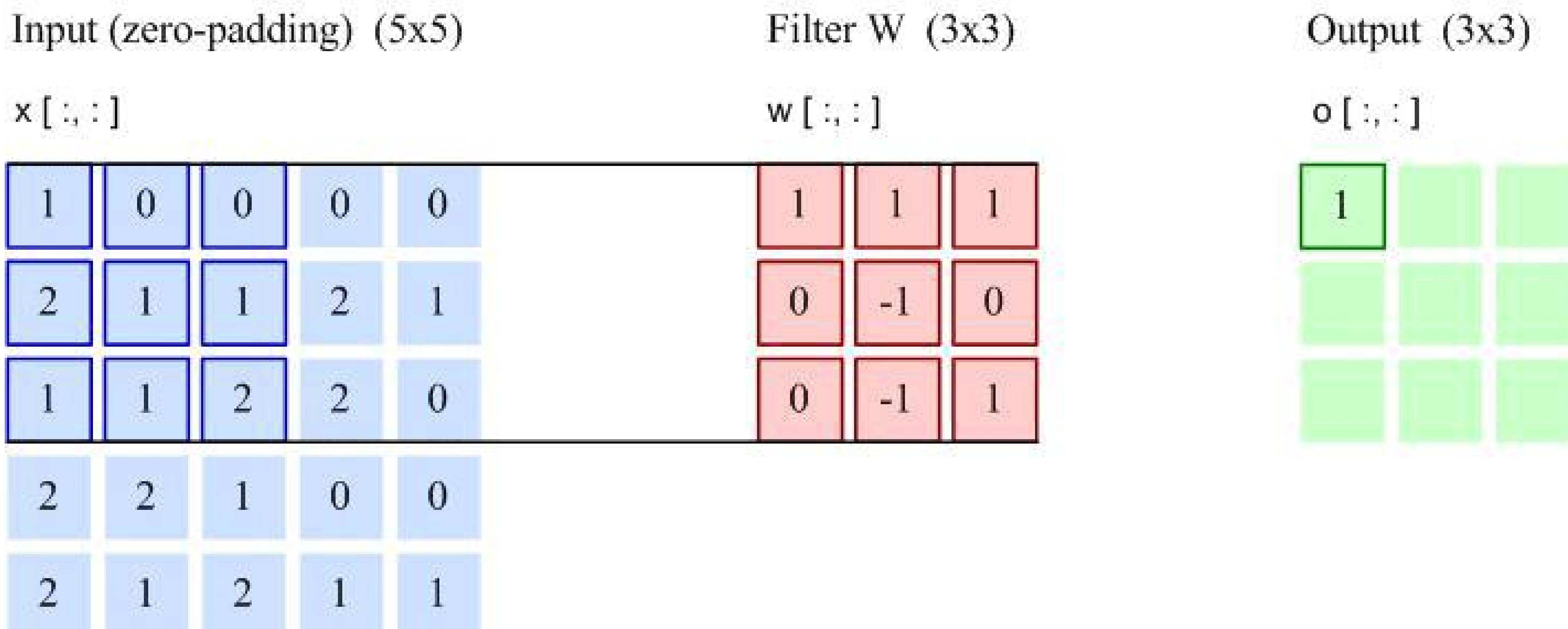
- Deal with 2-dimension signal, i.e., image

Concept: 2d Convolution

$$(f * g)[n] = \sum_{m=-M}^{M} f[n]g[m]$$

Concept: 2d Convolution

$$(f * g)[n] = \sum_{m=-M}^{M} f[n]g[m]$$

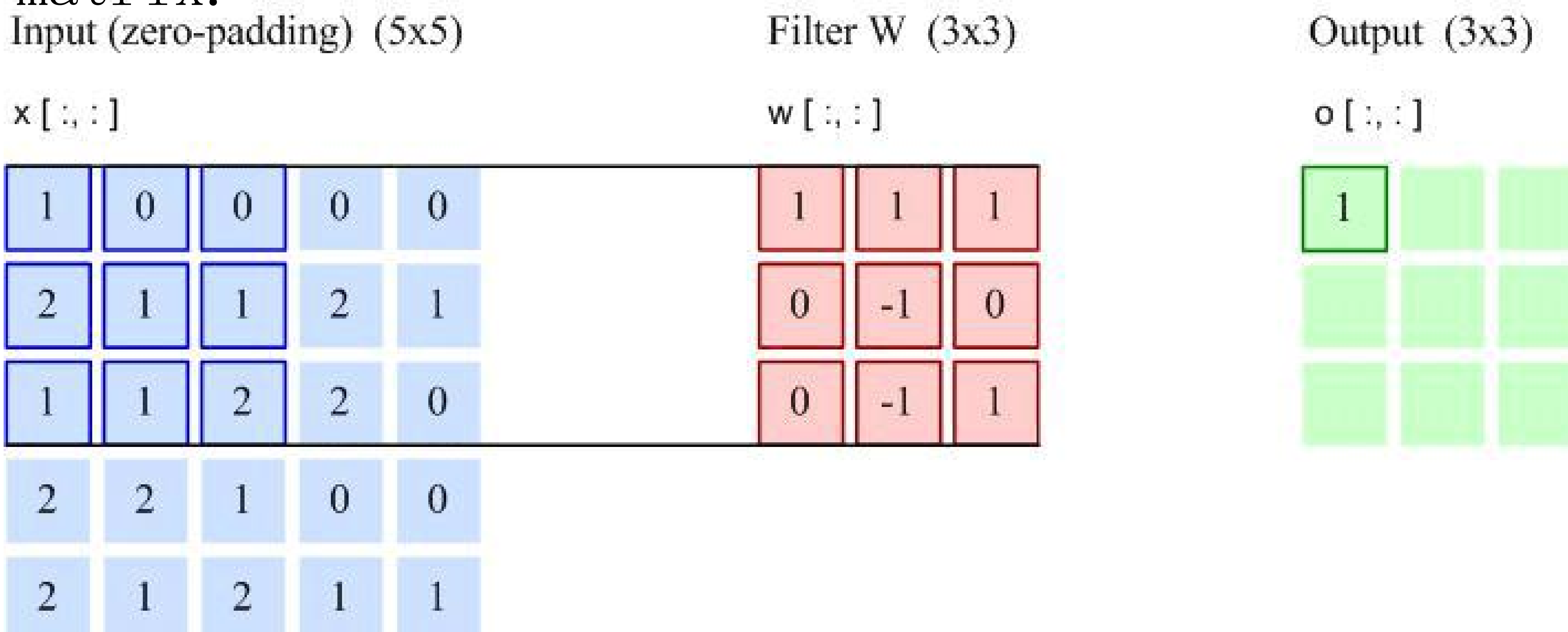


Concept: Stride

Stride: the number of units shifts over the input matrix.

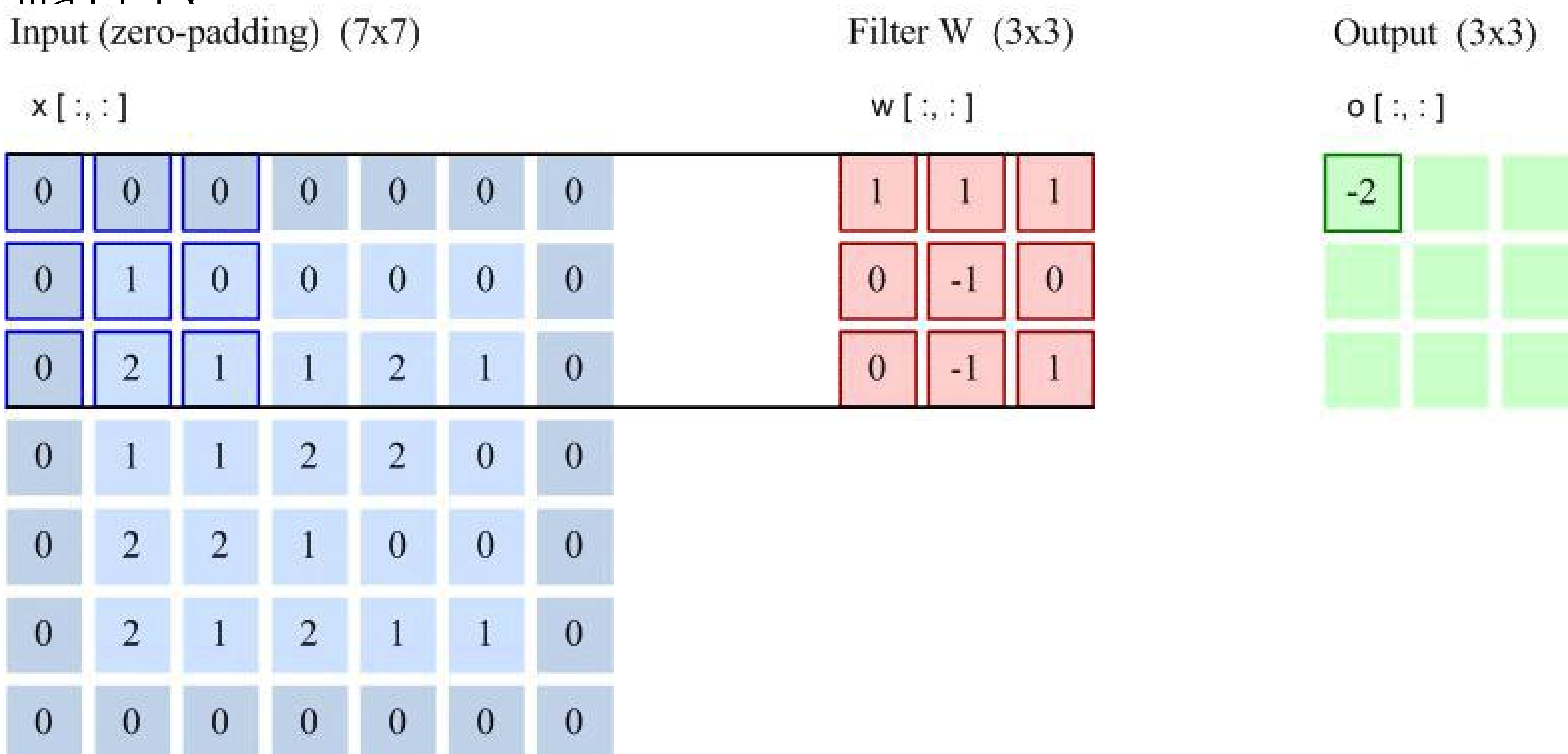
Concept: Stride

Stride: the number of units shifts over the input matrix.



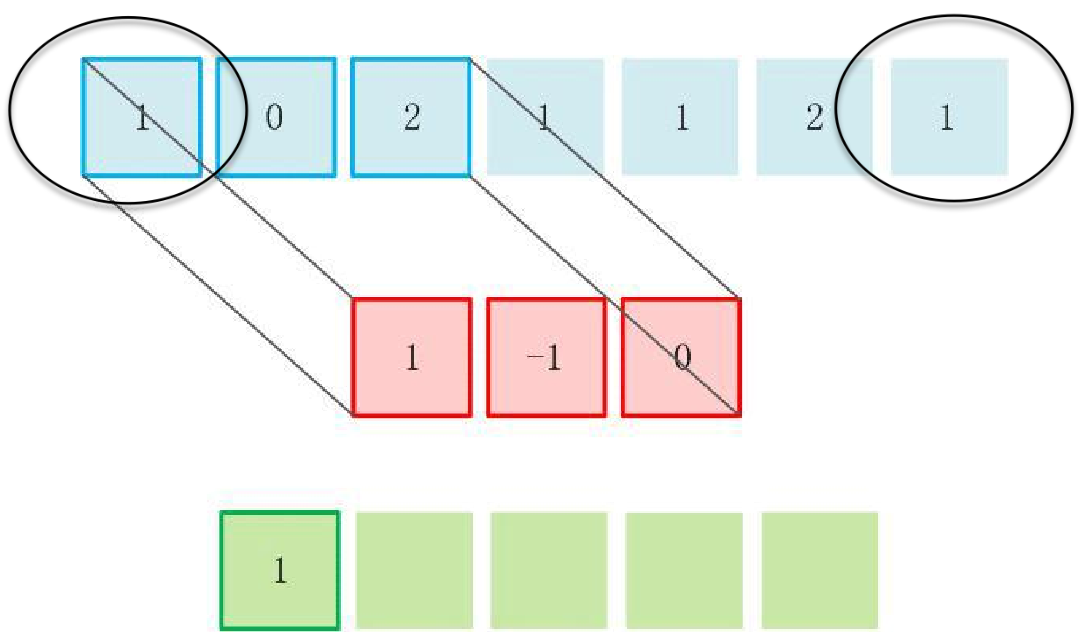
Concept: Stride

Stride: the number of units shifts over the input matrix



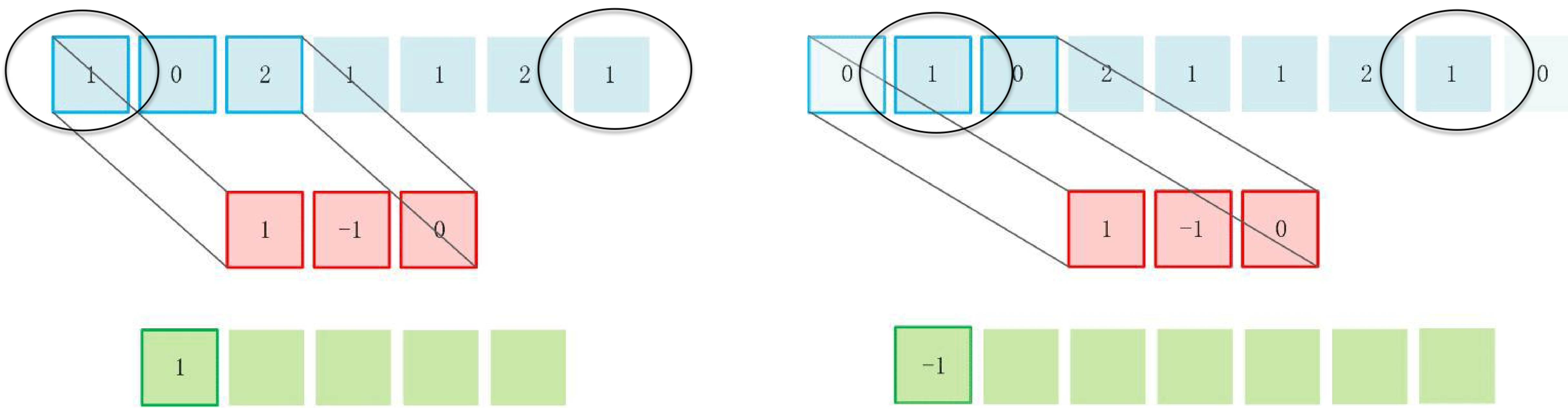
Concept: Padding

Padding: dealing with the units at the boundary of input vector.



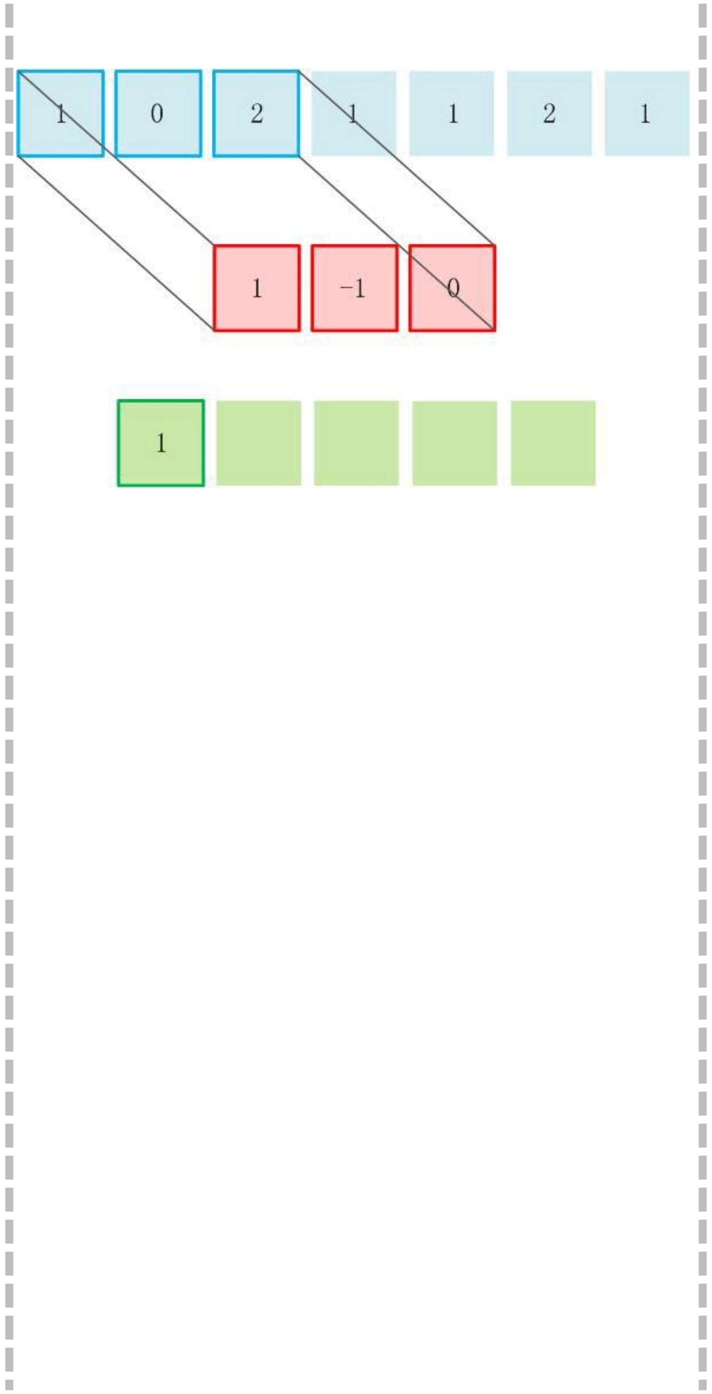
Concept: Padding

Padding: dealing with the units at the boundary of input vector.



