A Convolutional Neural Network for Modelling Sentences

ADVISOR: JIA-LING, KOH

SOURCE: ACL 2014

SPEAKER: SHAO-WEI, HUANG

DATE: 2019/11/01

OUTLINE

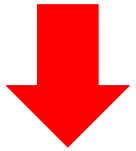
- Introduction
- Method
- Experiment
- Conclusion

INTRODUCTION

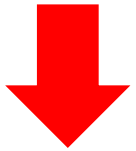
- Sentence model：to analyse and represent the semantic content of a sentence.

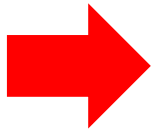
(Ex):

The cat sat on the red mat.



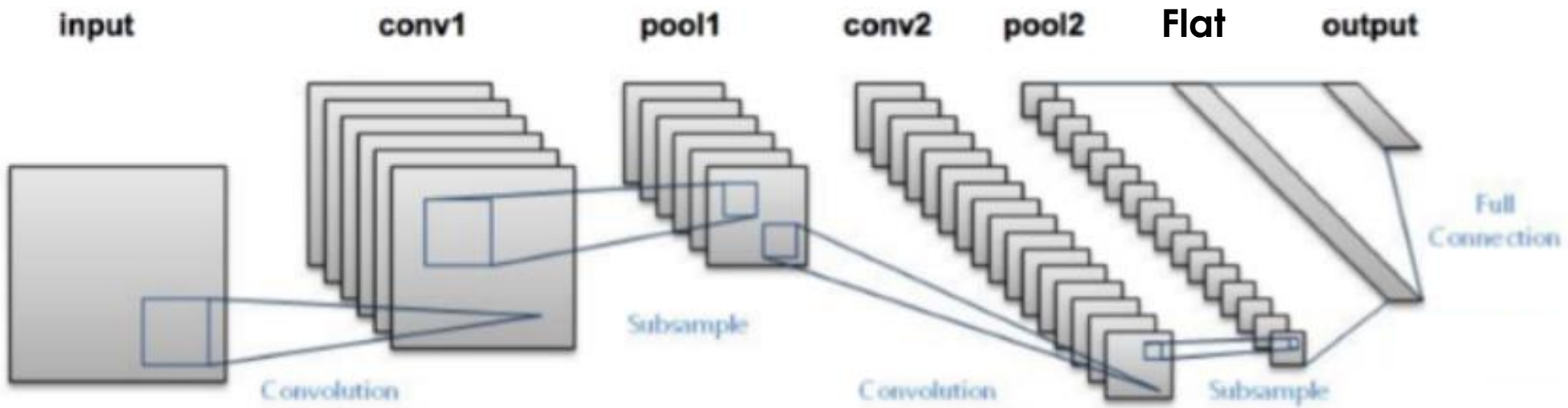
Dynamic Convolutional Neural Network（DCNN）

