Character-Aware Neural Language Models

Paper authors: Y. Kim, Y. Jernite, D. Sontag, A. M. Rush

2018. 10. 20.

5th flipped school, NLP bootcamp

Modulabs Research Scientist

Il Gu Yi

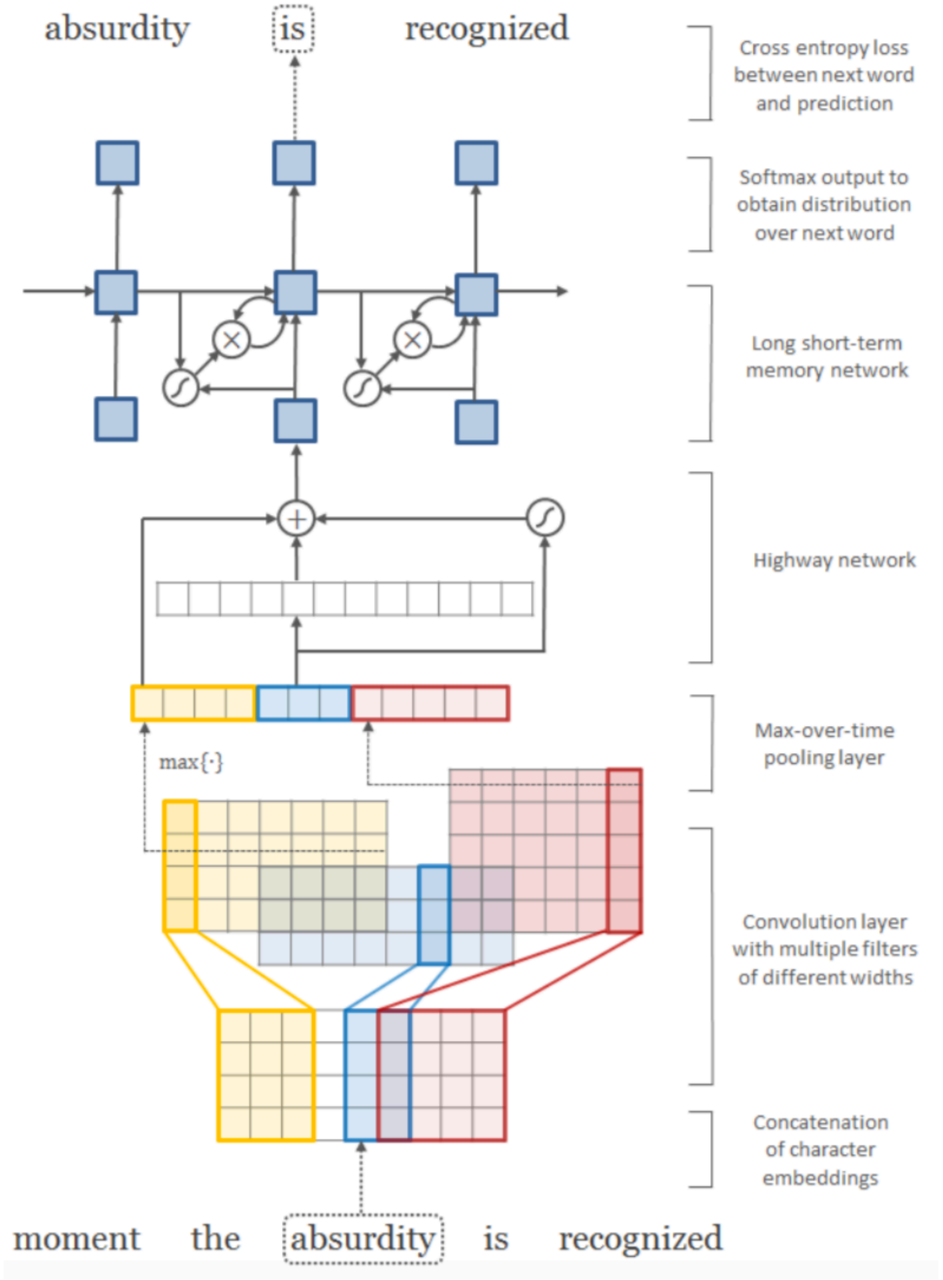
Contents

- Introduction