Three Types of Convolutions

Narrow

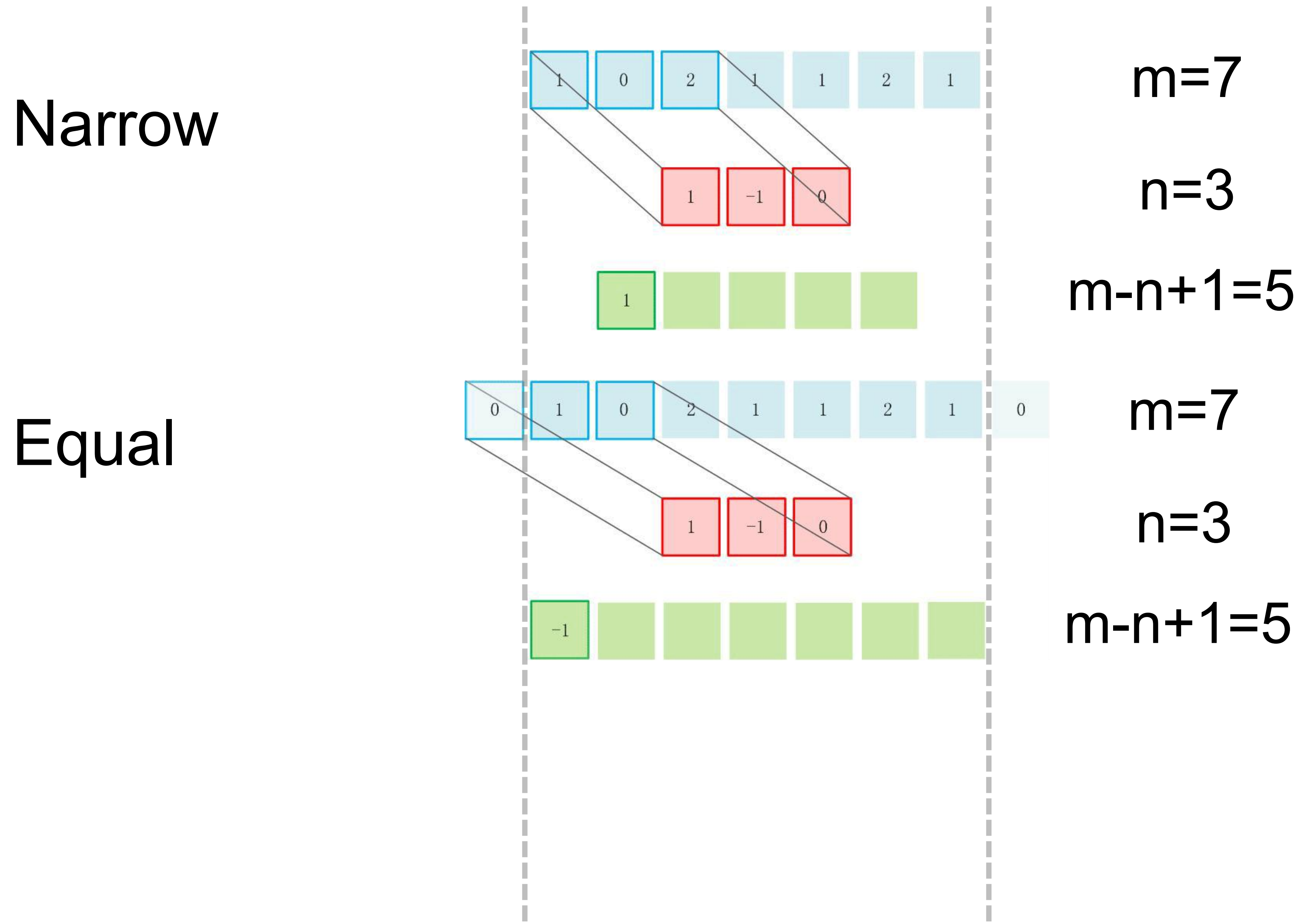


m=7

n=3

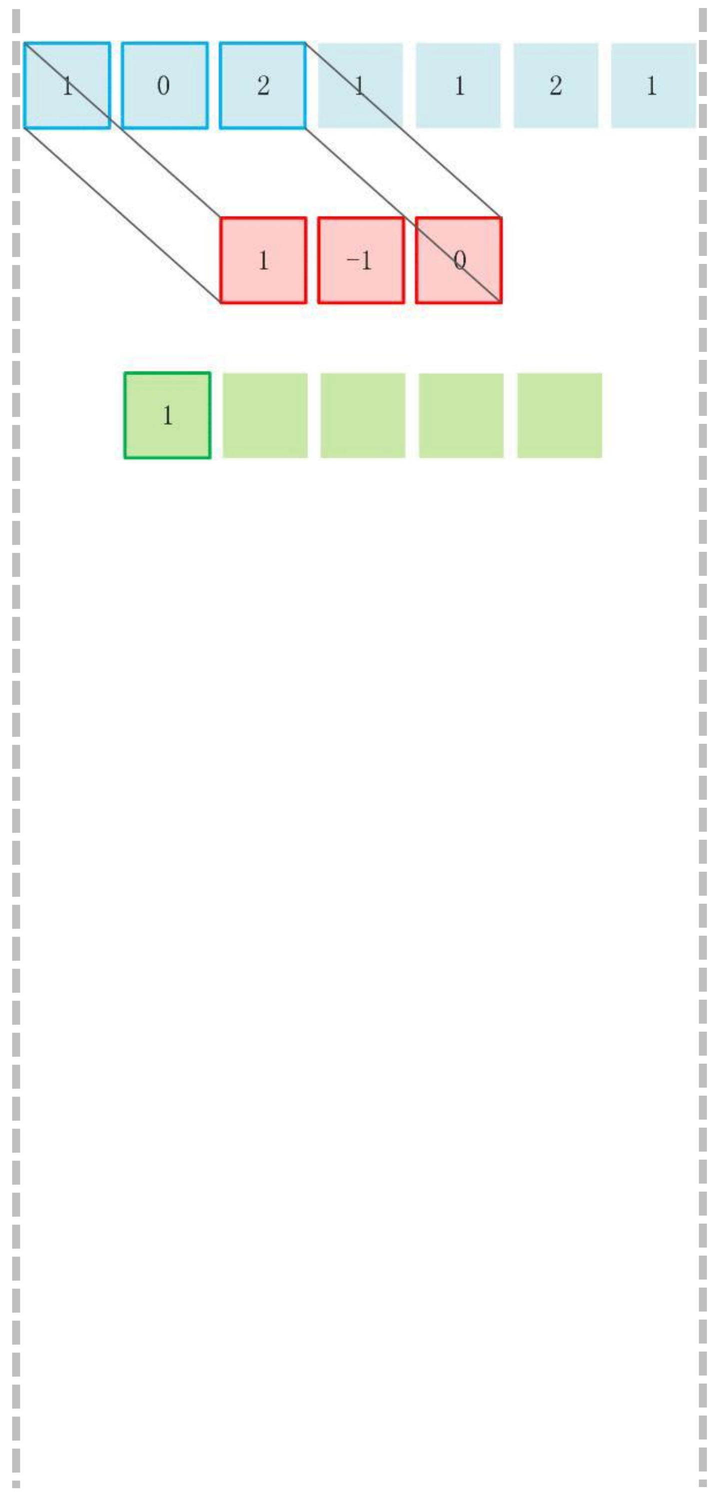
m-n+1=5

Three Types of Convolutions



Three Types of Convolutions

Narrow



m=7

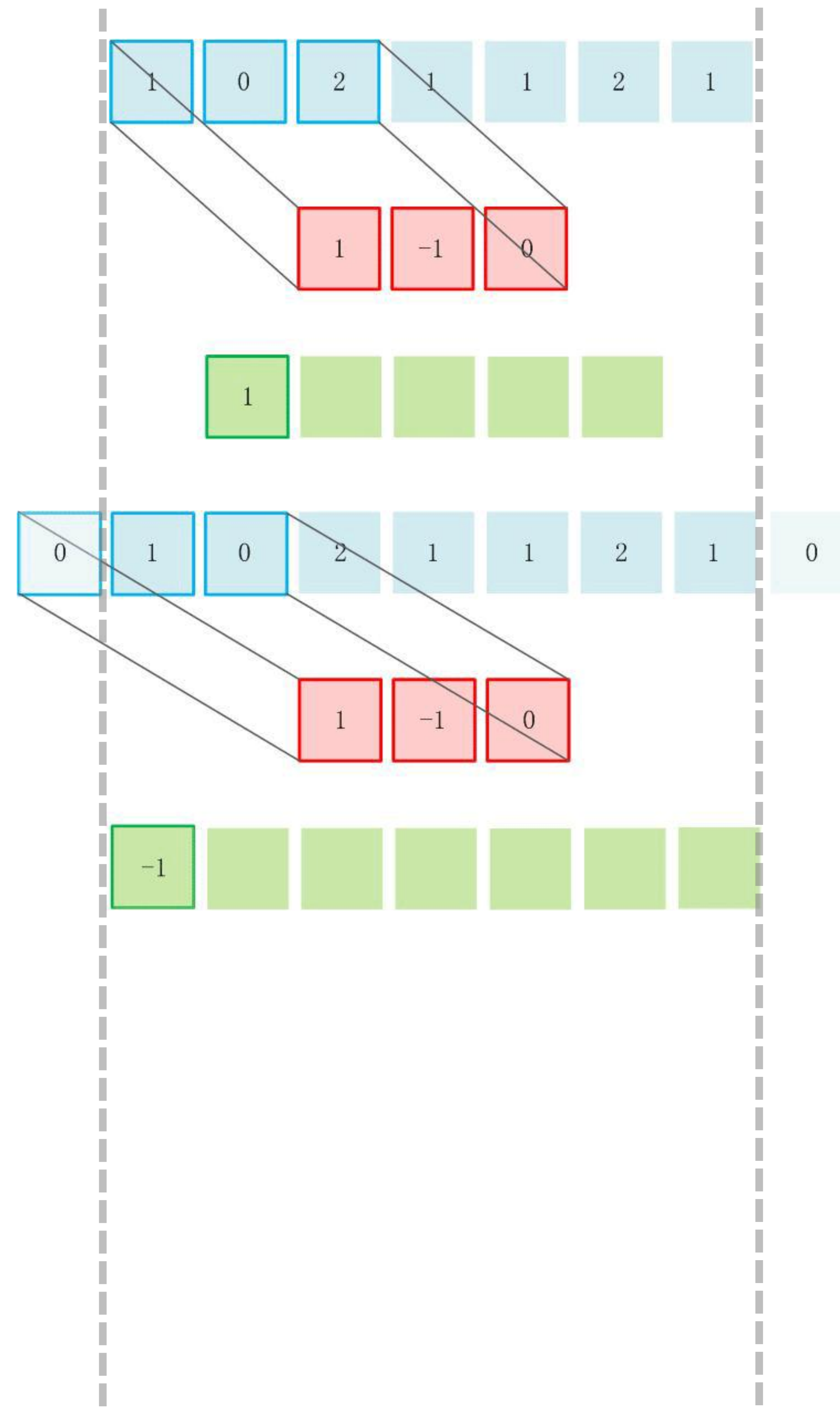
n=3

m-n+1=5

Three Types of Convolutions

Narrow

Equal



m=7

n=3

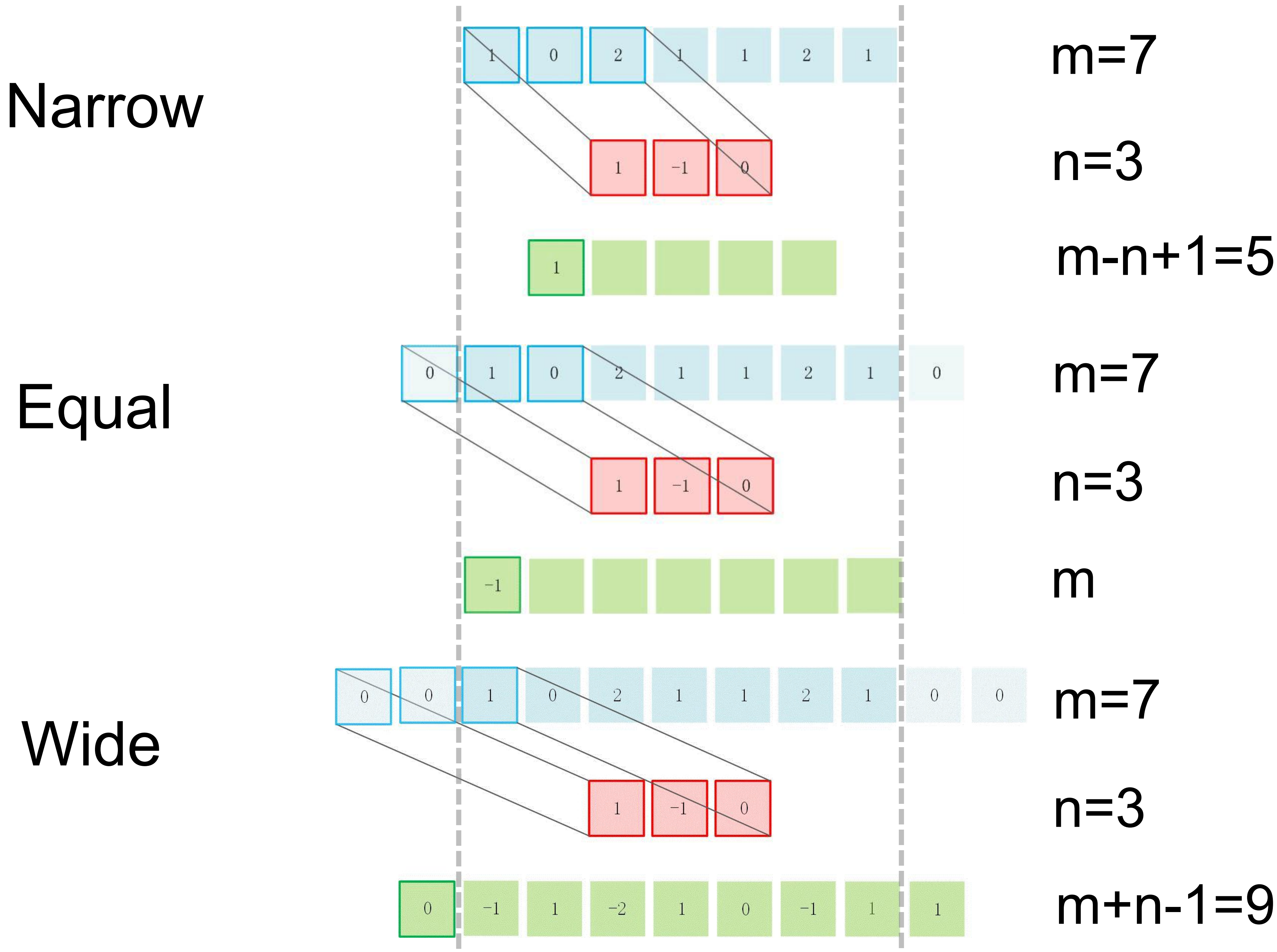
m-n+1=5

m=7

n=3

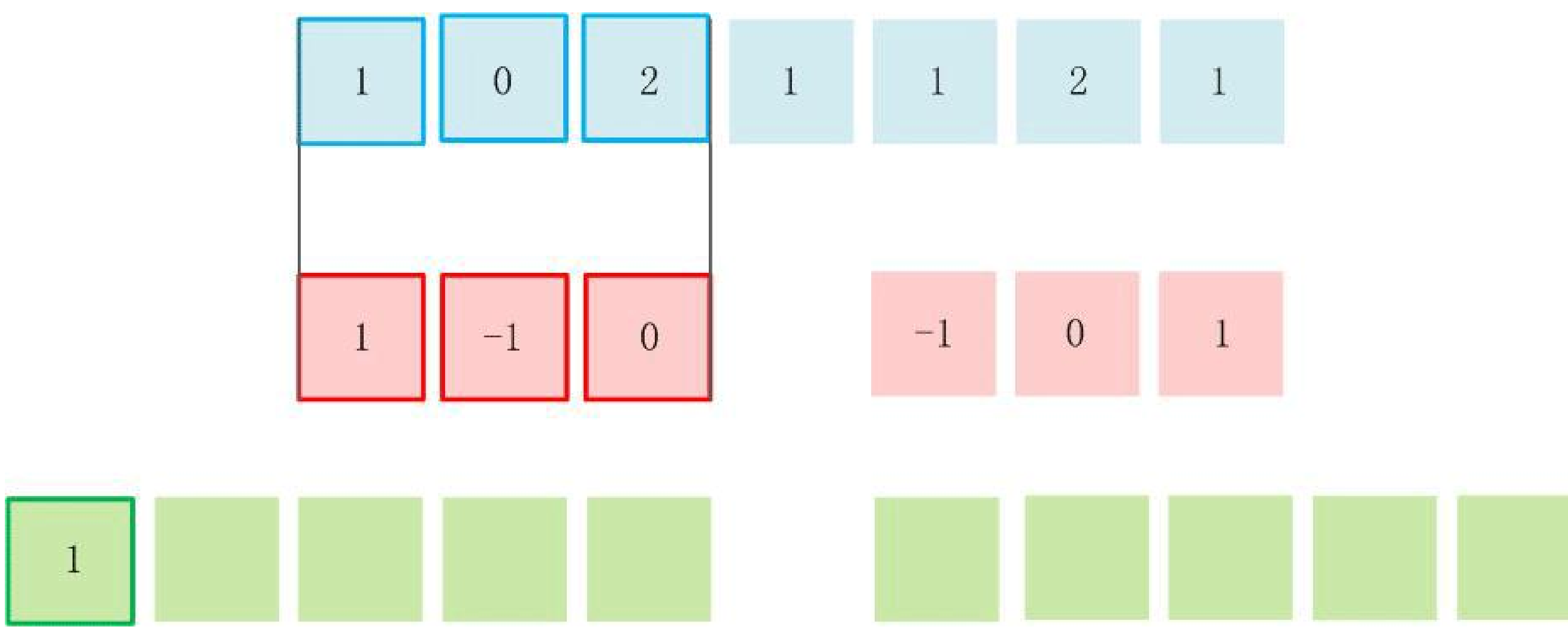
m

Three Types of Convolutions



Concept: Multiple Filters

Motivation: each filter represents a unique feature of the convolution window.