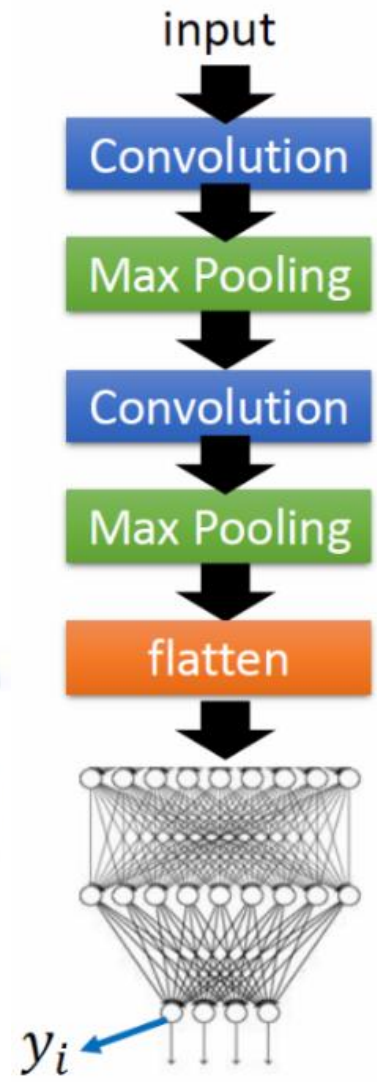


Representation of the sentence  For classification or generation

INTRODUCTION

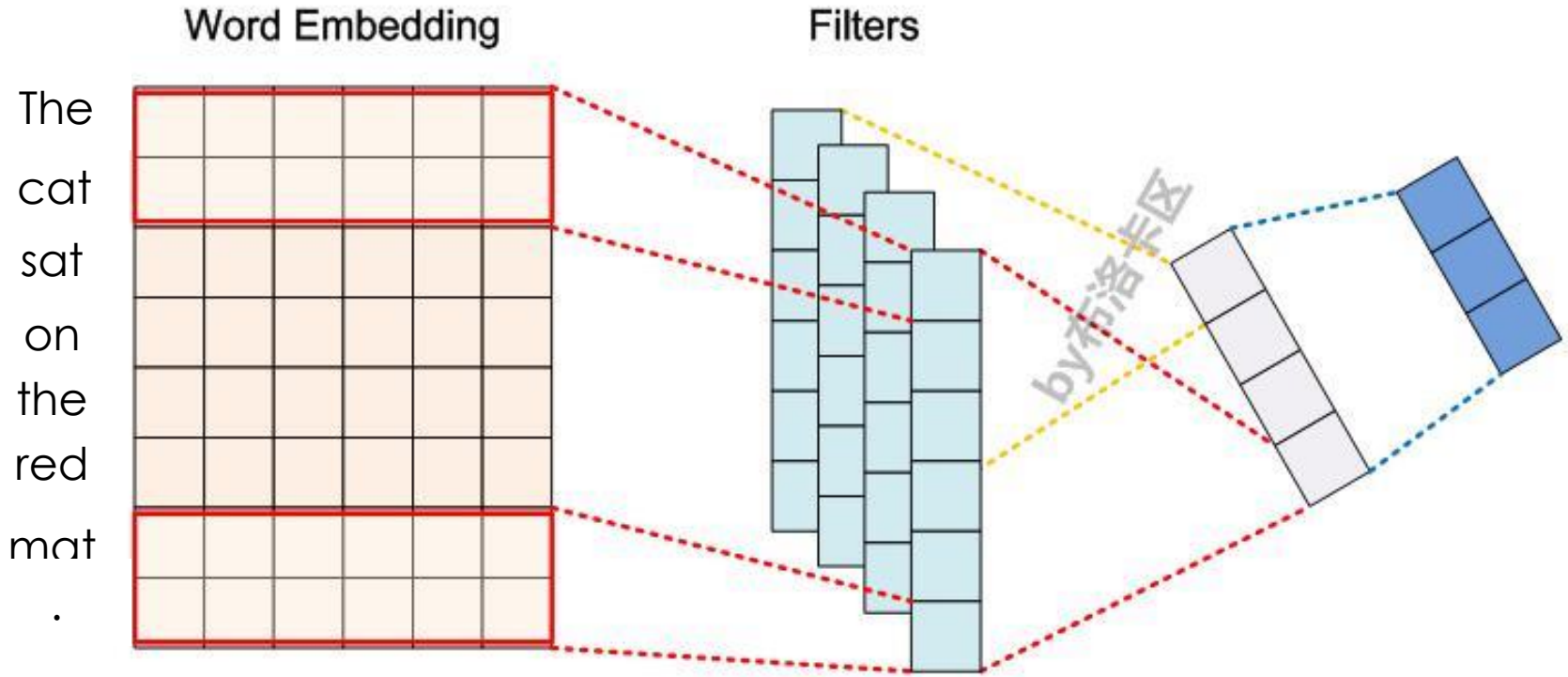
Traditional CNN model - Architecture





INTRODUCTION

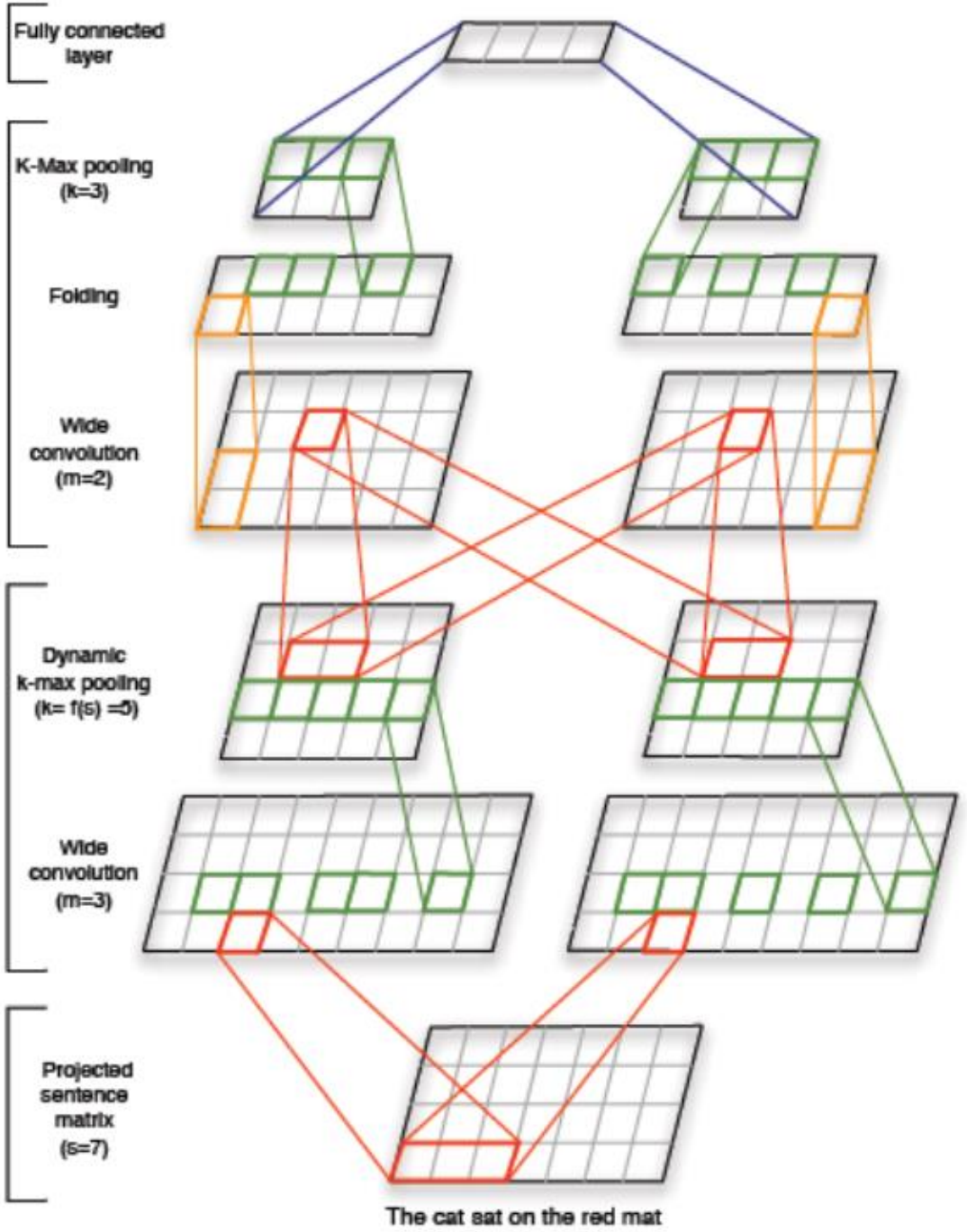
Traditional CNN model – for NLP



OUTLINE