- Architecture
- Results

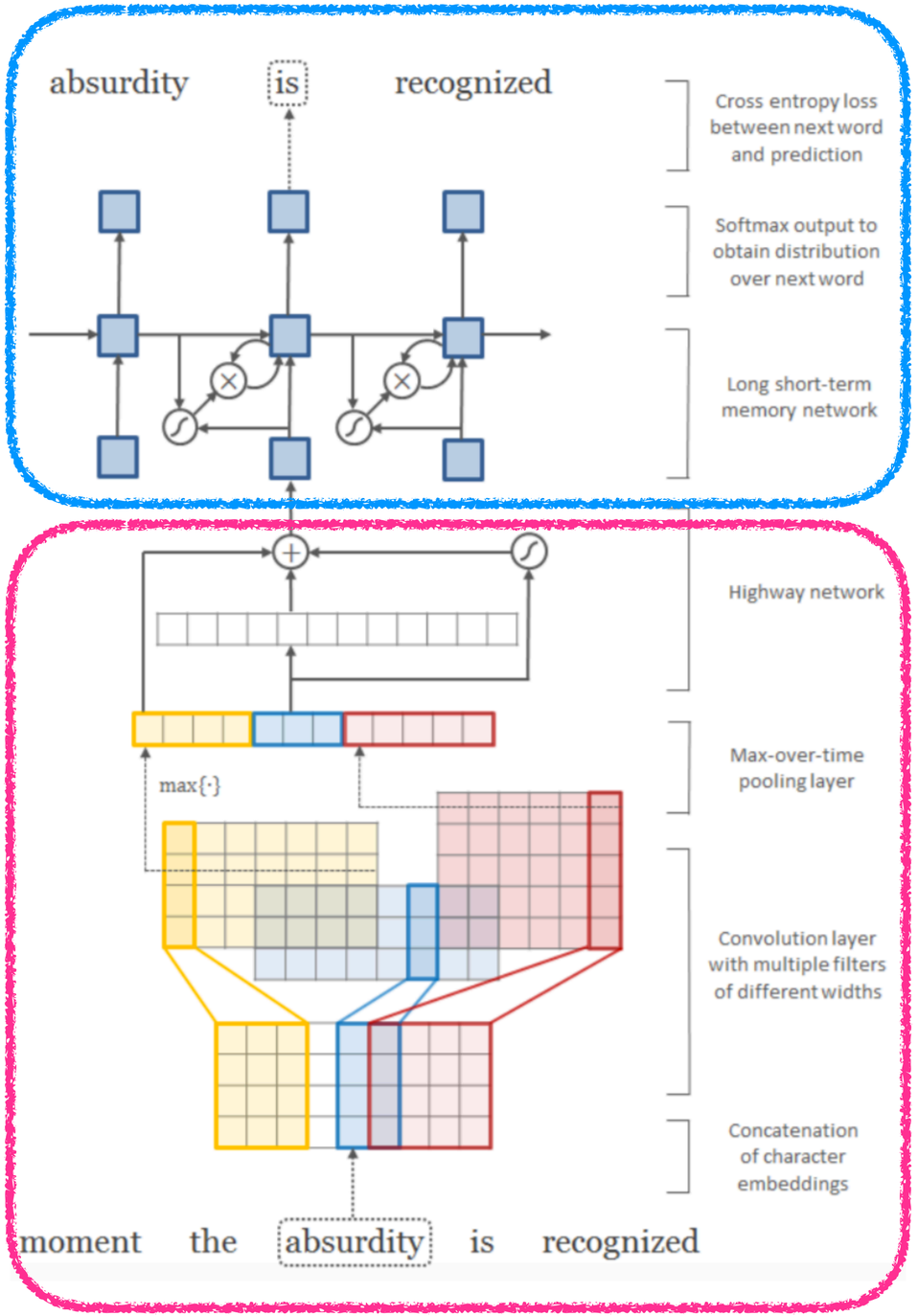
Introduction

- Neural Language Models (NLM)
  - Words as vectors (word embeddings)
- In this work (한 줄 요약)
  - RNN input 으로 들어가는 word embedding vector를 character-level CNN을 통해 만들었다