Concept: Pooling

- **Pooling** is an aggregation operation, aiming to select informative features

Concept: Pooling

- **Pooling** is an aggregation operation, aiming to select informative features

- **Max pooling:** “Did you see this feature anywhere in the range?” (most common)

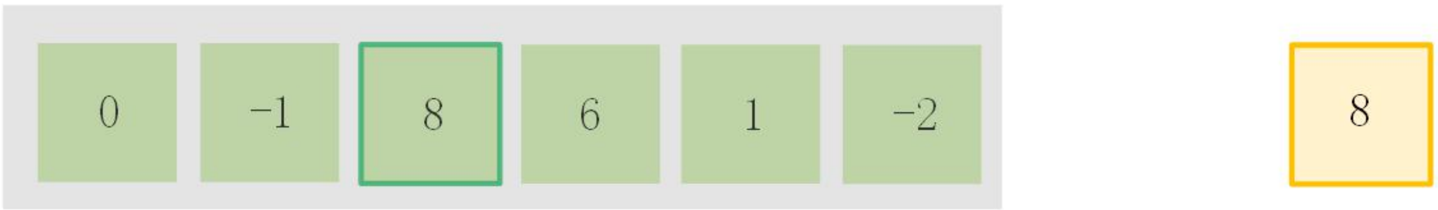
- Average pooling: “How prevalent is this feature over the entire range”

- **k-Max pooling:** “Did you see this feature up to k times?”

- **Dynamic pooling:** “Did you see this feature in the beginning? In the middle? In the end?”

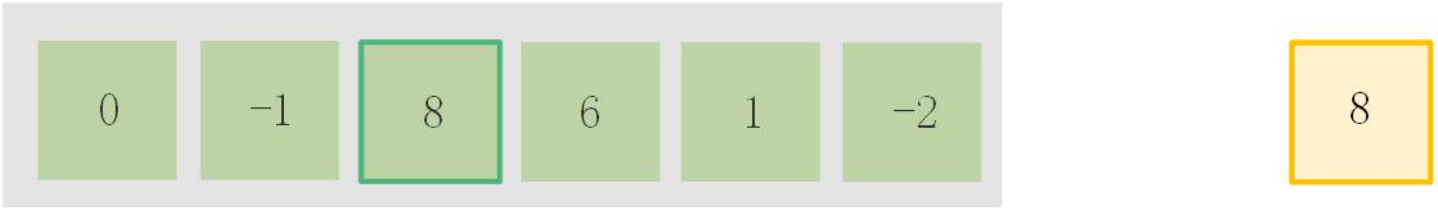
Concept: Pooling

Max pooling:

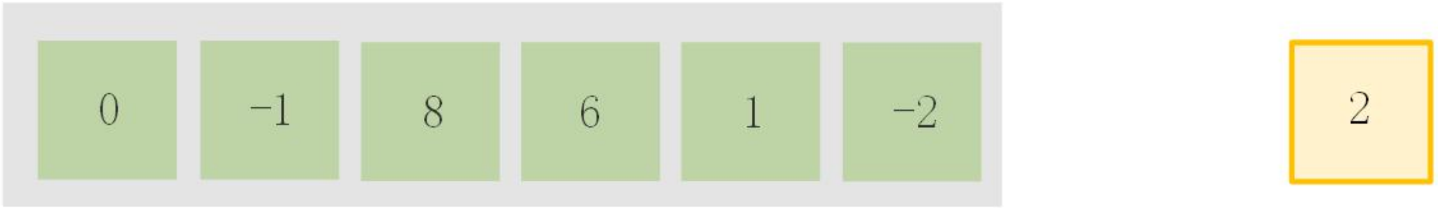


Concept: Pooling

Max pooling:

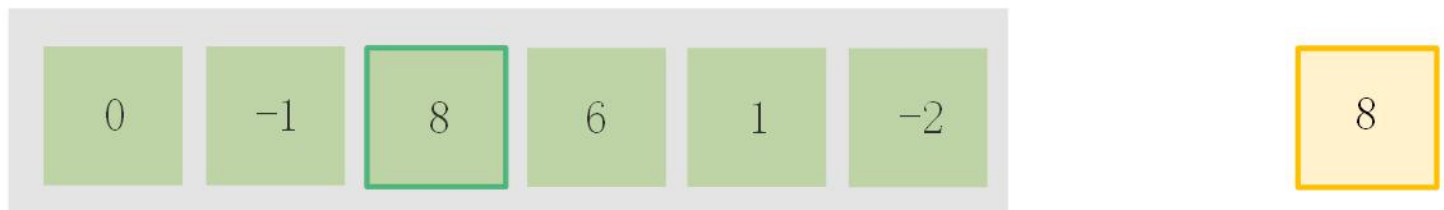


Mean pooling:

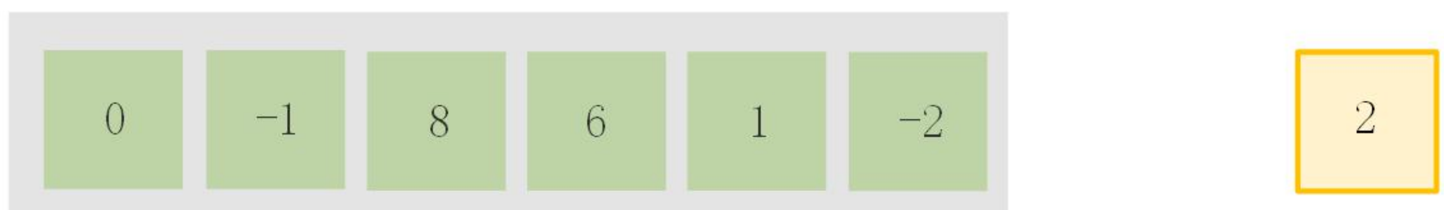


Concept: Pooling

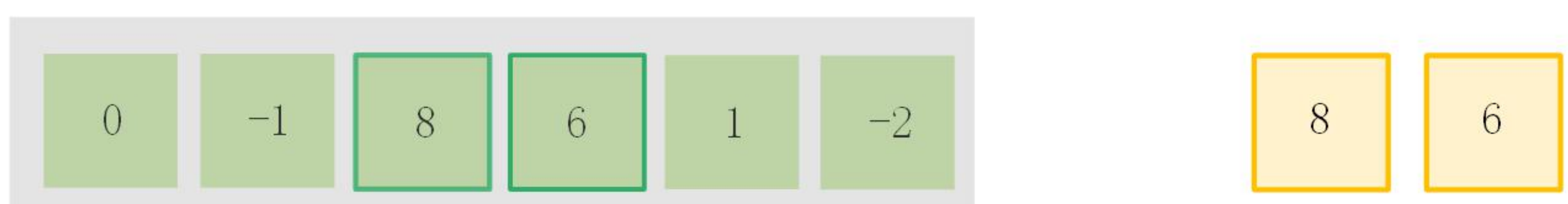
Max pooling:



Mean pooling:

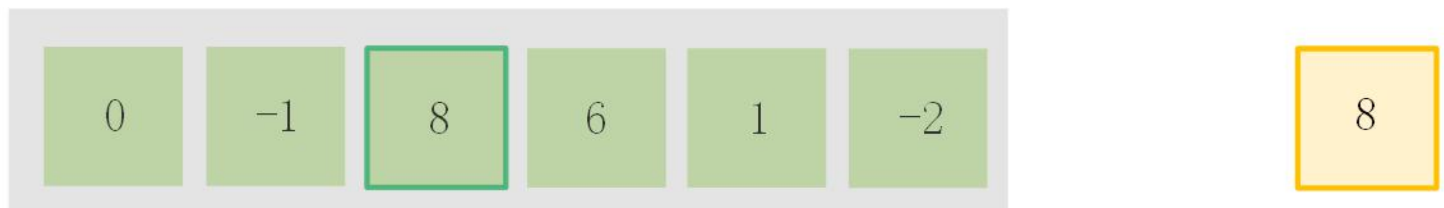


K-max pooling

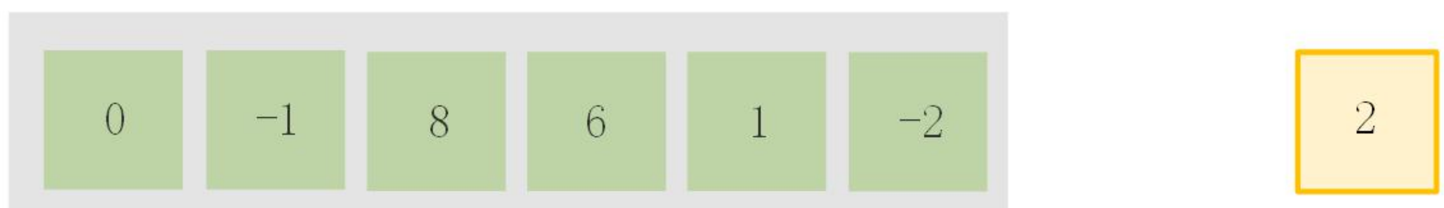


Concept: Pooling

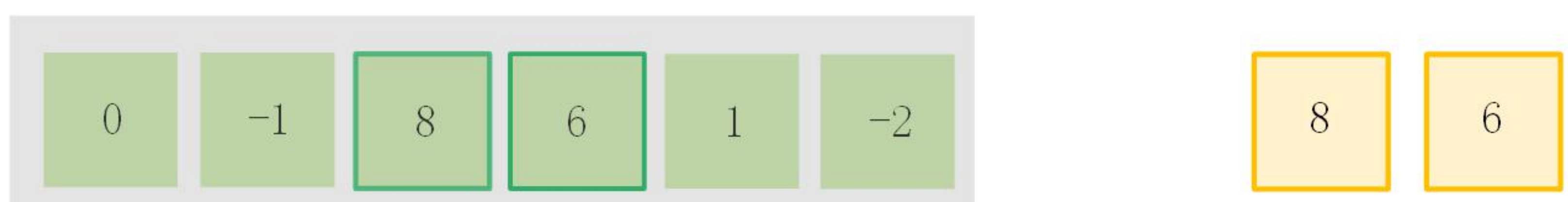
Max pooling:



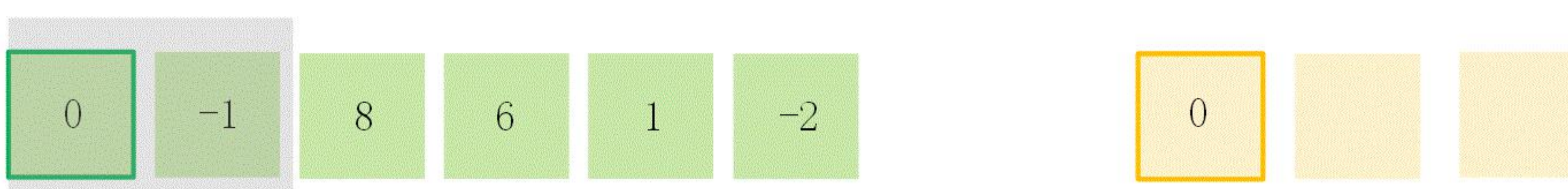
Mean pooling:



K-max pooling



Dynamic pooling:



Case Study:
Convolutional Networks for Text Classification (Kim 2015)

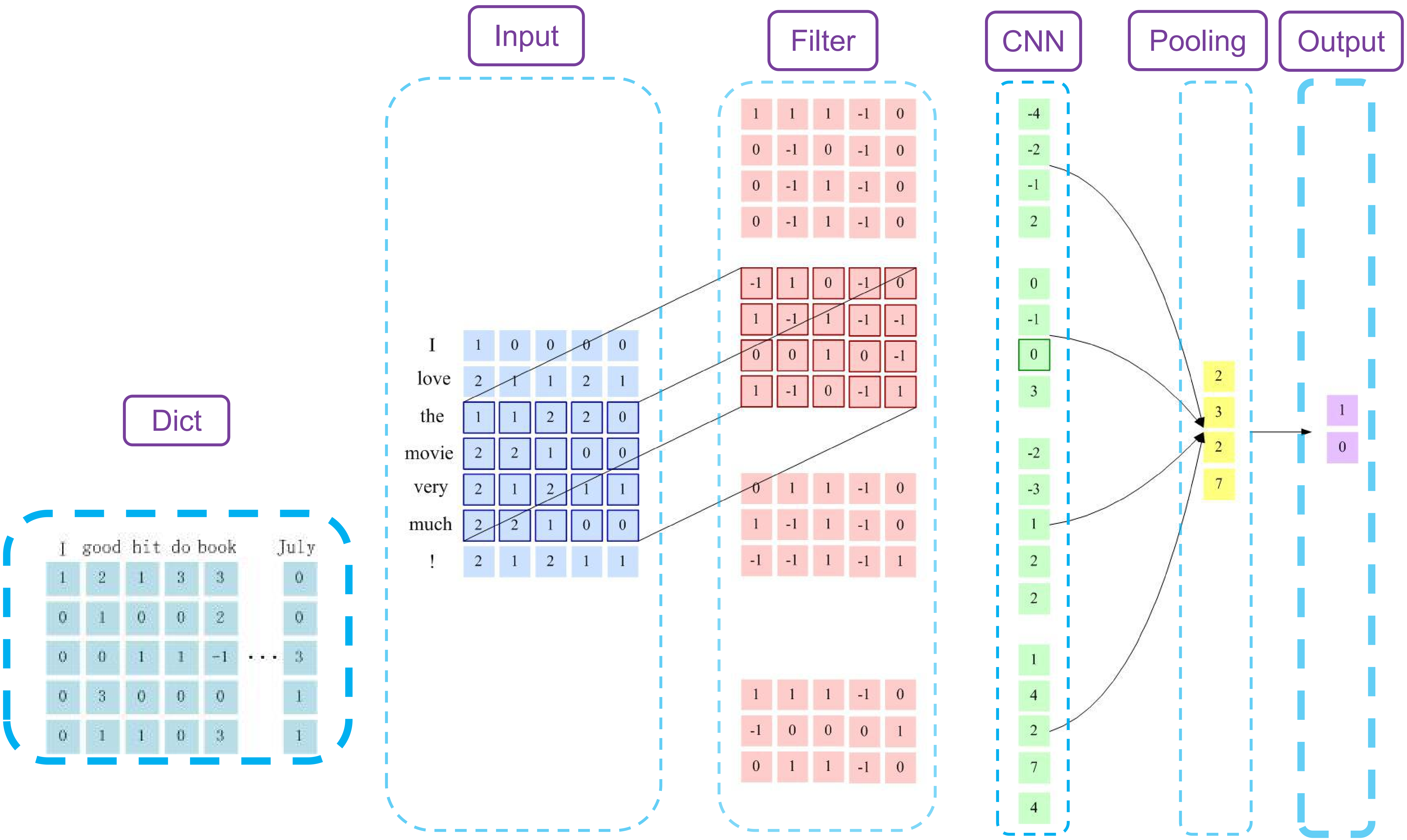
CNNs for Text Classification (Kim 2015)

- Task: sentiment classification
  - Input: a sentence
  - Output: a class label (positive/negative)

CNNs for Text Classification (Kim 2015)

- Task: sentiment classification
  - Input: a sentence
  - Output: a class label (positive/negative)
- Model:
  - Embedding layer
  - Multi-Channel CNN layer
  - Pooling layer/Output layer