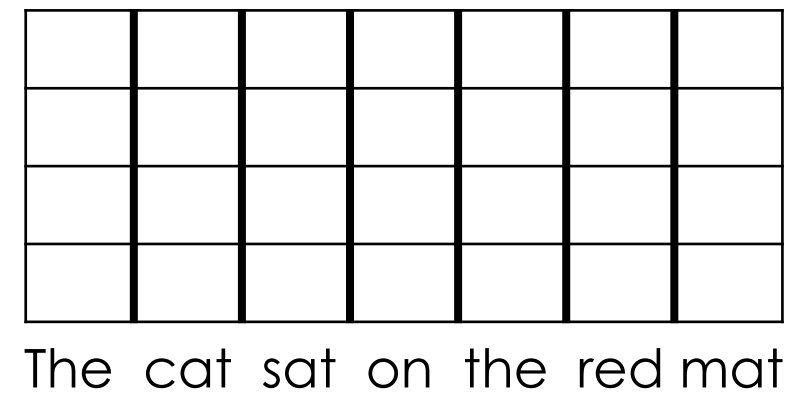
- Introduction
- Method
- Experiment
- Conclusion

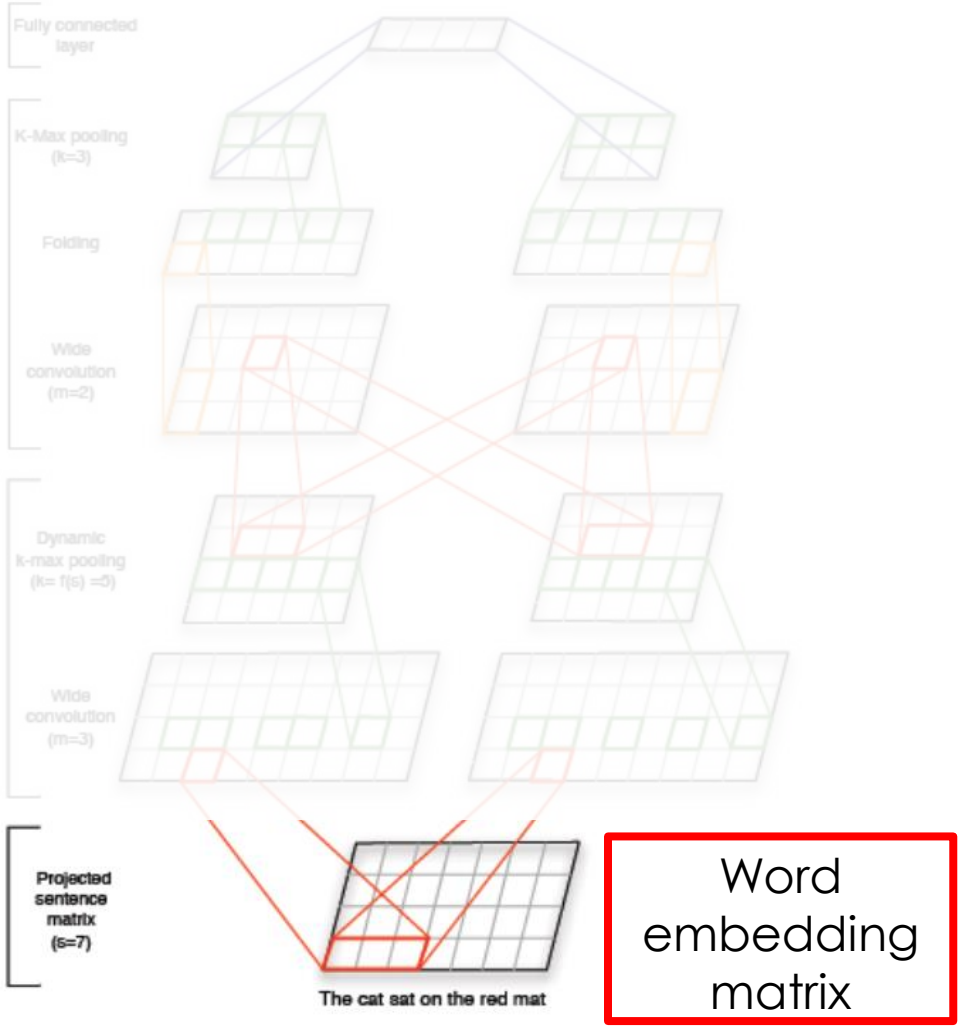
FRAMEWORK


Fully connected layer
K-Max pooling (k=3)
Folding
Wide convolution (m=2)
Dynamic k-max pooling (k= f(s) =5)
Wide convolution (m=3)
Projected sentence matrix (s=7)
The cat sat on the red mat