Architecture



Architecture



RNN

Word embedding by Character-aware CNN

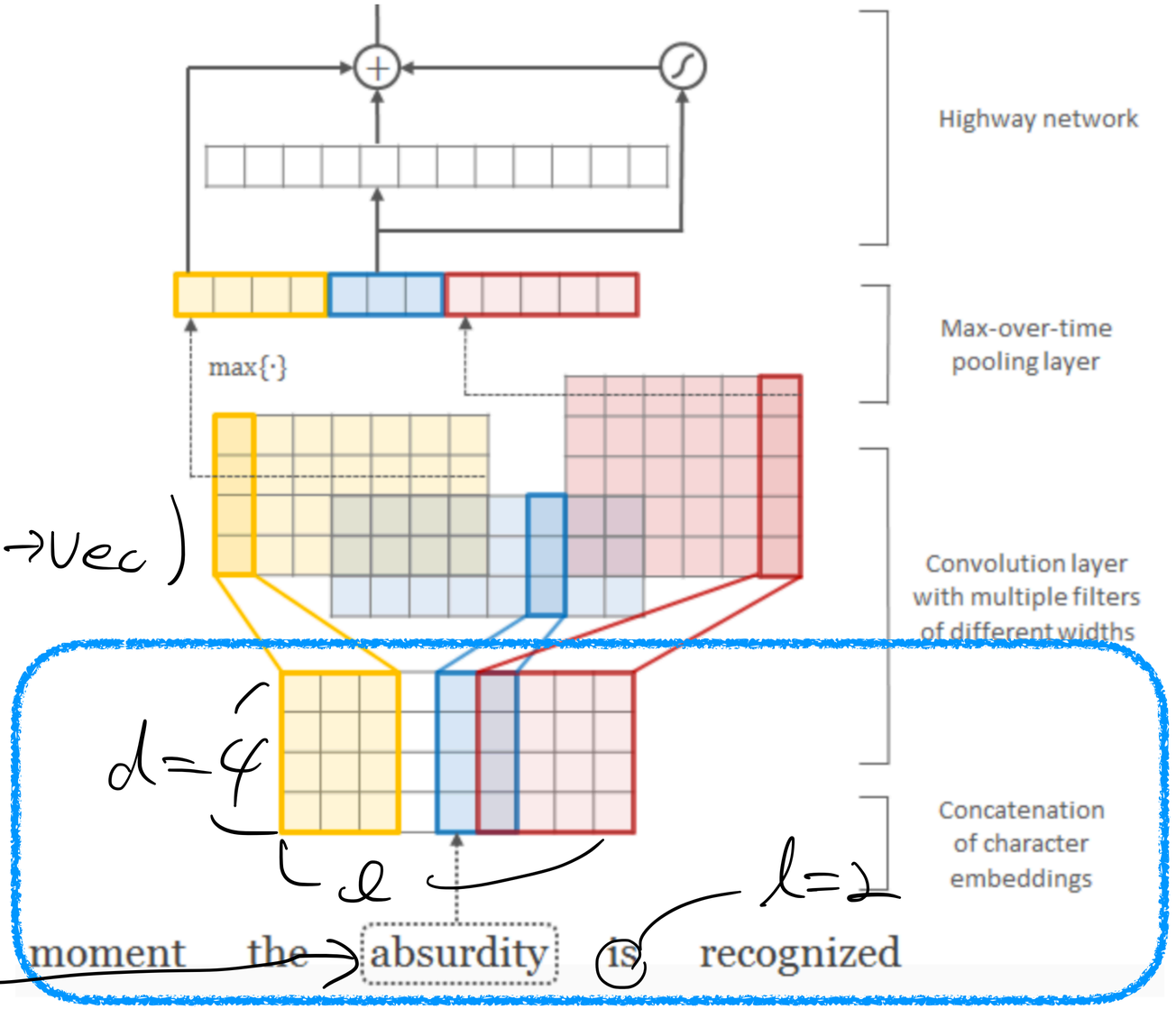
1. Character Embedding

$C^k_{d \times l}$: character embedding

k : word index

d : embedding dim. (char → vec)

l : length of word (depending on word)