Overview of the Architecture



Embedding Layer

- Build a look-up table (pre-trained? Fine-tuned?)
- Discrete → distributed

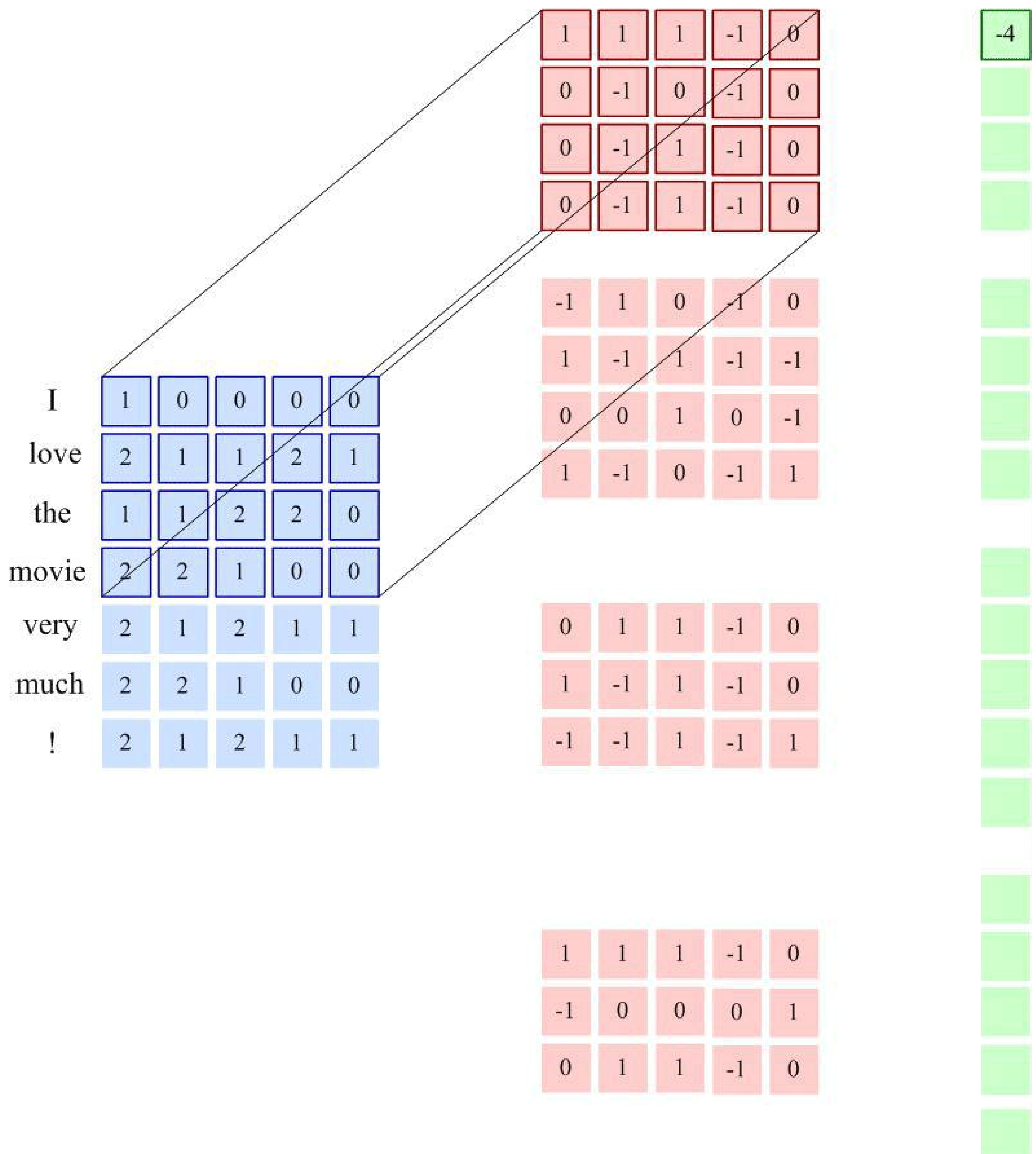
Input

I	1	0	0	0	0
love	2	1	1	2	1
the	1	1	2	2	0
movie	2	2	1	0	0
very	2	1	2	1	1
much	2	2	1	0	0
!	2	1	2	1	1

Look-up Table

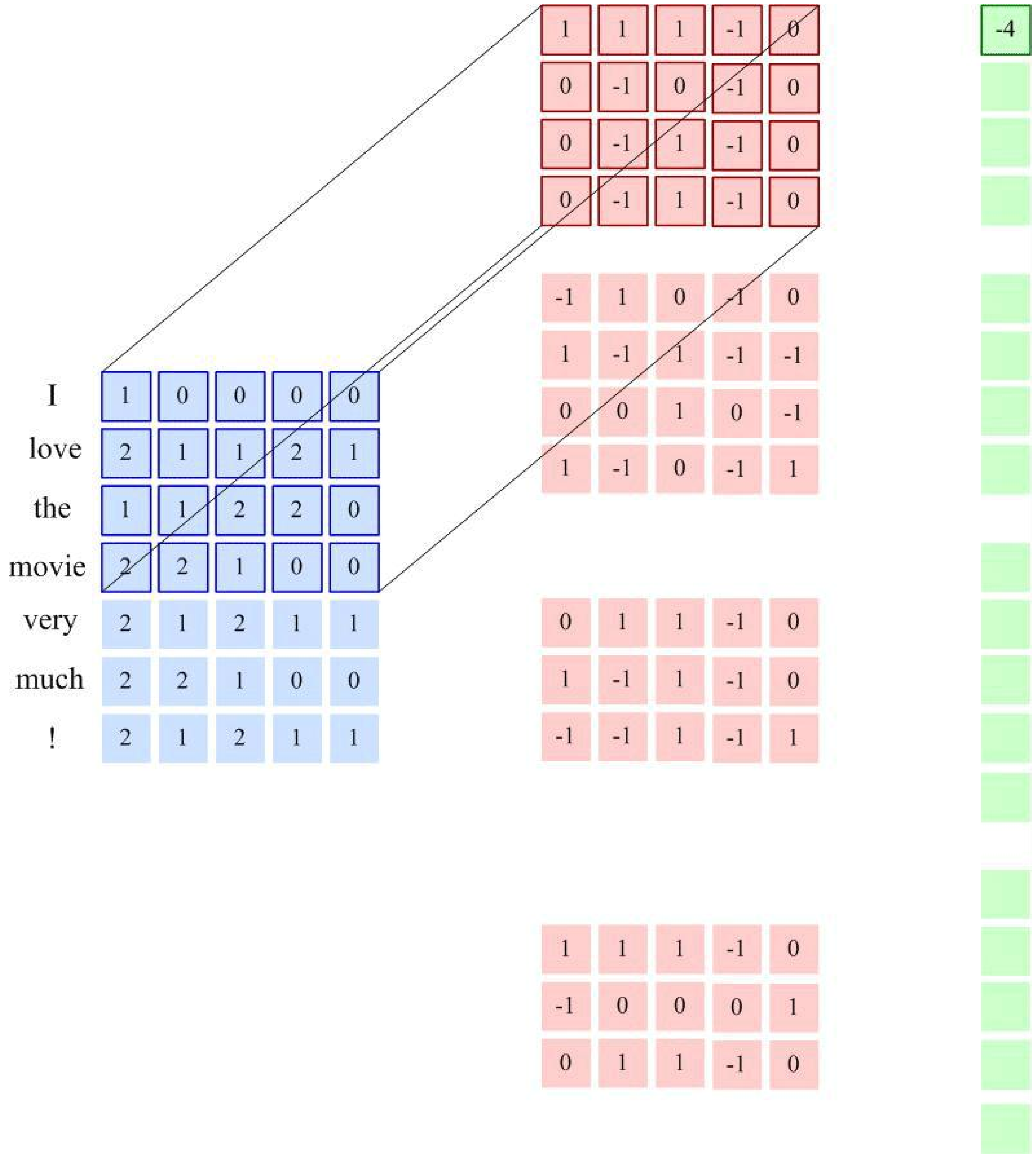
I	good	hit	do	book		July
1	2	1	3	3		0
0	1	0	0	2		0
0	0	1	1	-1	…	3
0	3	0	0	0		1
0	1	1	0	3		1

Conv. Layer



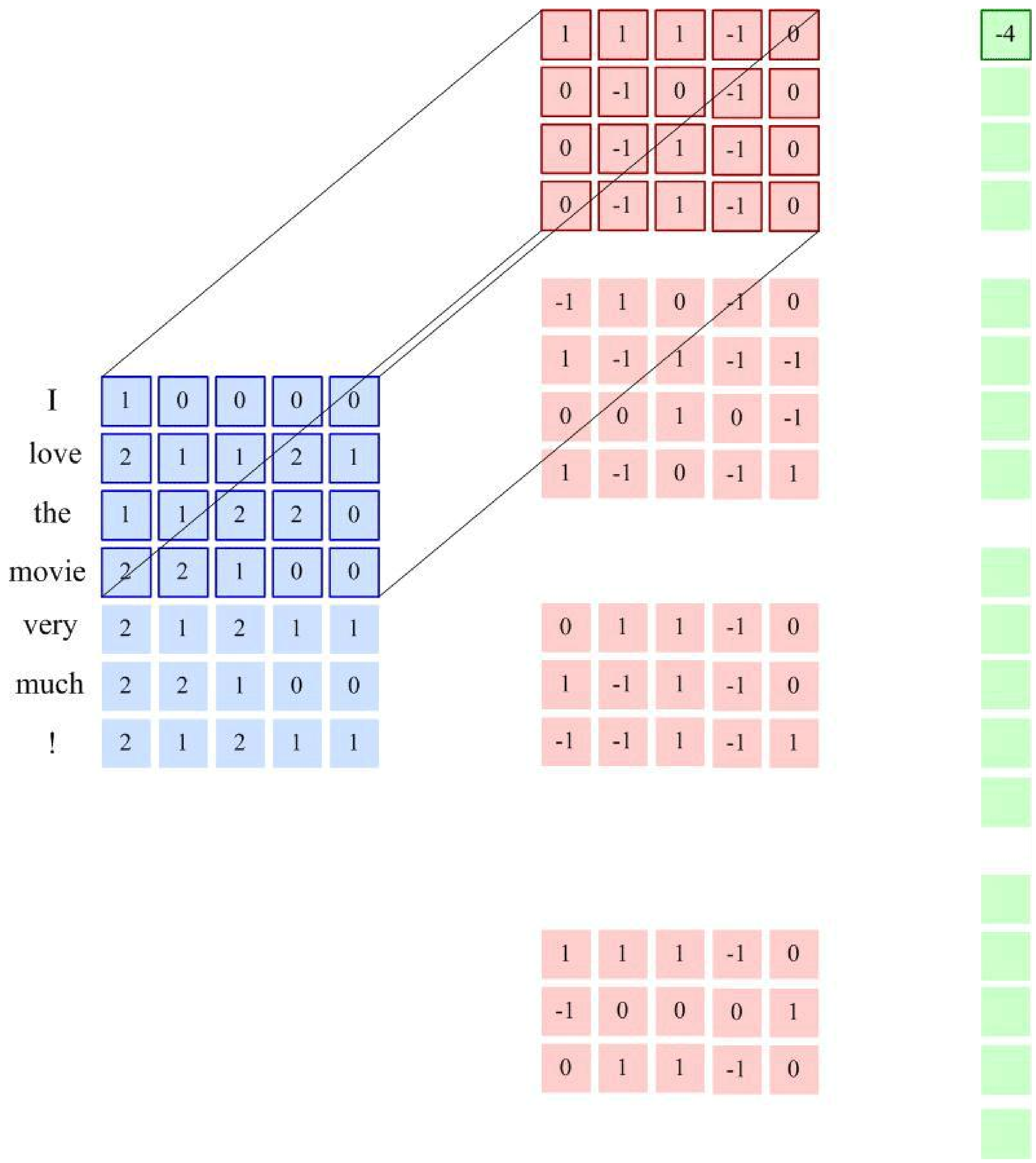
I 1 0 0 0 0
love 2 1 1 2 1
the 1 1 2 2 0
movie 2 2 1 0 0
very 2 1 2 1 1
much 2 2 1 0 0
! 2 1 2 1 1
1 1 1 -1 0
0 -1 0 -1 0
0 -1 1 -1 0
0 -1 1 -1 0
-1 1 0 -1 0
1 -1 1 -1 -1
0 0 1 0 -1
1 -1 0 -1 1
0 1 1 -1 0
1 -1 1 -1 0
-1 -1 1 -1 1
1 1 1 -1 0
-1 0 0 0 1
0 1 1 -1 0
-4

Conv. Layer



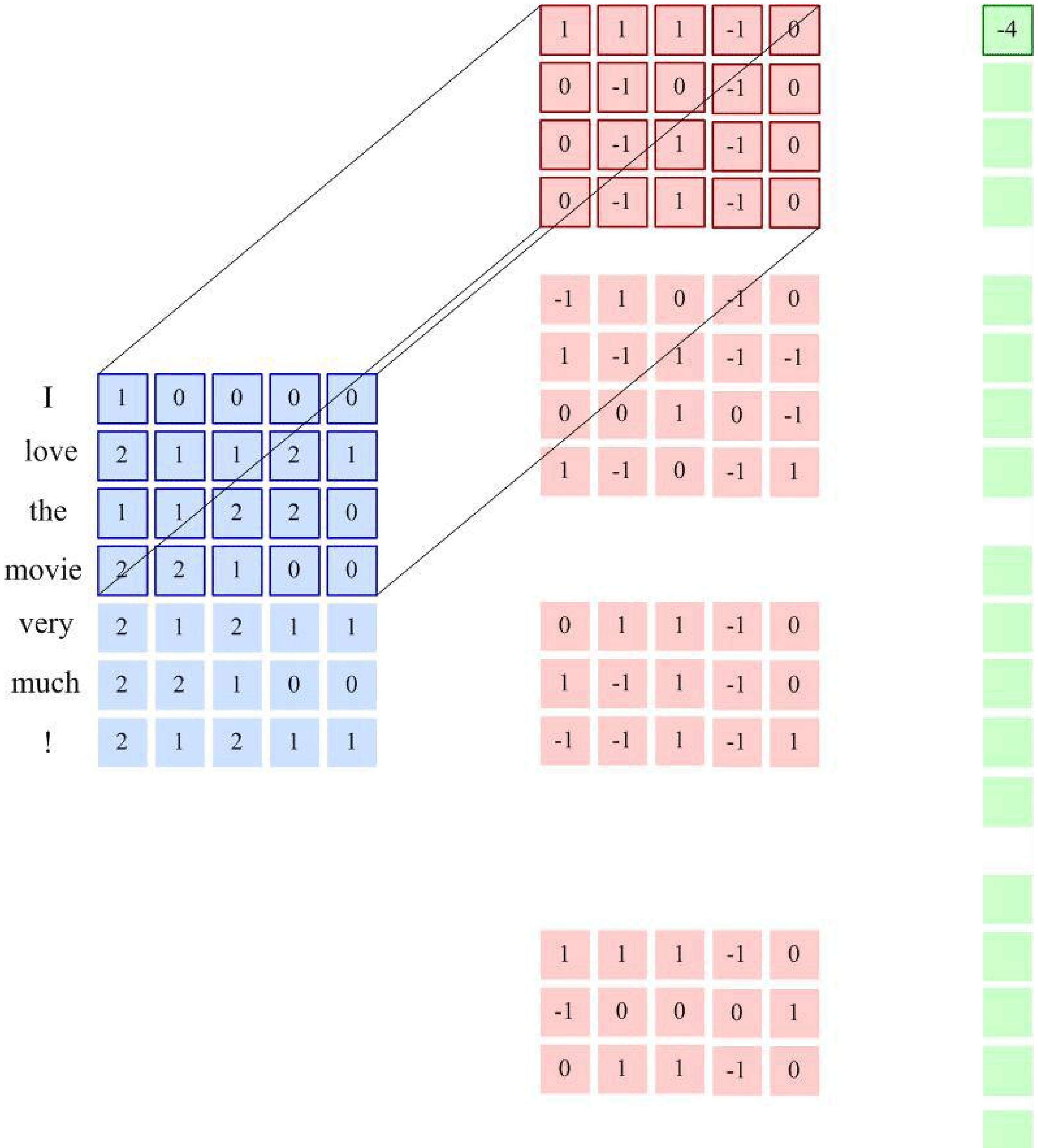
- Stride size?