METHOD Word embedding matrix

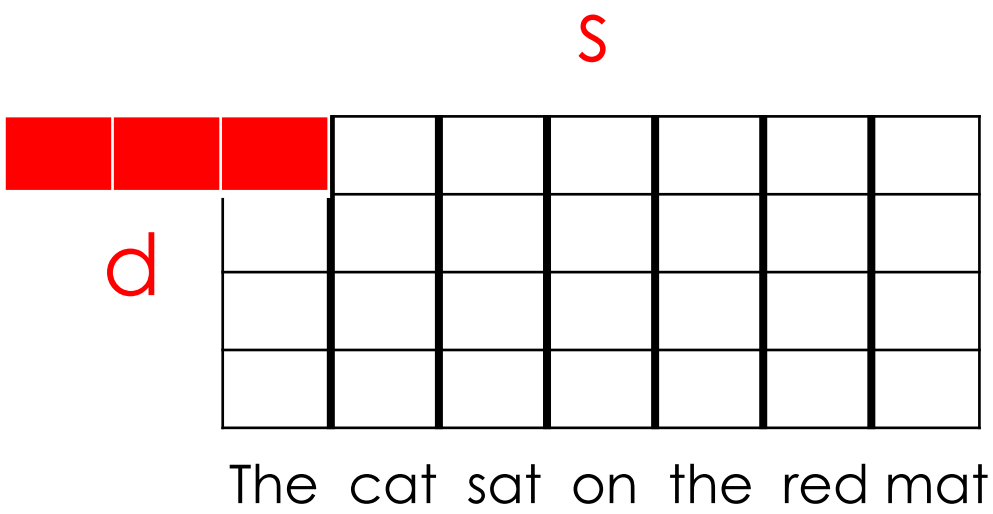
- Take the embedding $w^i \in \mathbb{R}^d$ for each word in the sentence and construct the sentence matrix $s \in \mathbb{R}^{d*s}$
- Sentence matrix :

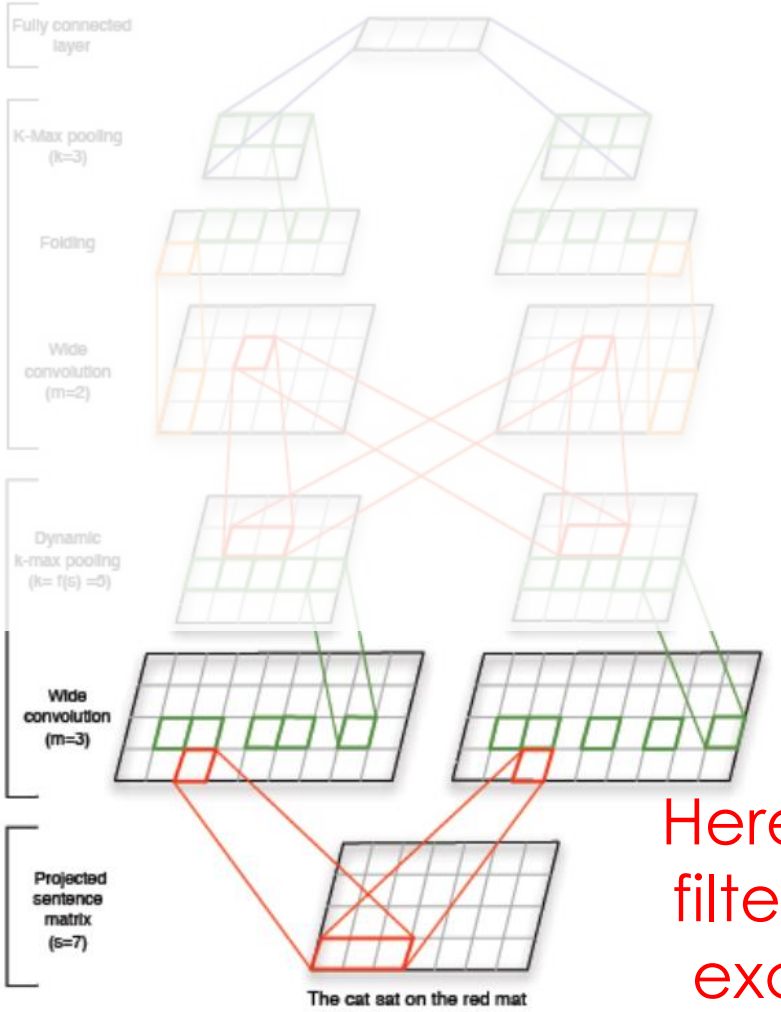




METHOD Wide convolution

- A convolutional layer is obtained by convolving a matrix m with the matrix at the layer below.
- (Ex) : If filter size m = 3.





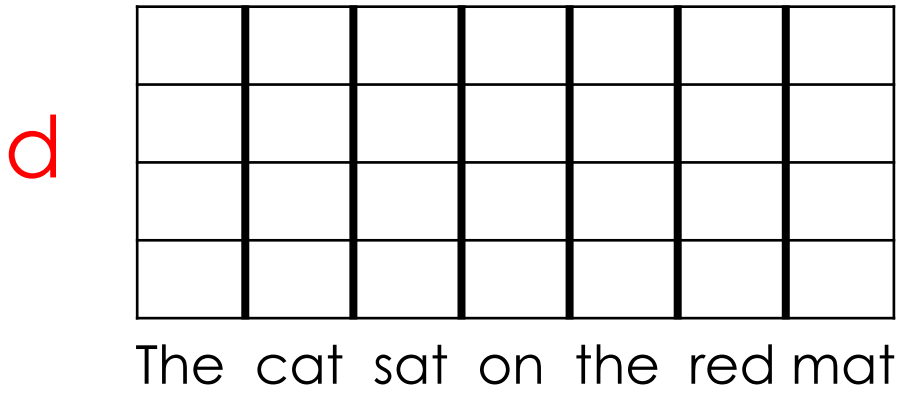
*** Here is two filter in this example

METHOD Wide convolution

- (Ex) : if filter size m = 3.