2. Convolution Layer

Use Conv1D

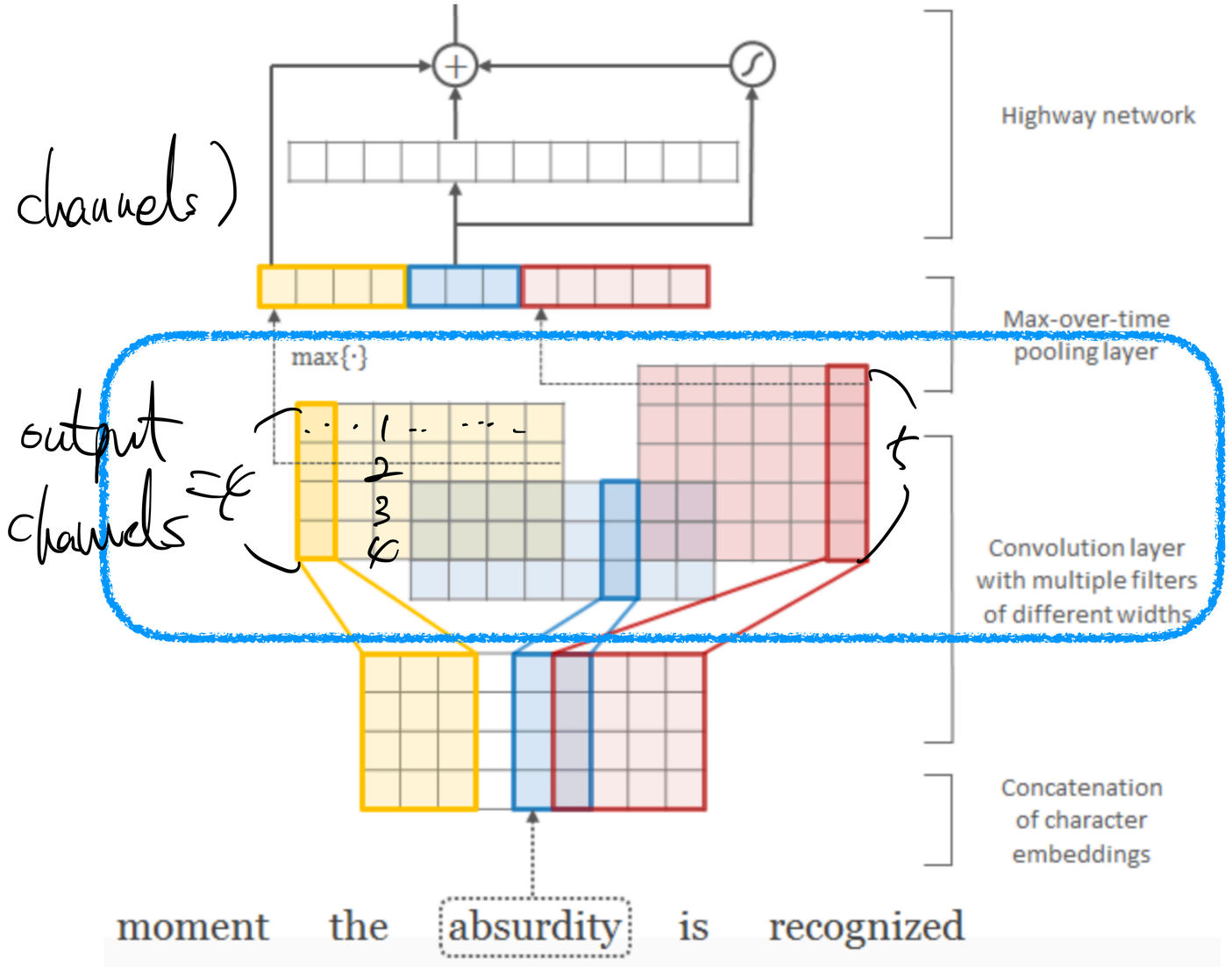
(kernel size, output channels)

blue : (2, 3)

yellow : (3, 4)

red : (4, 5)

(like inception)