Conv. Layer



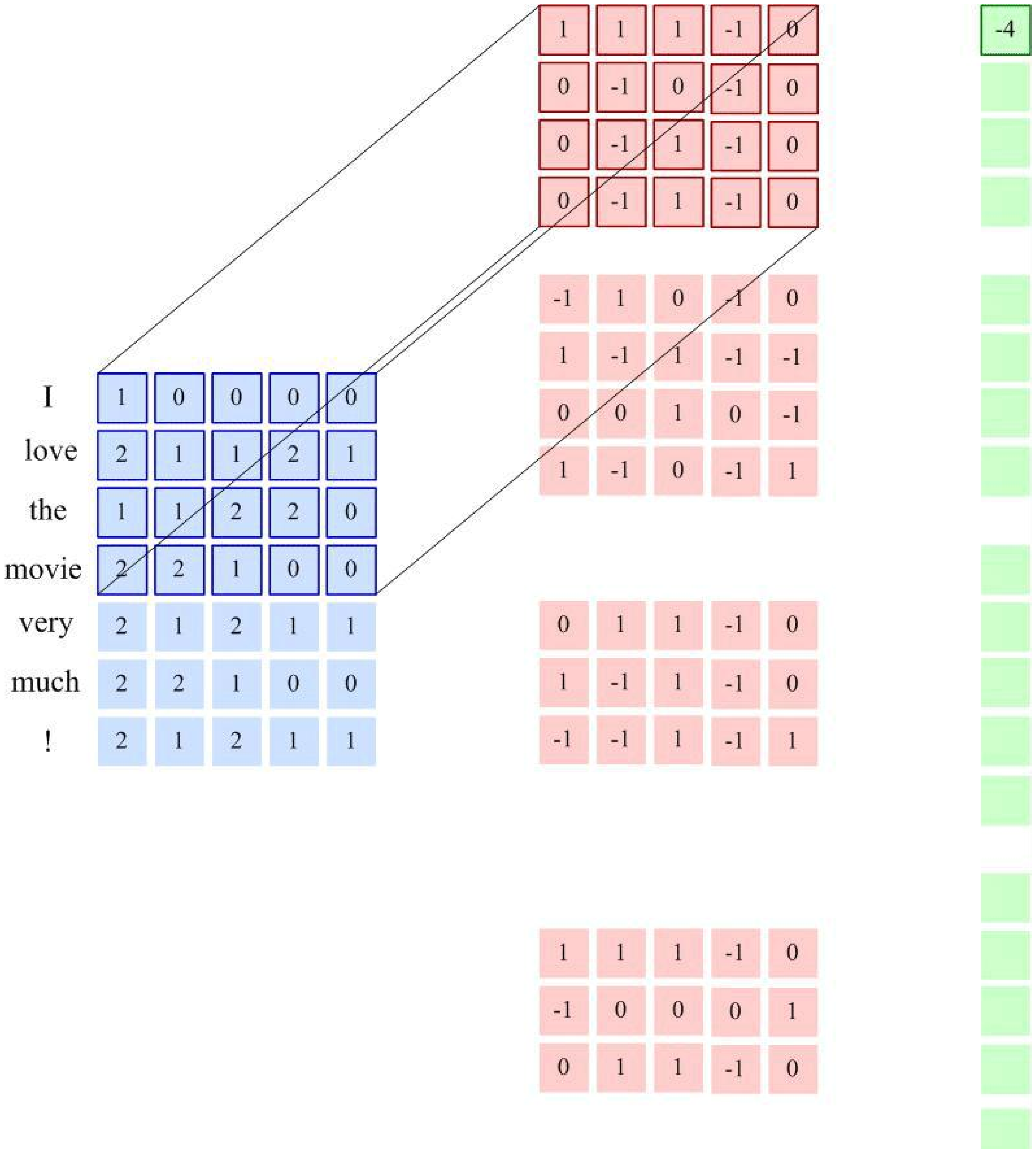
- Stride size?
  - 1

Conv. Layer



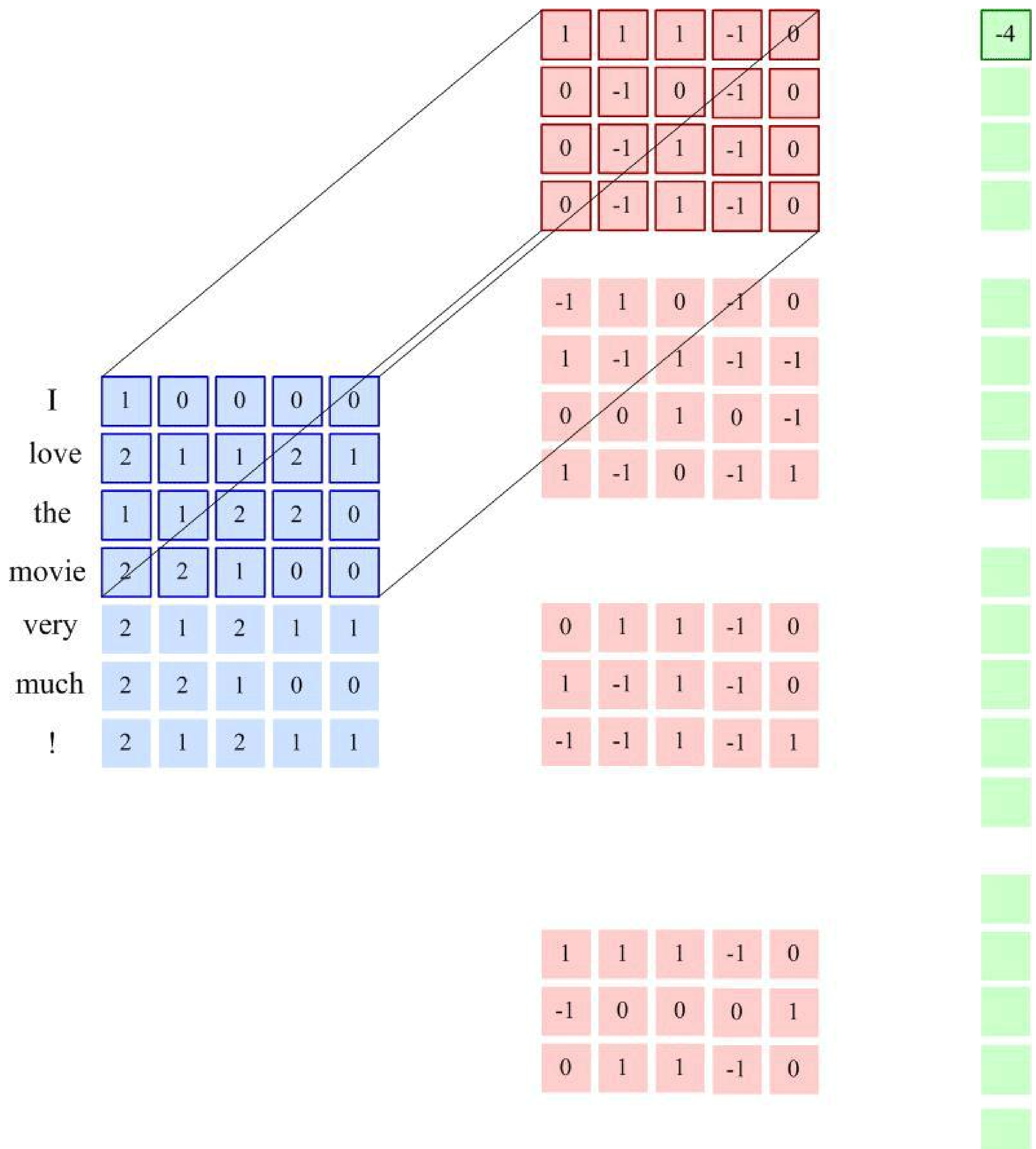
- Wide, equal, narrow?

Conv. Layer



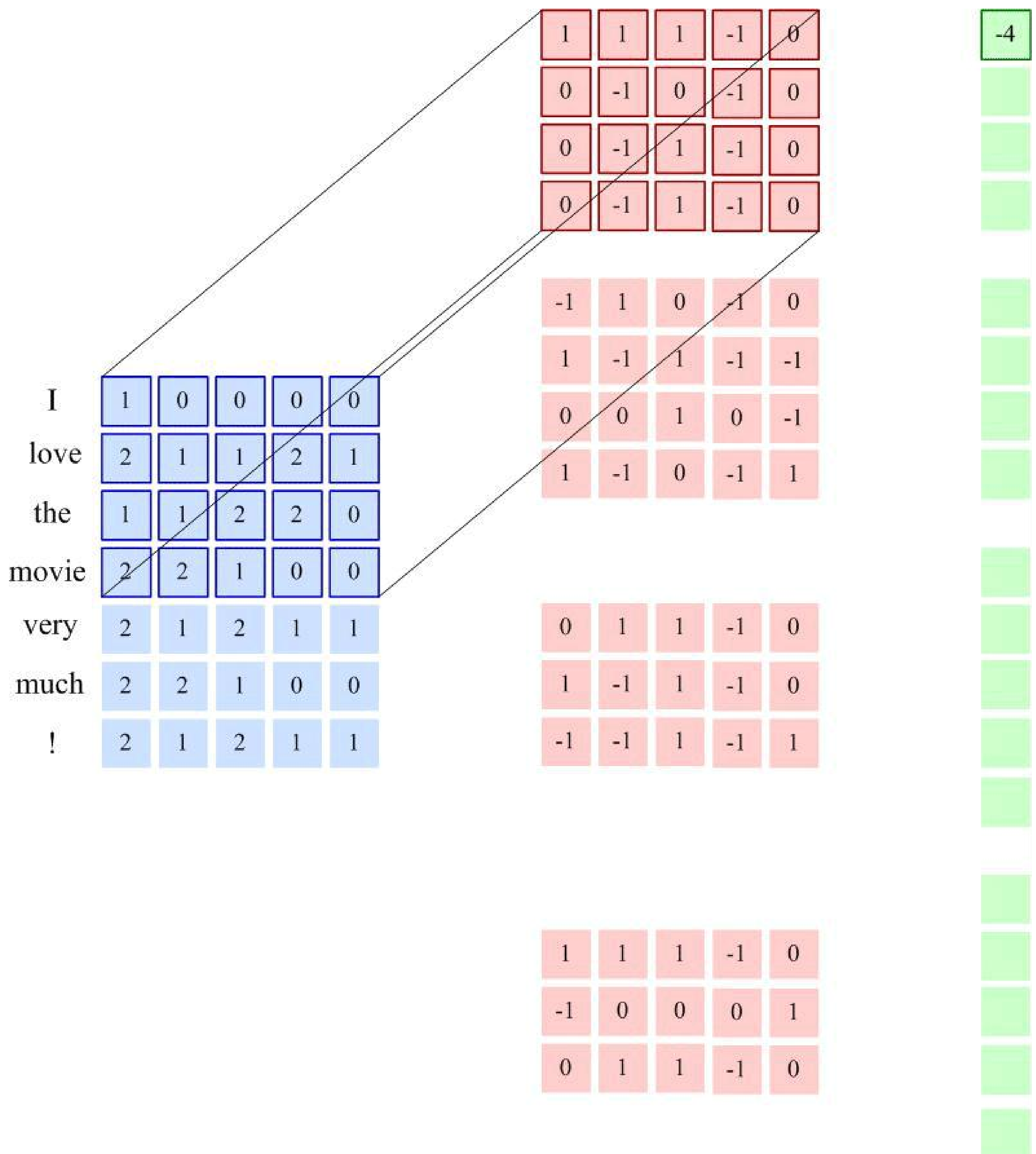
- Wide, equal, narrow?
  - narrow

Conv. Layer



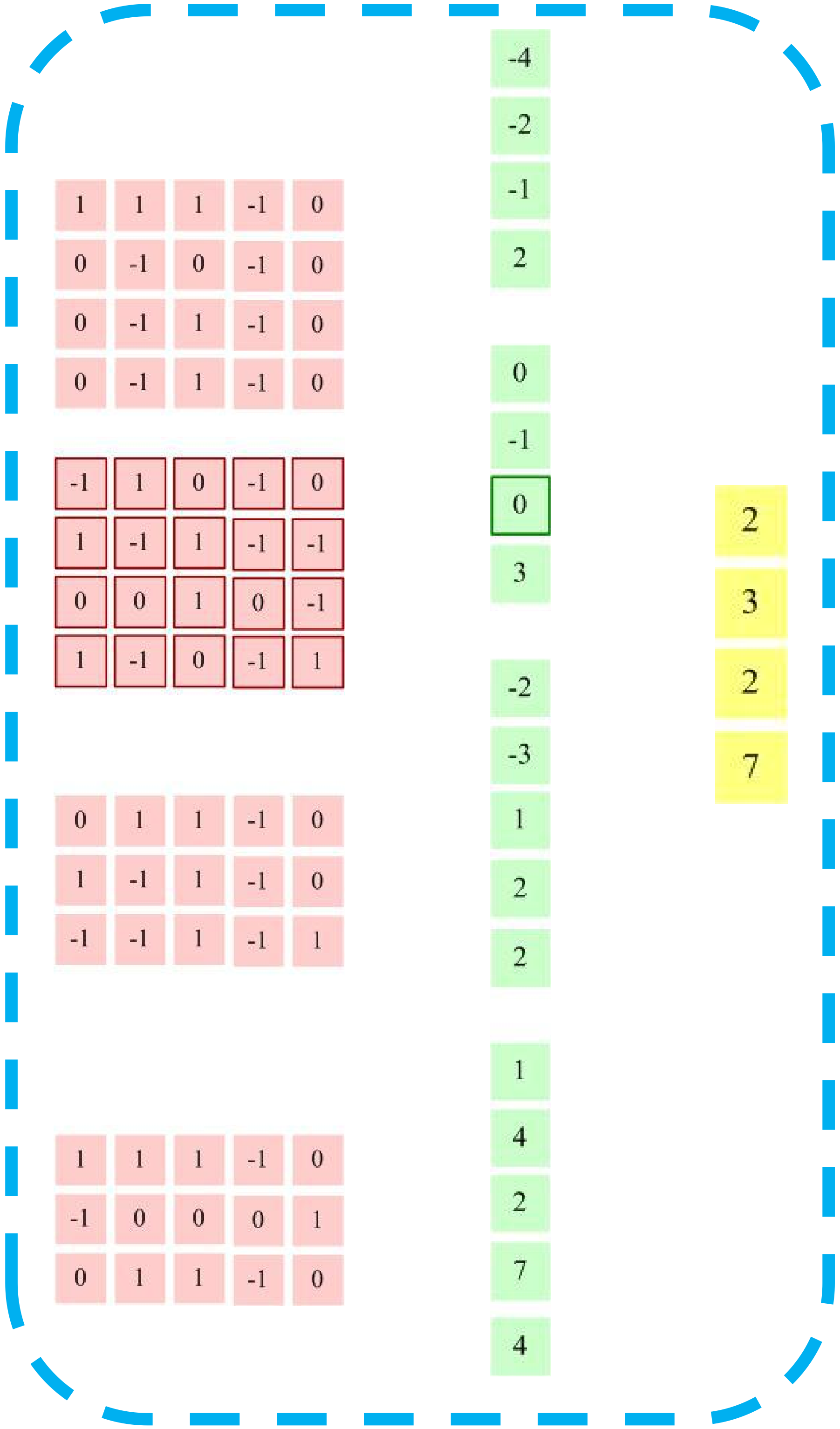
- How many filters?