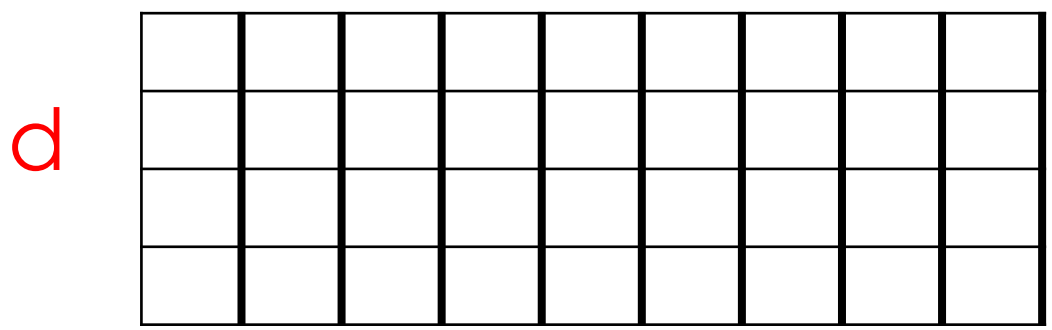
The result of convolution : d * (s+m-1)

Sentence matrix :



Wide Convolution

Result matrix of wide convolution :



METHOD K-Max pooling

- K-max pooling selects the subsequence $\mathbf{p}_{max}^{k}$ of the k highest values of p.
- The order of the values in $\mathbf{p}_{max}^{k}$ corresponds to their original order in p.
- (Ex) : If k = 5

S + m -1

d

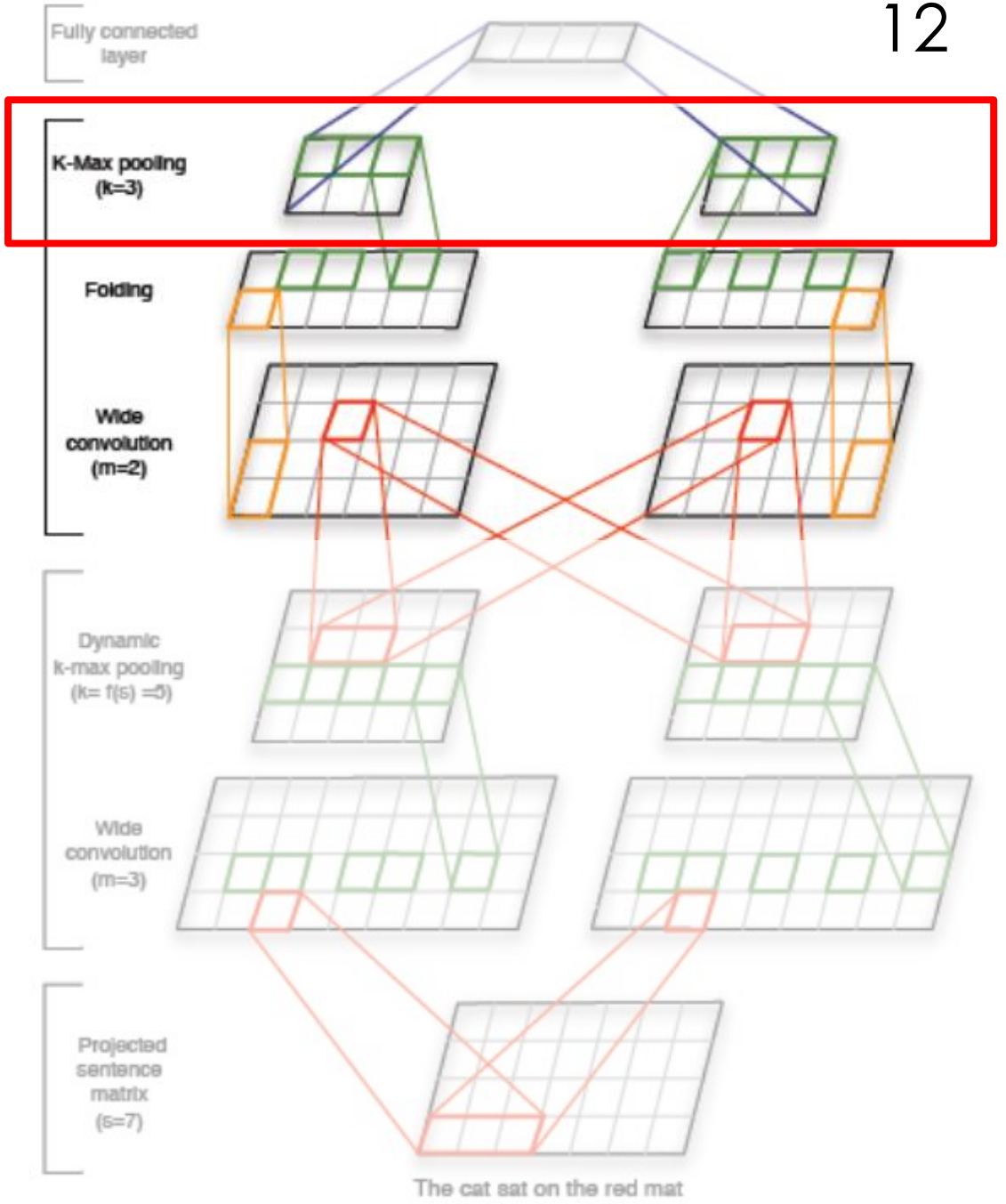
4	3	5	8	2	7	6	1	1
3	7	8	1	5	2	9	6	4
1	6	8	2	5	9	3	4	7
4	3	2	5	6	7	8	9	1

5-Max pooling

4	5	8	7	6
7	8	5	9	6
6	8	5	9	7
5	6	7	8	9

METHOD K-Max pooling

- The k-max pooling operator is only applied in the network after the topmost convolutional layer.