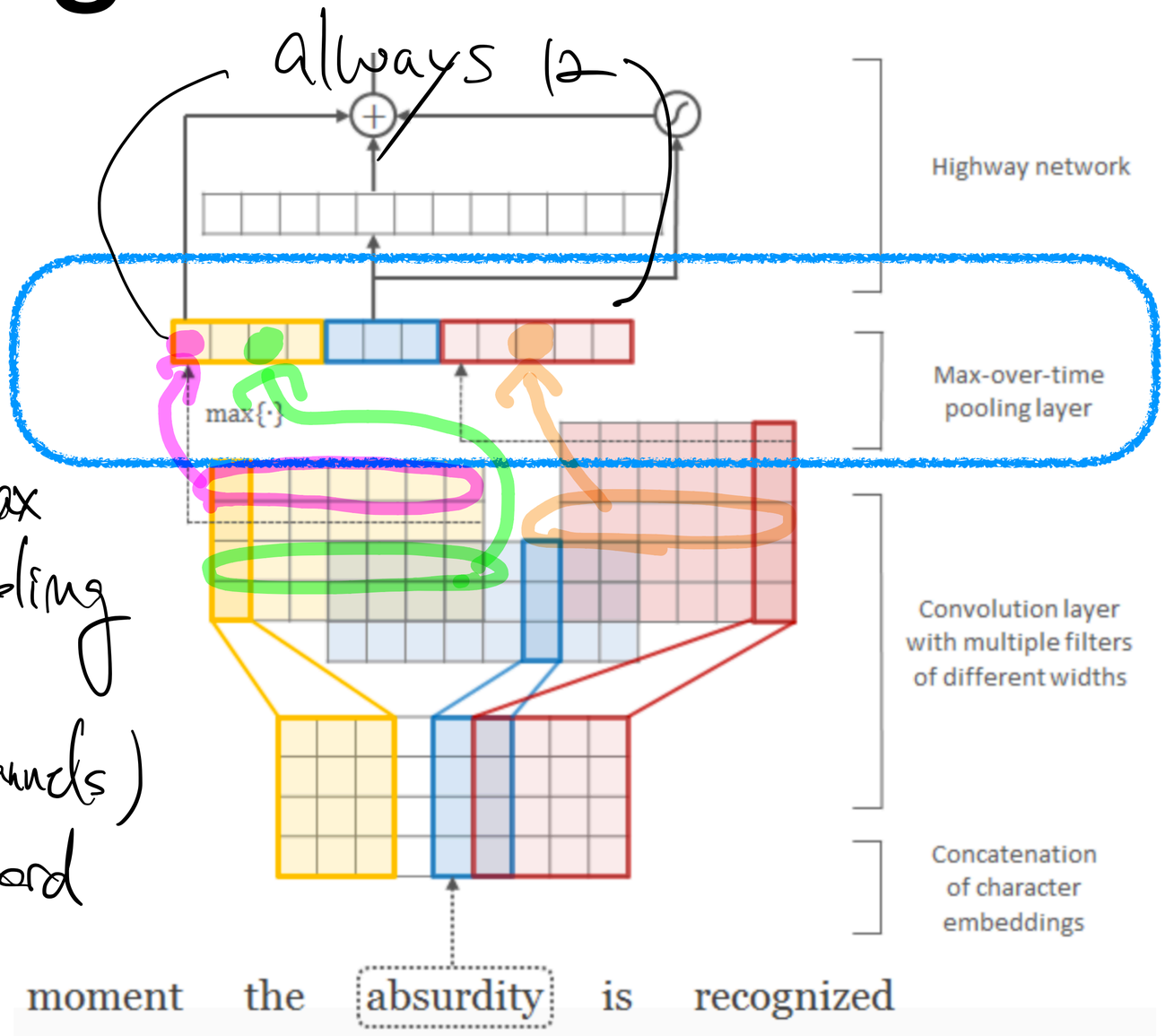


3. Max Pooling over Time

Max Pooling
in each output channel

length of output vec. = # output channels
is always 12 regardless of length of input word