Conv. Layer



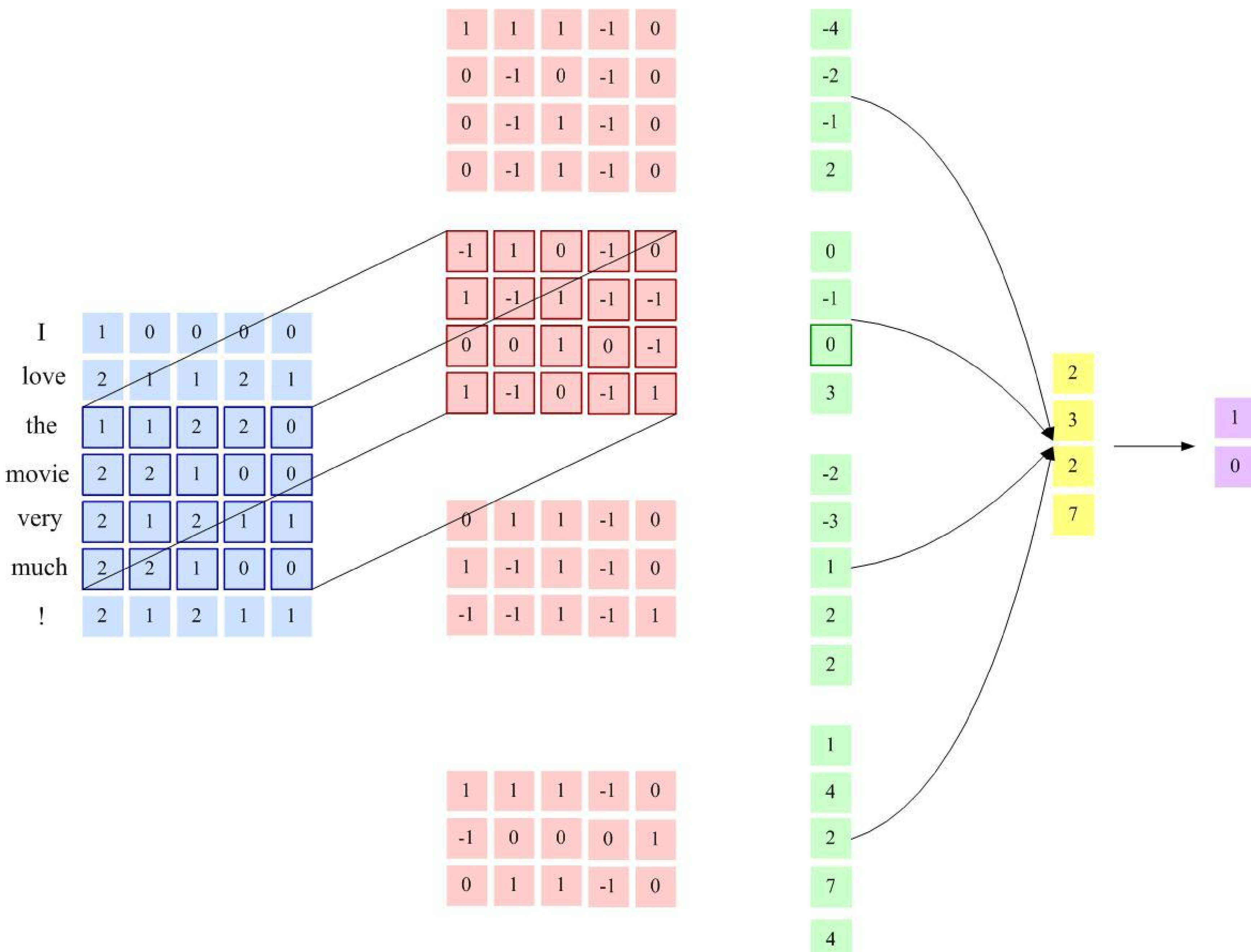
- How many filters?
  - 4

Pooling Layer



- Max-pooling
- Concatenate

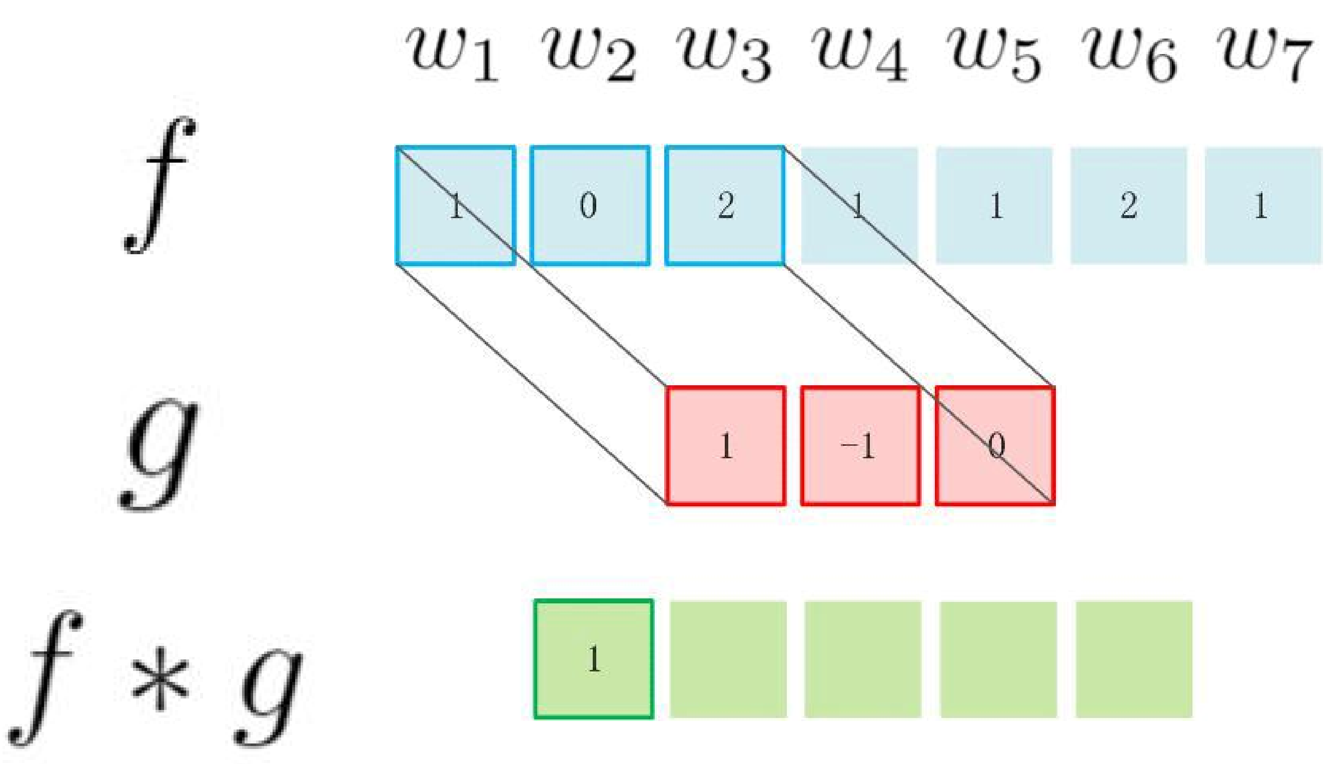
Output Layer



- MLP layer
- Dropout
- Softmax

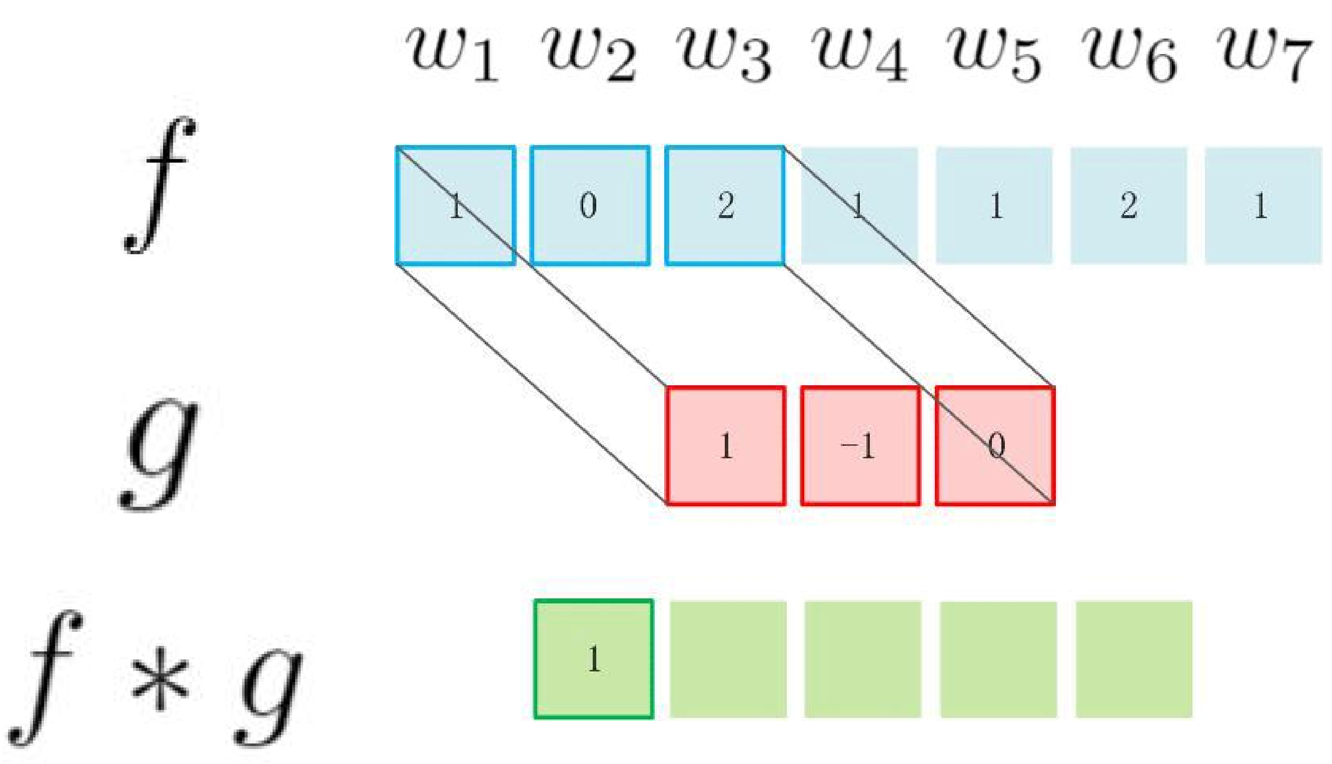
CNN Variants

Priori Entailed by CNNs



- Local bias
- Parameter sharing

Priori Entailed by CNNs

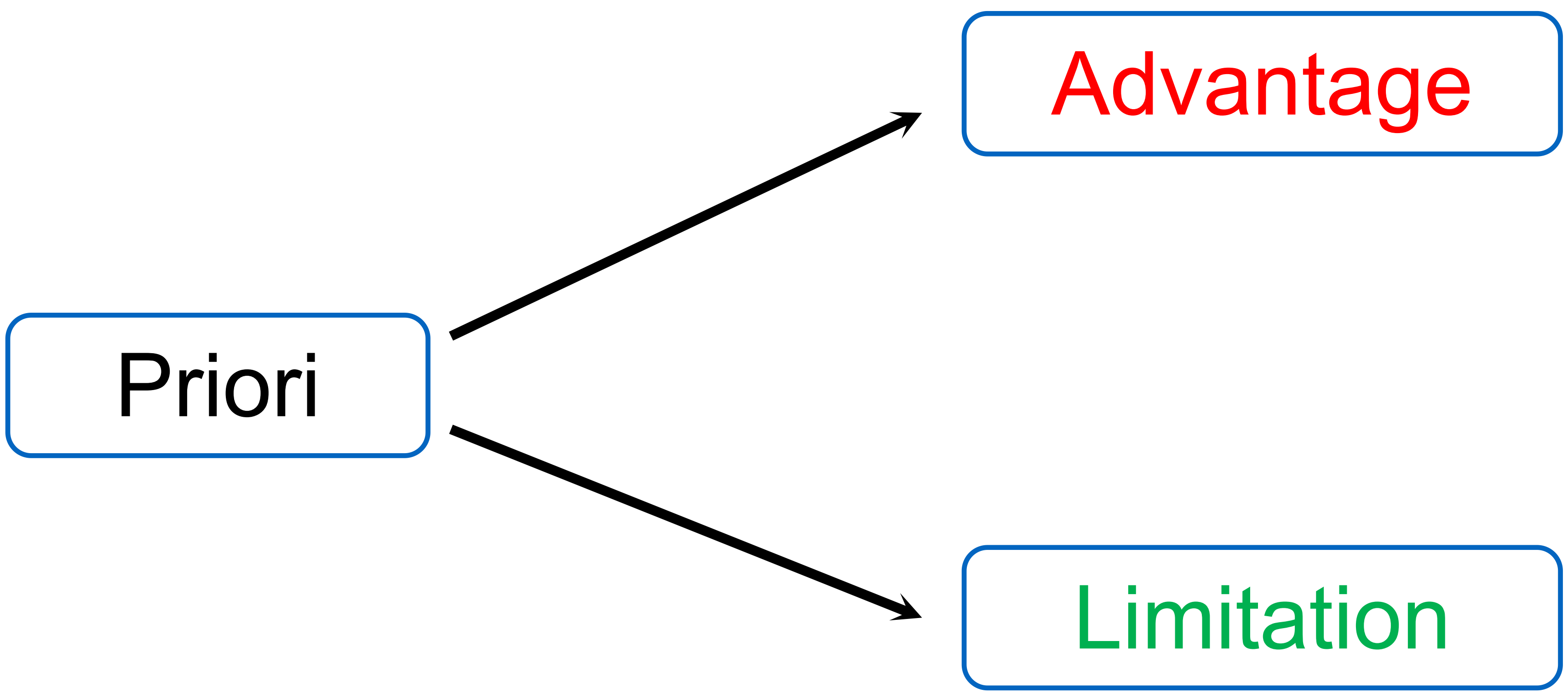


- Local bias
- Parameter sharing

How to handle long-term dependencies?

How to handle different types of compositionality?

Priori Entailed by CNNs



CNN Variants

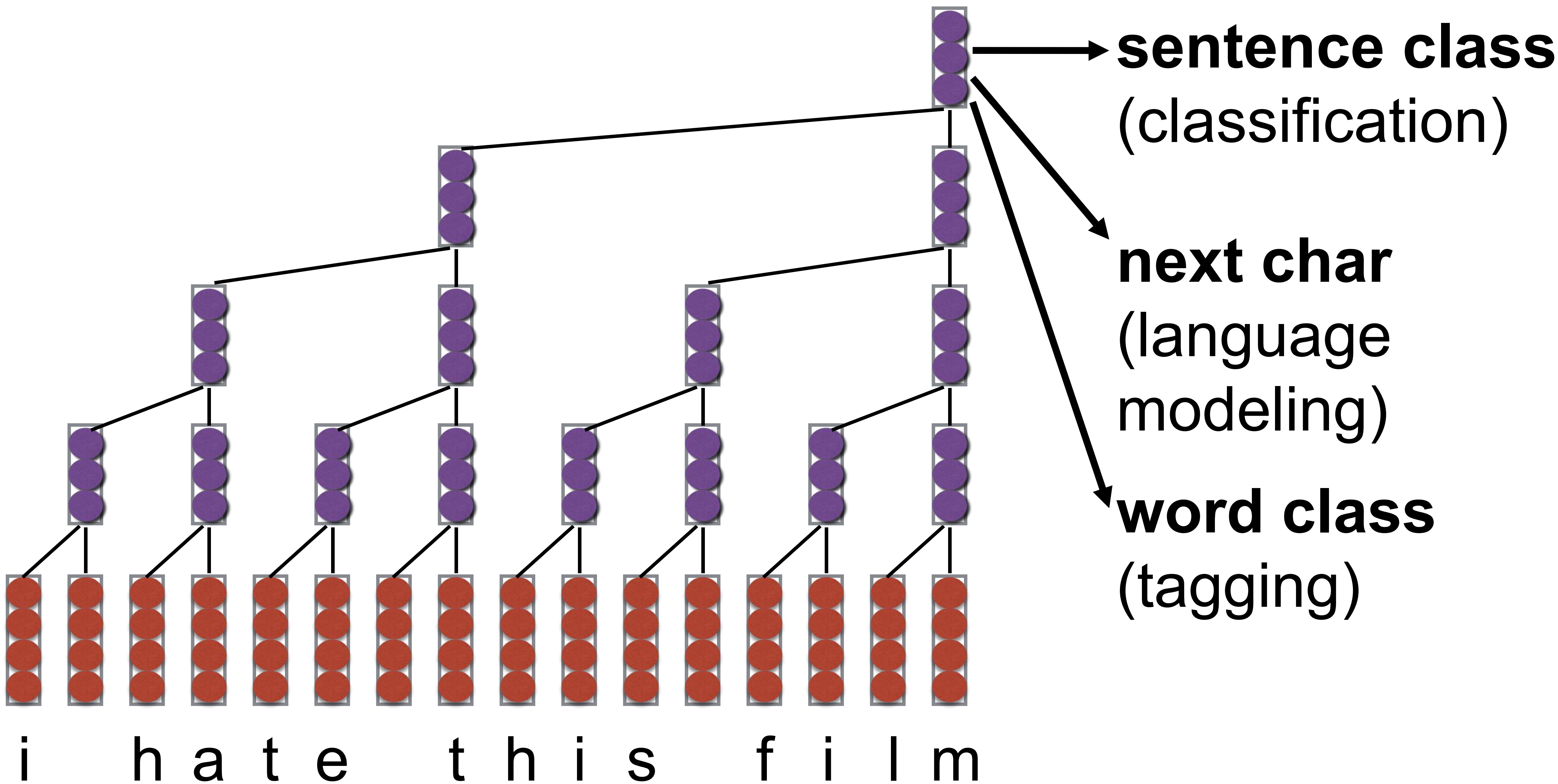
Locality Bias

- Long-term dependency
  - increase receptive fields (dilated)
- Complicated Interaction
  - dynamic filters

Sharing Parameters

Dilated Convolution
(e.g. Kalchbrenner et al. 2016)

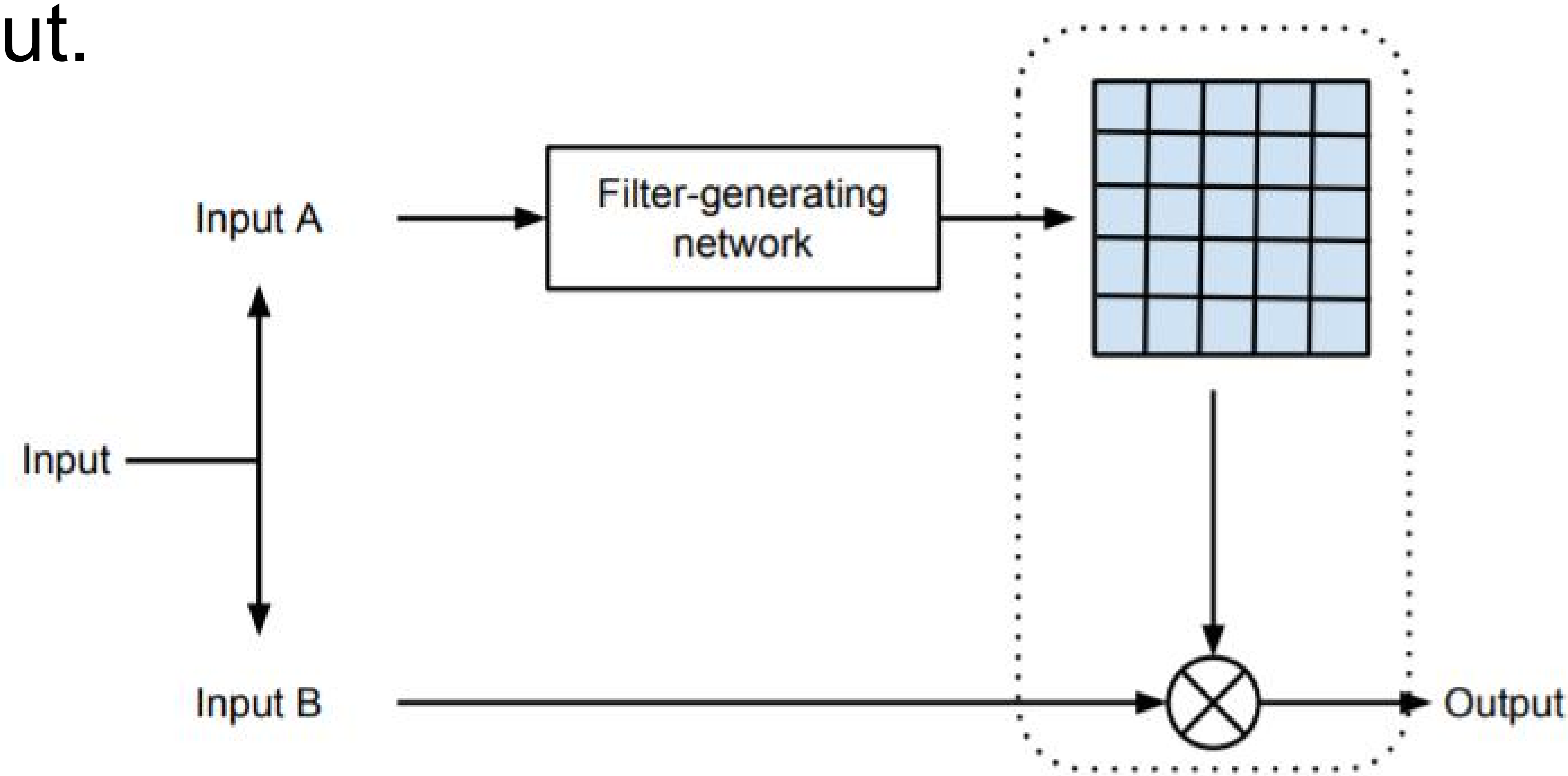
- Long-term dependency with less layers



Dynamic Filter CNN (e.g. Brabandere et al. 2016)

- Parameters of filters are static, failing to capture rich interaction patterns.