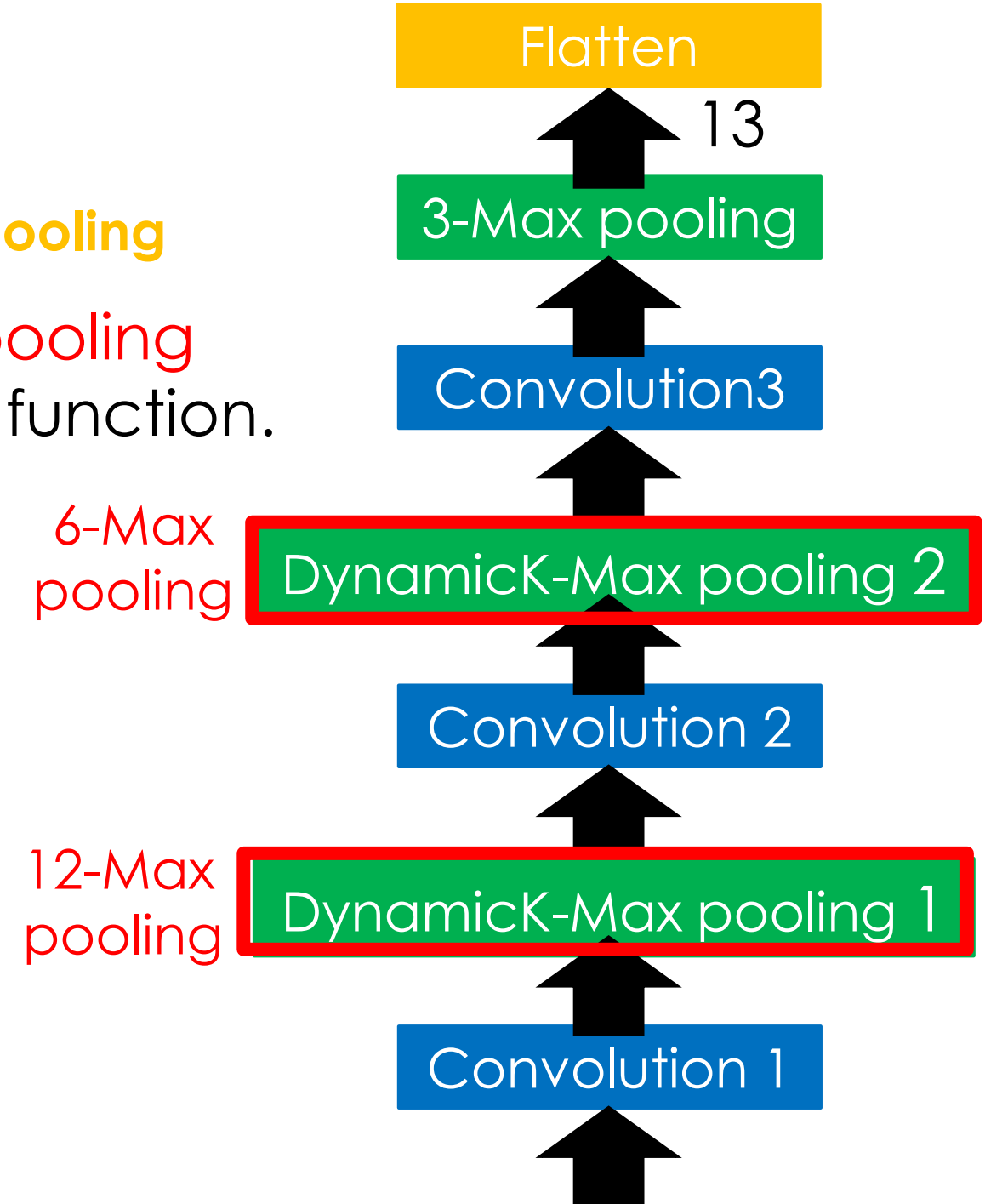


METHOD Dynamic k-Max pooling

- Dynamic k-max pooling is a k-max pooling where k can dynamic change by a function.

- $k_l = \max(k_{top}, \lceil \frac{L-l}{L} s \rceil)$

- (Ex) : If the length of input s = 18
  - k1= max(3, $\frac{3-1}{3} * 18$) = 12
  - k2= max(3, $\frac{3-2}{3} * 18$) = 6



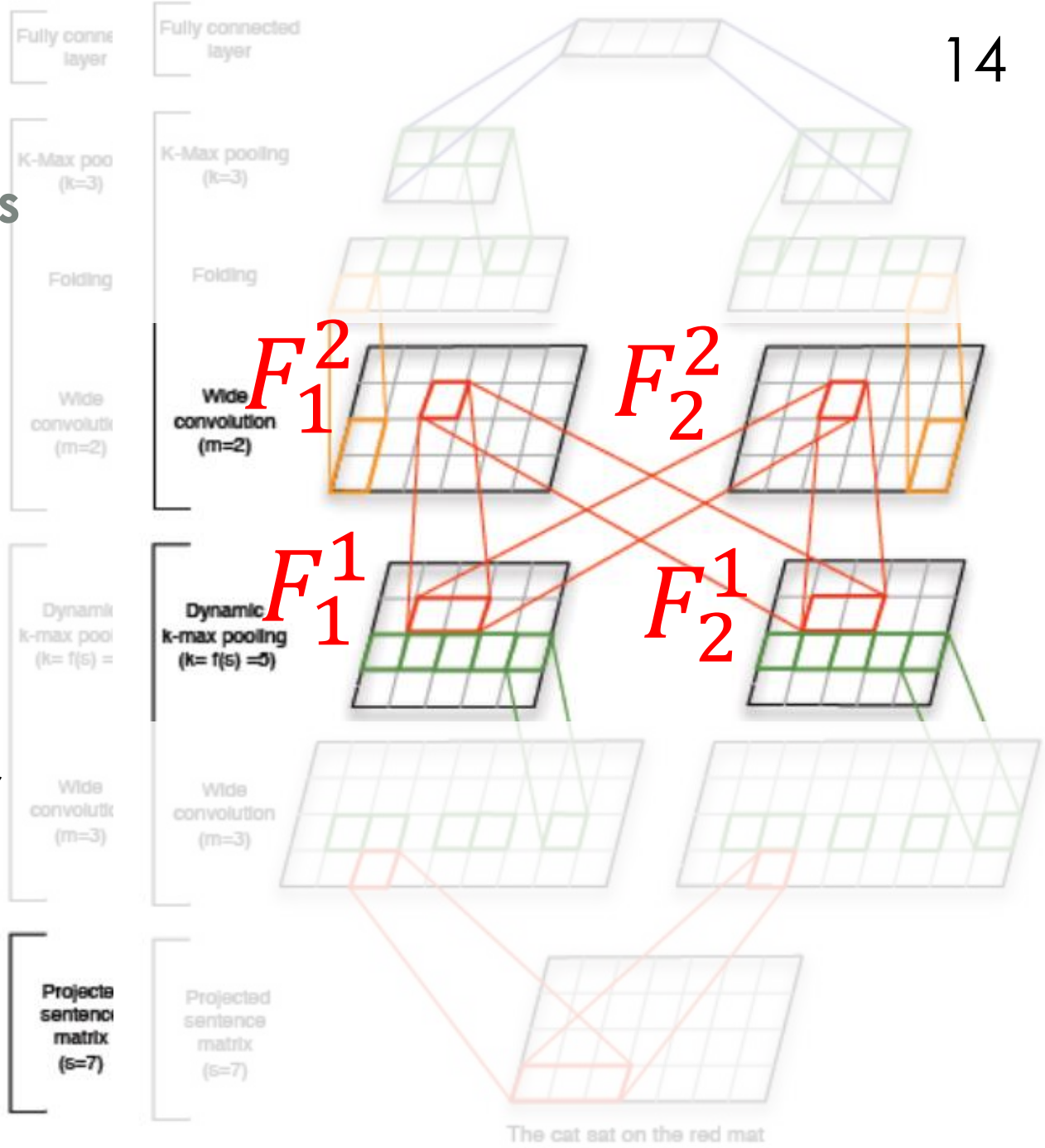
METHOD Multiple feature maps

$$\mathbf{F}_j^i = \sum_{k=1}^{n} \mathbf{m}_{j,k}^i * \mathbf{F}_k^{i-1}$$

- Weight

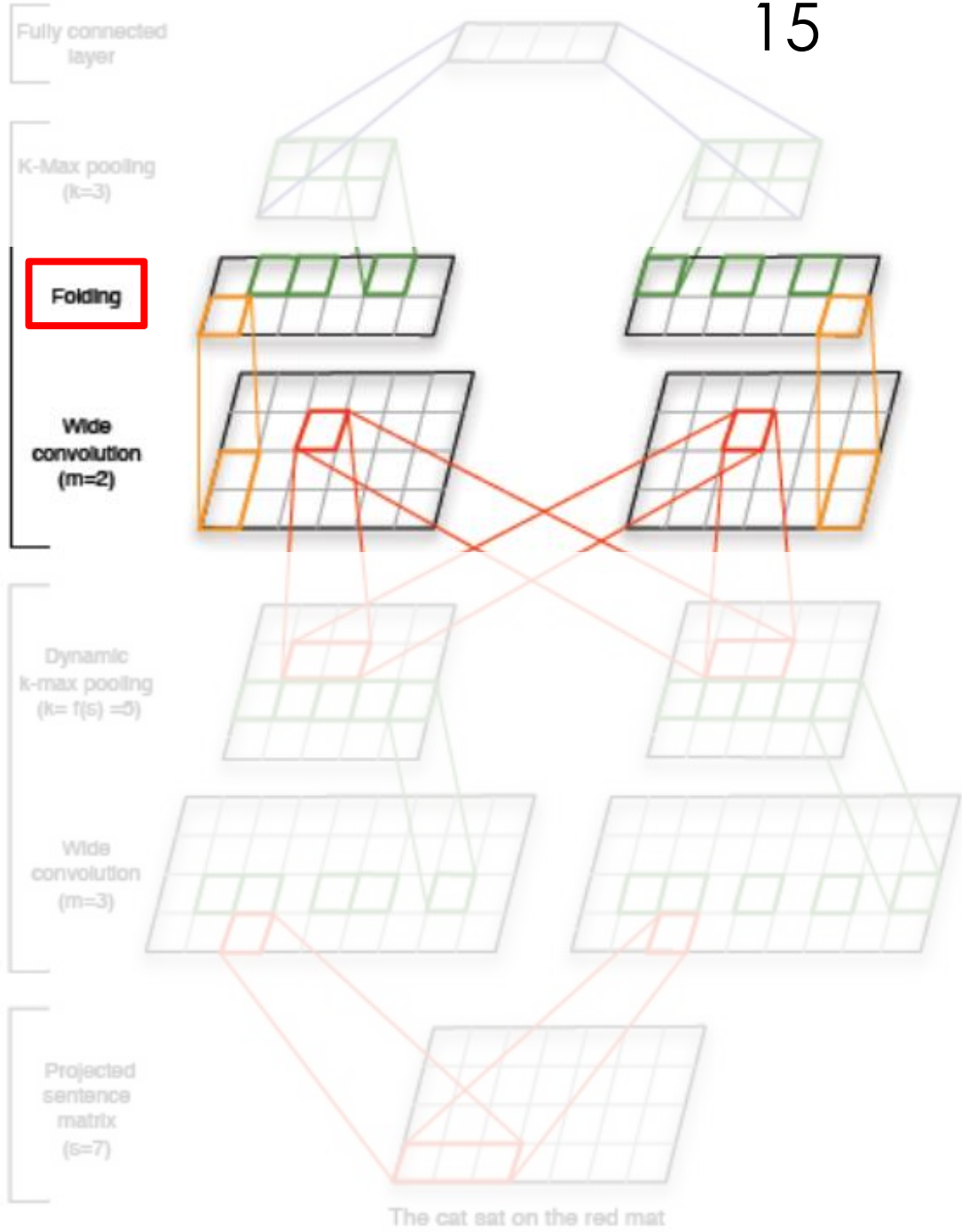
- i-th layer
- j-th feature map

- (i-1)-th layer
- k-th feature map



METHOD **Folding**

- Convolution operation acts independently on each row.
- Before full connected layer, different rows are independent of each other.
- Sums every two rows in a feature map component-wise.



OUTLINE

- Introduction
- Method
- Experiment
- Conclusion

EXPERIMENT

Dataset 1

- **Stanford Sentiment Treebank（movie reviews classification）：**

1. Binary：
   - 6920 training, 872 validation and 1821 test sentences.
2. Fine-grained：
   - 8544 training, 1101 validation and 2210 test sentences.

EXPERIMENT

- Accuracy of sentiment prediction in the movie reviews dataset :

	Classifier	Fine-grained (%)	Binary (%)
Traditional	NB	41.0	81.8
	BiNB	41.9	83.1
	SVM	40.7	79.4
Neural	RecNTN	45.7	85.4
	Max-TDNN	37.4	77.1
	NBoW	42.4	80.5
	DCNN	**48.5**	**86.8**

EXPERIMENT

Dataset 2

- **TREC（question type classification）：**
  - Training dataset consists 5452 labelled questions whereas the test dataset consists of 500 questions.

EXPERIMENT

- Accuracy of six-way question classification on the TREC questions dataset :

Classifier	Features	Acc. (%)
HIER	unigram, POS, head chunks NE, semantic relations	91.0
MAXENT	unigram, bigram, trigram POS, chunks, NE, supertags CCG parser, WordNet	92.6
MAXENT	unigram, bigram, trigram POS, wh-word, head word word shape, parser hypernyms, WordNet	93.6
SVM	unigram, POS, wh-word head word, parser hypernyms, WordNet 60 hand-coded rules	95.0
MAX-TDNN	unsupervised vectors	84.4
NBOW	unsupervised vectors	88.2
DCNN	unsupervised vectors	93.0