- Filters are generated dynamically conditioned on an input.



Common Applications

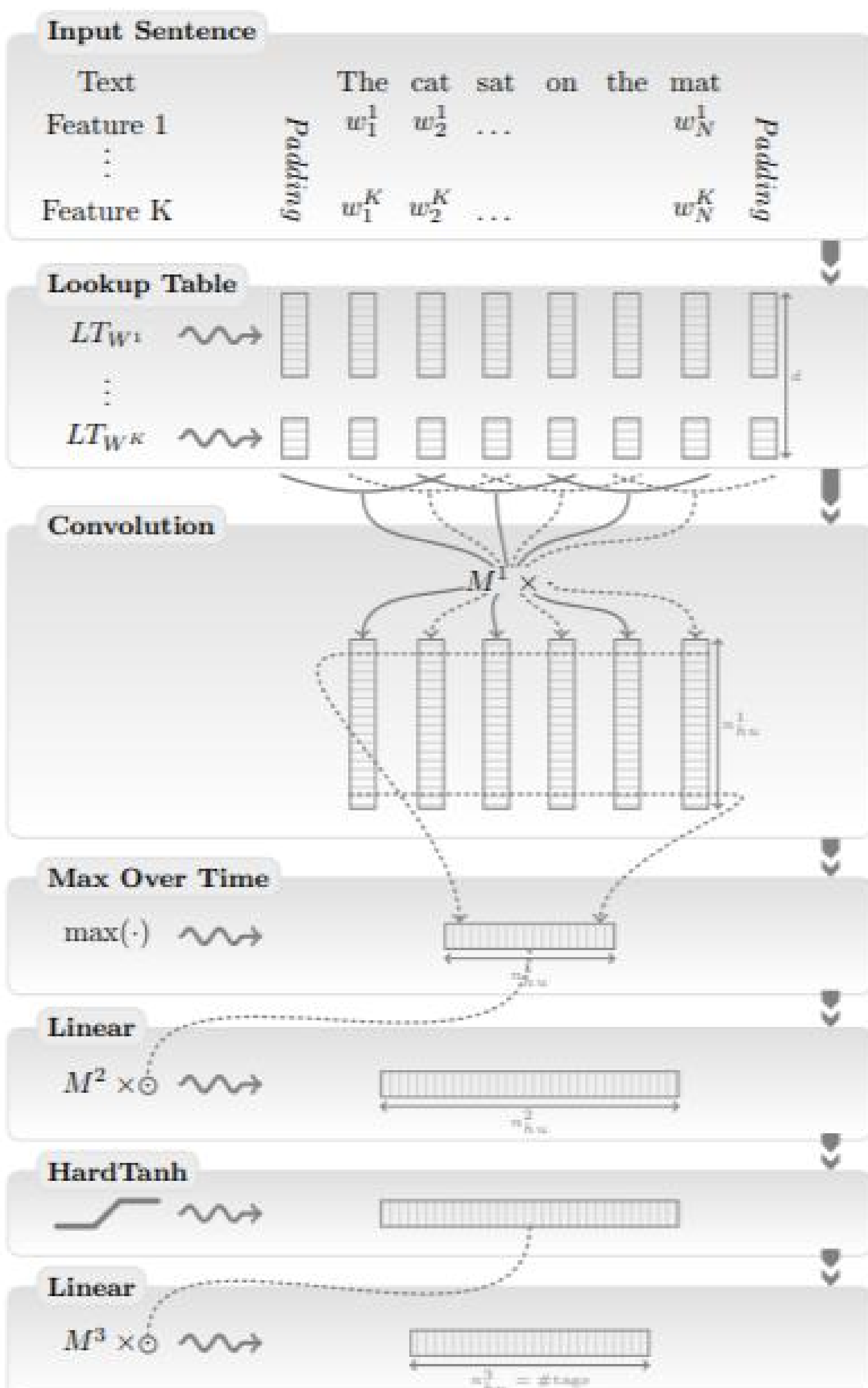
CNN Applications

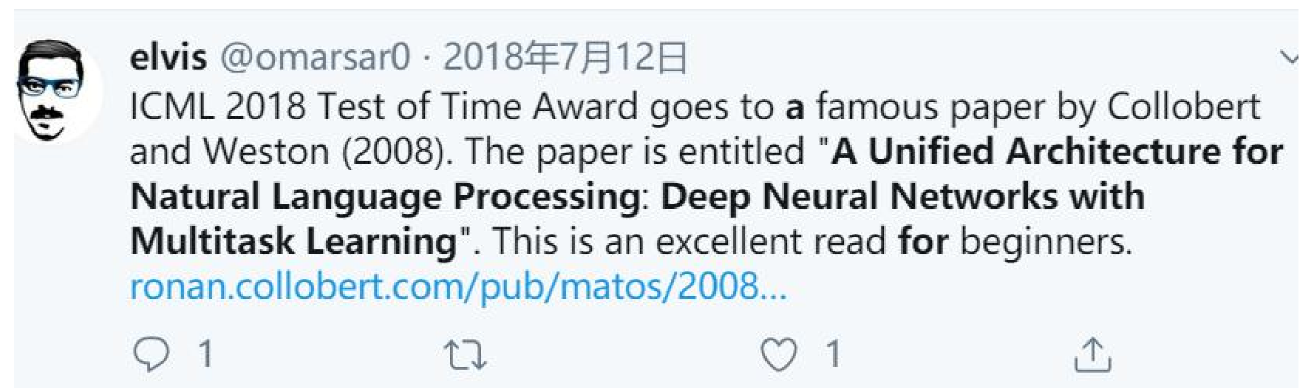
- **Word-level CNNs**
  - Basic unit: word
  - Learn the representation of a sentence
  - Phrasal patterns
- **Char-level CNNs**
  - Basic unit: character
  - Learn the representation of a word
  - Extract morphological patters

CNN Applications

- **Word-level CNN**
  - Sentence representation

NLP (Almost) from Scratch
(Collobert et al.2011)




elvis @omarsar0 · 2018年7月12日

ICML 2018 Test of Time Award goes to **a** famous paper by Collobert and Weston (2008). The paper is entitled "**A Unified Architecture for Natural Language Processing**: **Deep Neural Networks with Multitask Learning**". This is an excellent read **for** beginners. ronan.collobert.com/pub/matos/2008...

1 1

- One of the most important papers in NLP
- Proposed as early as 2008

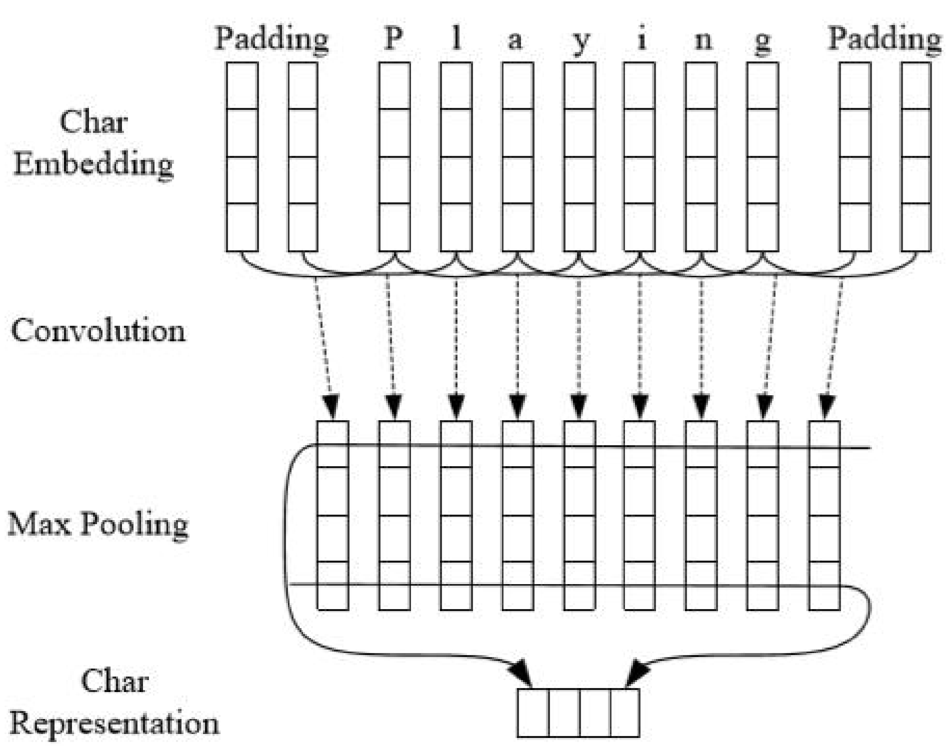
CNN Applications

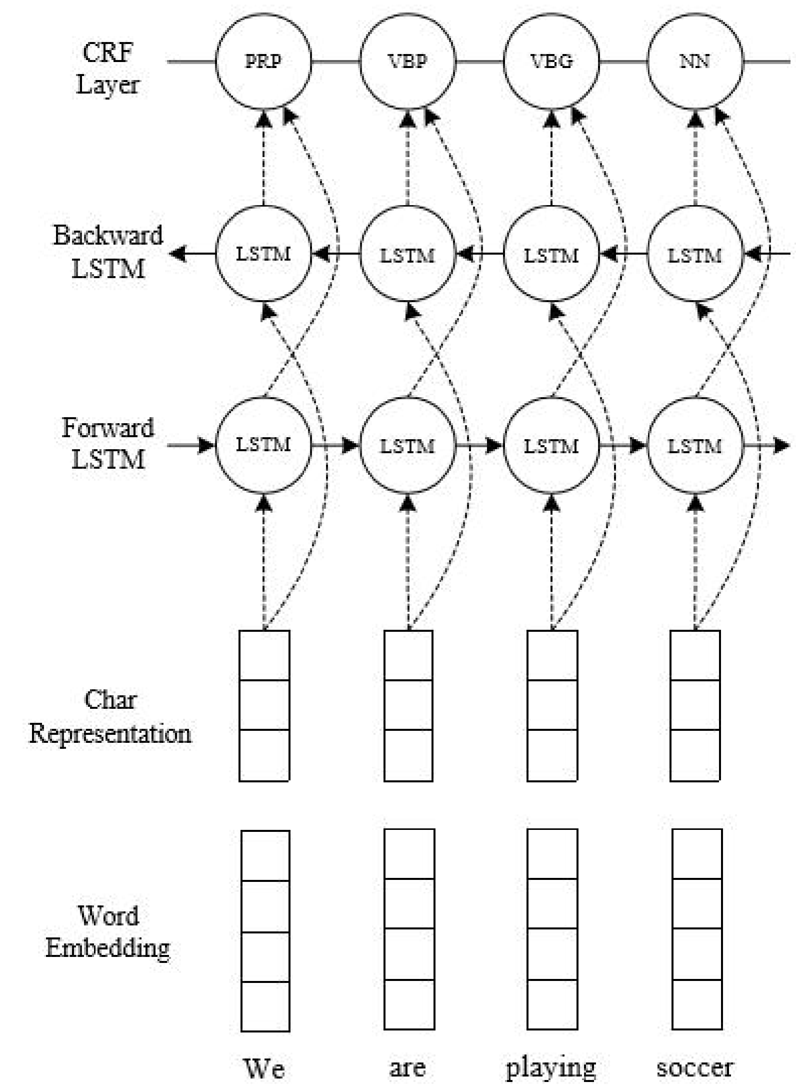
- **Word-level CNN**
  - Sentence representation
- **Char-level CNN**
  - Text Classification

CNN-RNN-CRF for Tagging
(Ma et al. 2016)

- A classic framework and de-facto standard for tagging

- Char-CNN is used to learn word representations (extract morphological information).
- Complementarity





Structured Convolution

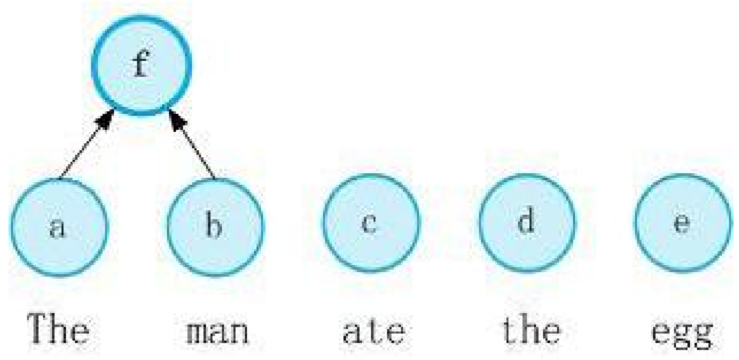
Why Structured Convolution?

The man ate the egg.

Why Structured Convolution?

The man ate the egg.

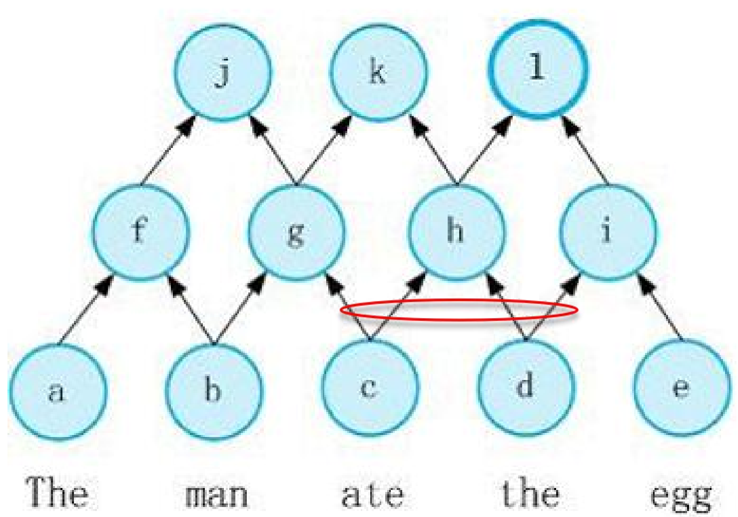
vanilla
CNNs



Why Structured Convolution?

The man ate the egg.

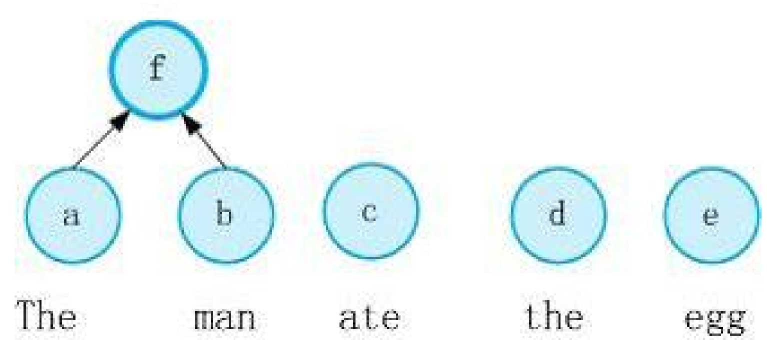
vanilla CNNs



- Some convolutional operations are not necessary
- e.g. noun-verb pairs very informative, but not captured by normal CNNs

Why Structured Convolution?

The man ate the egg.

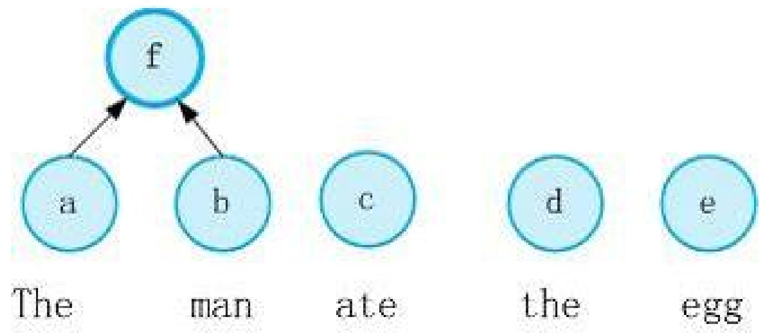


- Some convolutional operations are not necessary
- e.g. noun-verb pairs very informative, but not captured by normal CNNs
- Language has structure, would like it to localize features

Why Structured Convolution?

The man ate the egg.

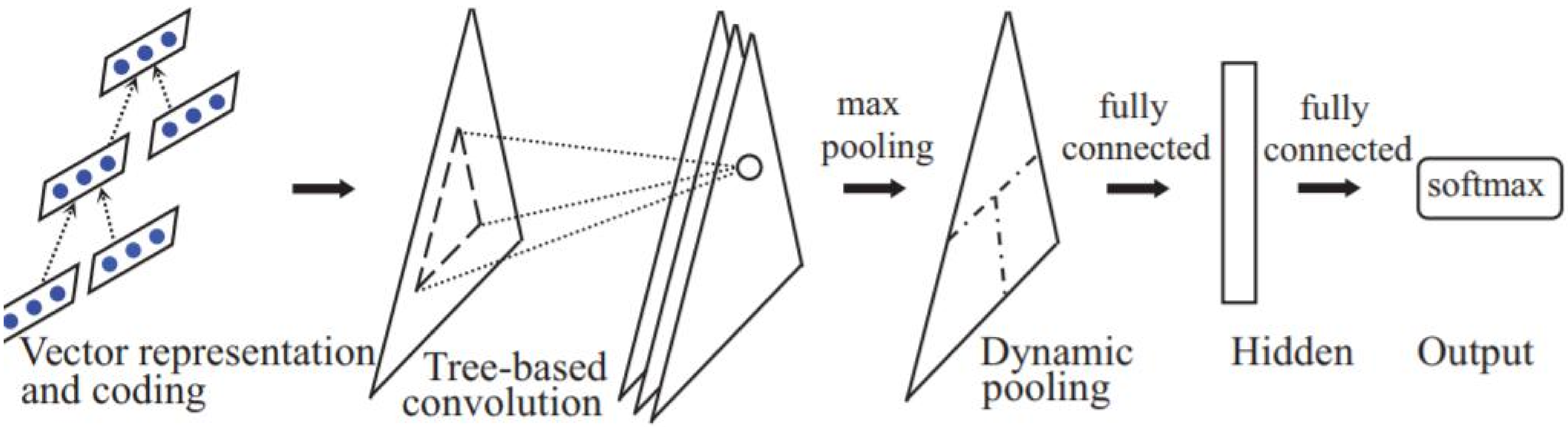
- Some convolutional operations are not necessary
- e.g. noun-verb pairs very informative, but not captured by normal CNNs
- Language has structure, would like it to localize features



The “Structure” provides stronger prior!

Tree-structured Convolution
(Mou et al. 2014, Ma et al. 2015)

- Convolve over parents, grandparents, siblings



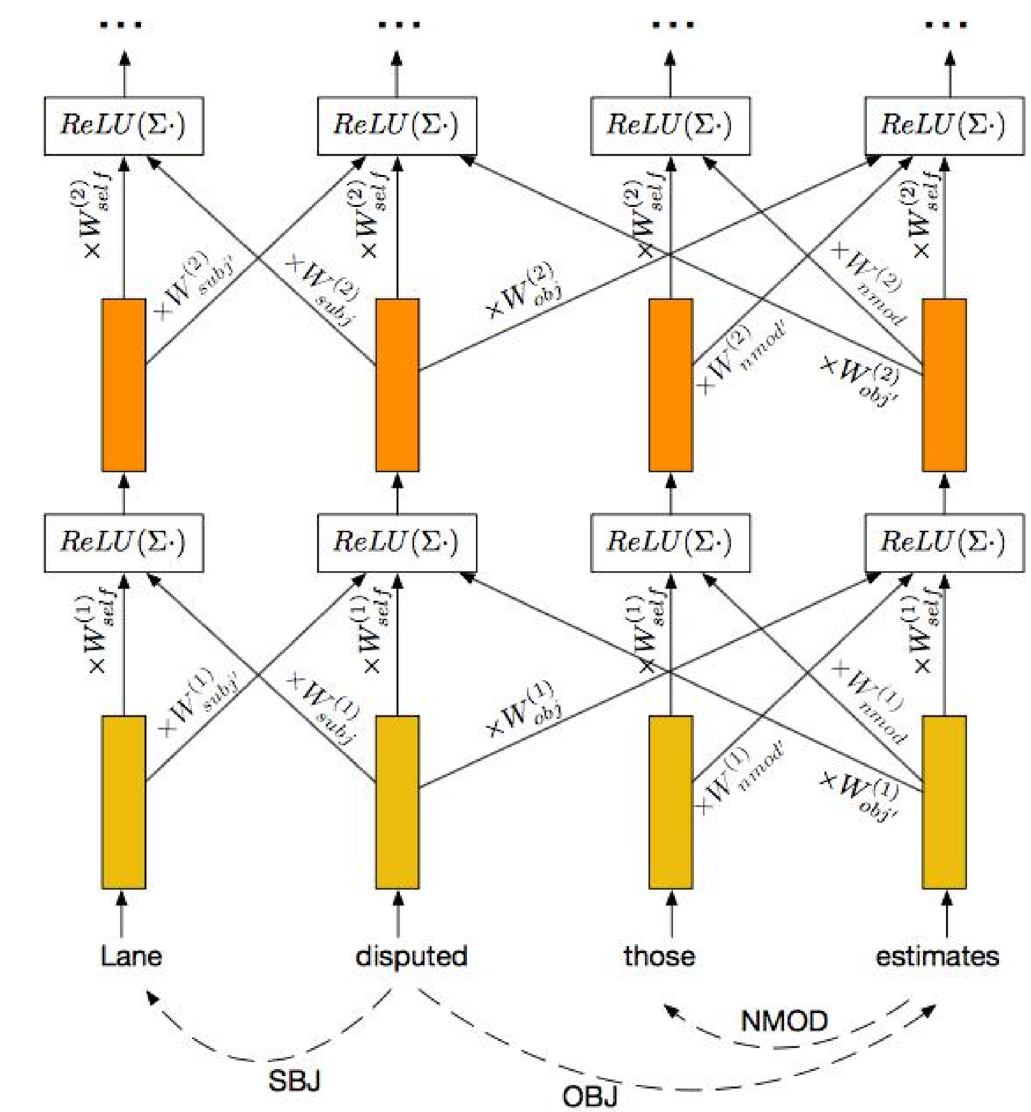
Graph Convolution
(e.g. Marcheggiani et al. 2017)

- Convolution is shaped by graph structure
- For example, dependency tree is a graph with

  1) Self-loop connection

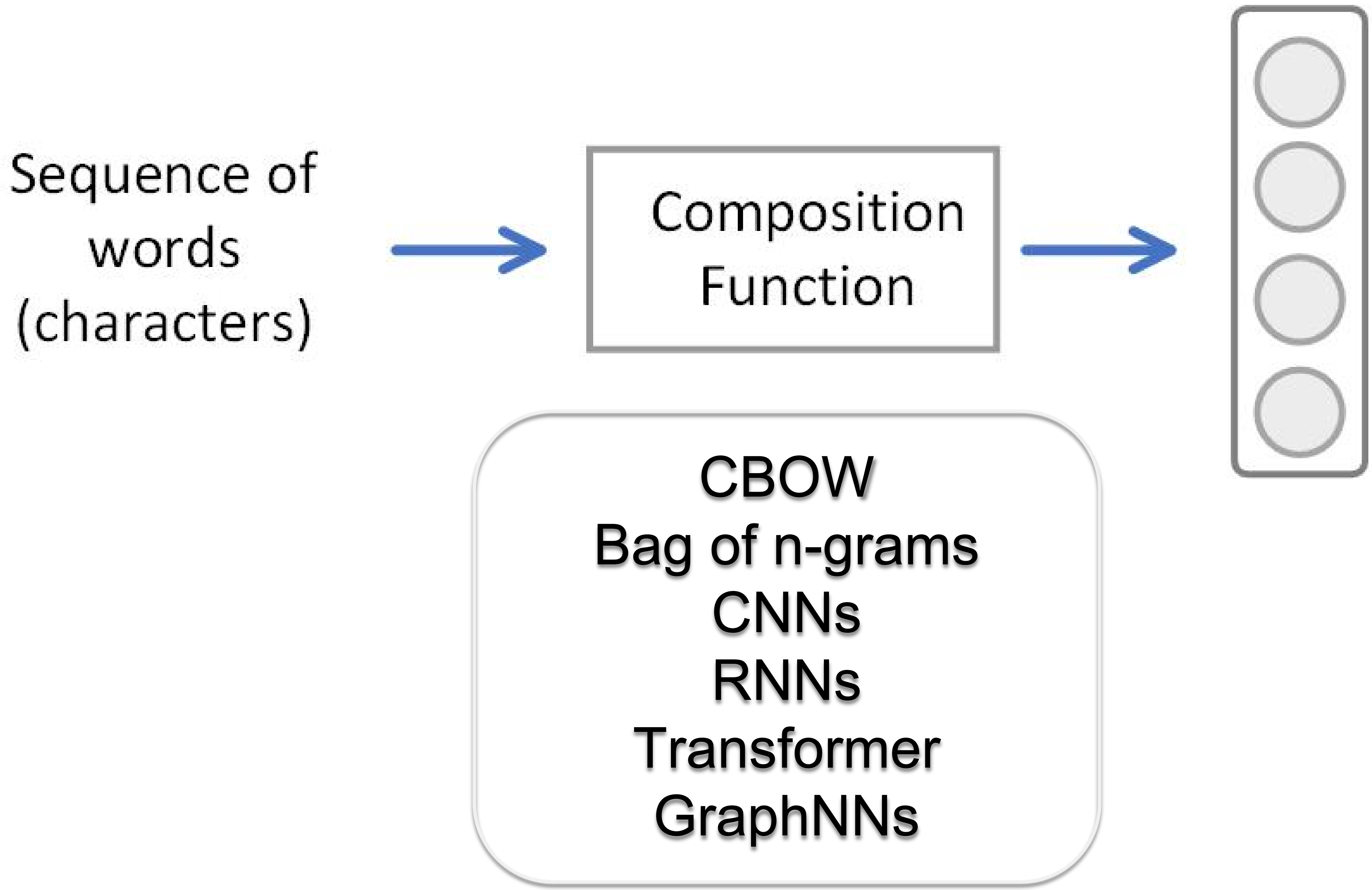
  2) Dependency connections

  3) Reverse connections

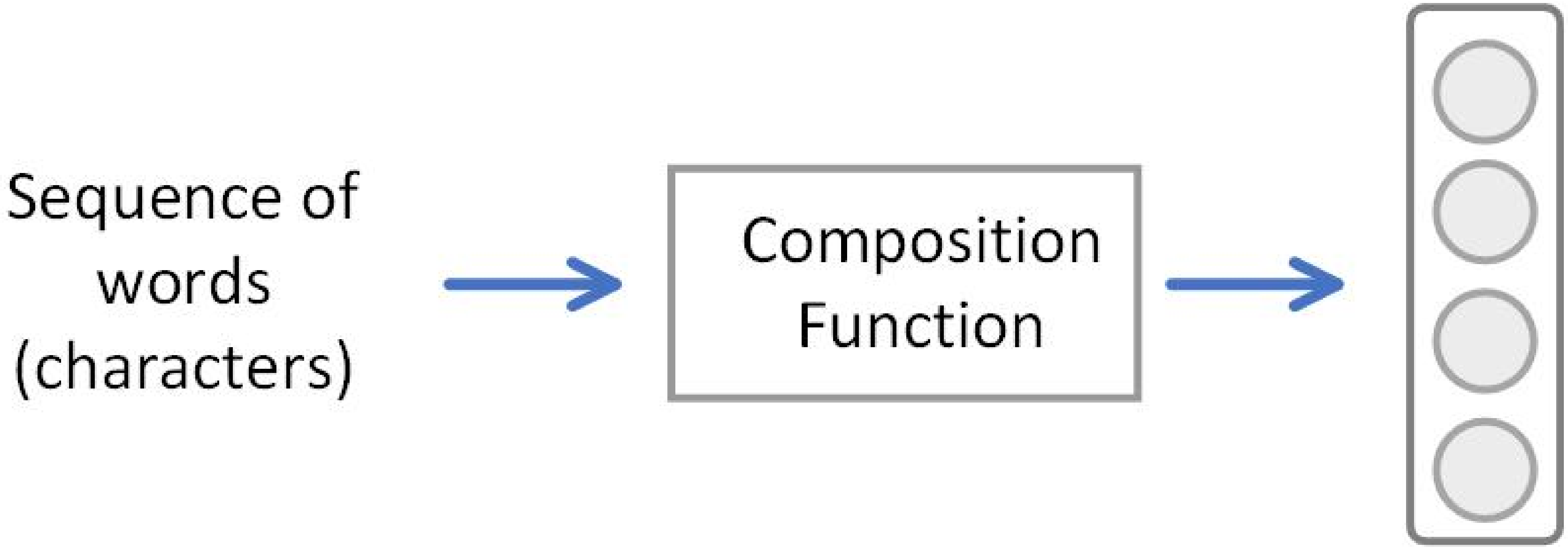


Summary

Neural Sequence Models



Neural Sequence Models



How do we make the choices of different neural sequence models?

Understand the design philosophy of a model

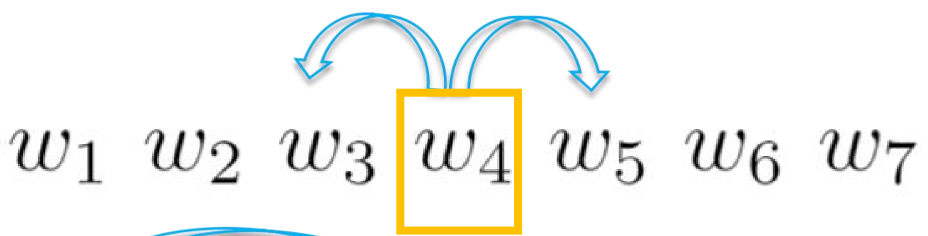
- **Inductive bias**: the set of assumptions that the learner uses to predict outputs given inputs that it has not encountered (from wikipedia)

- **Structural bias**: a set of prior knowledge incorporated into your model design

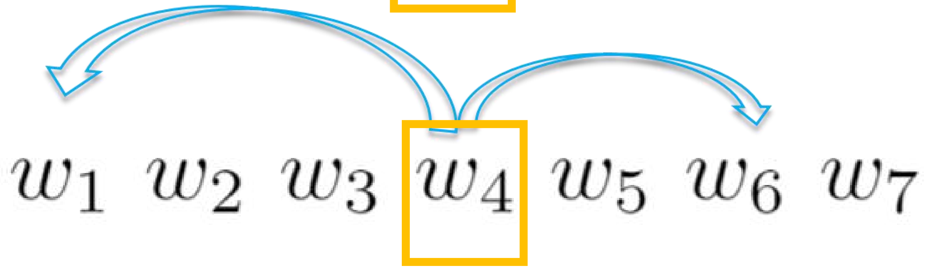
Structural Bias

- **Structural bias**: a set of prior knowledge incorporated into your model design
  - Locality

    Local



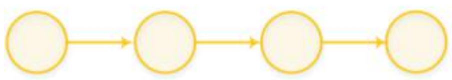
    Non-local



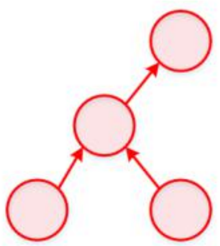
Structural Bias

- **Structural bias**: a set of prior knowledge incorporated into your model design
  - Topological structure

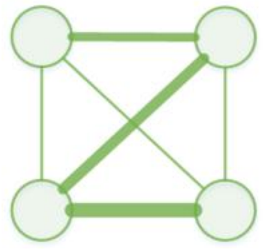
Sequential



Tree



Graph



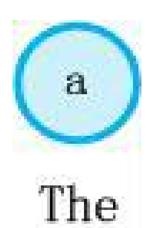
What inductive bias does a neural component entail?

Locality Bias	Local	Non-local	
Topological Structure	Seq.	Tree	Graph

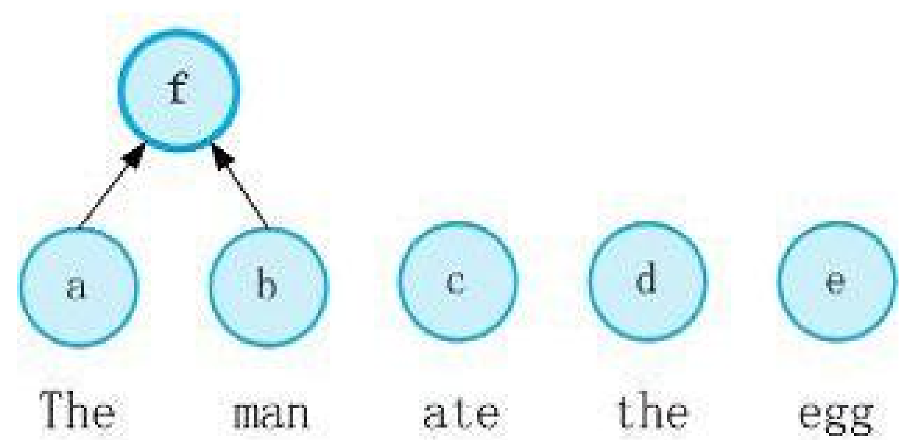
What inductive bias does a neural component entail?

Locality Bias | Local | Non-local

Topological Structure | Seq. | Tree | Graph



RNN

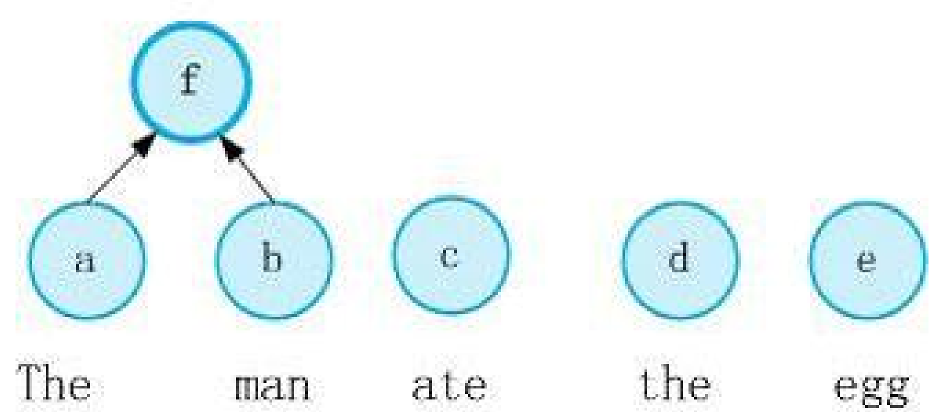


CNN

What inductive bias does a neural component entail?

Locality Bias | Local | Non-local

Topological Structure | Seq. | Tree | Graph



Structured CNN

What inductive bias does a neural component entail?

Locality Bias — Local | Non-local

Topological Structure — Seq. | Tree | Graph

?

What inductive bias does a neural component entail?

Locality Bias | Local | Non-local

Topological Structure | Seq. | Tree | Graph

?

What inductive bias does a neural component entail?

Locality Bias	Local	Non-local	
Topological Structure	Seq.	Tree	Graph

?

Questions?