Traditional

Neural

EXPERIMENT

Dataset 3

- **Twitter Sentiment Prediction：**
  - Use distant supervision.
  - Training set consists of 1.6 million tweets with emoticon-based labels and the test set of about 400 hand-annotated tweets.

EXPERIMENT

- Accuracy on the Twitter sentiment dataset :

	Classifier	Accuracy (%)
Traditional	SVM	81.6
	BINB	82.7
Neural	MAXENT	83.0
	MAX-TDNN	78.8
	NBOW	80.9
	DCNN	**87.4**

EXPERIMENT

- Top five 7-grams feature detectors in the first layer of the network :

			POSITIVE			
lovely	comedic	moments	and	several	fine	performances
good	script	,	good	dialogue	,	funny
sustains	throughout	is	daring	,	inventive	and
well	written	,	nicely	acted	and	beautifully
remarkably	solid	and	subtly	satirical	tour	de
			NEGATIVE			
,	nonexistent	plot	and	pretentious	visual	style
it	fails	the	most	basic	test	as
so	stupid	,	so	ill	conceived	,
,	too	dull	and	pretentious	to	be
hood	rats	butt	their	ugly	heads	in

EXPERIMENT

- Top five 7-grams feature detectors in the first layer of the network :

			'NOT'			
n't	have	any	huge	laughs	in	its
no	movement	,	no	,	not	much
n't	stop	me	from	enjoying	much	of
not	that	kung	pow	is	n't	funny
not	a	moment	that	is	not	false
			'TOO'			
,	too	dull	and	pretentious	to	be
either	too	serious	or	too	lighthearted	,
too	slow	,	too	long	and	too
feels	too	formulaic	and	too	familiar	to
is	too	predictable	and	too	self	consciou

OUTLINE

- Introduction
- Method
- Experiment
- Conclusion

CONCLUSION

- Described a dynamic convolutional neural network that uses the dynamic k-max pooling operator as a non-linear subsampling function.
- The network achieves high performance on question and sentiment classification without requiring external features as provided by parsers or other resources.