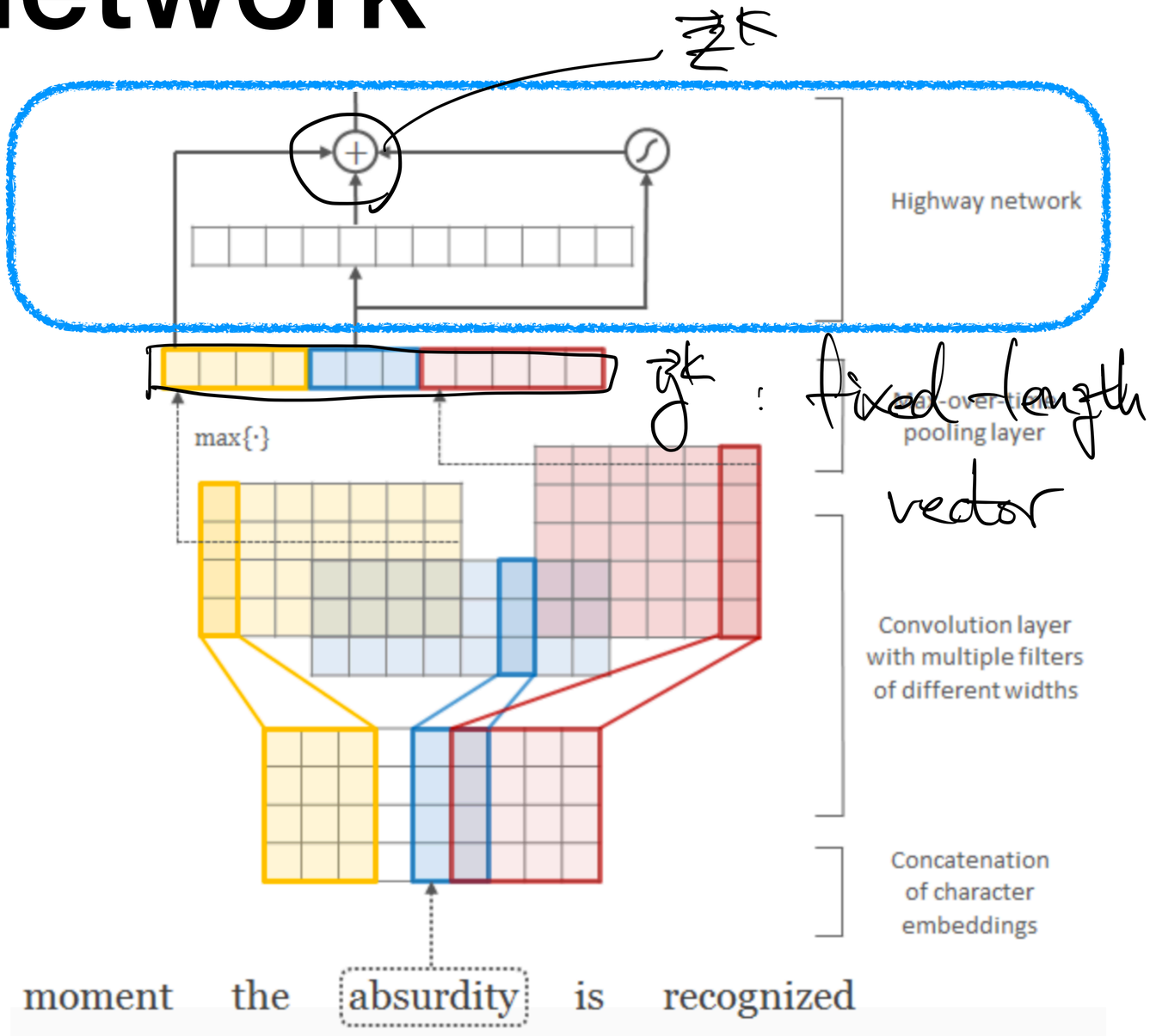
max pooling



4. Highway Network

$$\vec{z}^k = \vec{t}^k \odot g(\vec{W}_H \vec{y}^k + \vec{b}_H) + (1 - \vec{t}^k) \odot \vec{y}^k$$

$$\vec{t}^k = \sigma(\vec{W}_T \vec{y}^k + \vec{b}_T)$$



Optimization

- Truncated backpropagation through time
- Dropout rate: 0.5
- Norm of gradients to be below 5
- Use a hierarchical softmax $$\Pr(w_{t+1} = j | w_{1:t}) = \frac{\exp(\mathbf{h}_t \cdot \mathbf{s}^r + t^r)}{\sum_{r'=1}^{c} \exp(\mathbf{h}_t \cdot \mathbf{s}^{r'} + t^{r'})} \times \frac{\exp(\mathbf{h}_t \cdot \mathbf{p}_r^j + q_r^j)}{\sum_{j' \in \mathcal{V}_r} \exp(\mathbf{h}_t \cdot \mathbf{p}_r^{j'} + q_r^{j'})}$$

first term: Pr of picking cluster r
second term: Pr of picking word j
given that cluster r is picked

Datasets

	DATA-S			DATA-L		
	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{C}\rvert$	T	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{C}\rvert$	T
English (EN)	10 k	51	1 m	60 k	197	20 m
Czech (CS)	46 k	101	1 m	206 k	195	17 m
German (DE)	37 k	74	1 m	339 k	260	51 m
Spanish (ES)	27 k	72	1 m	152 k	222	56 m
French (FR)	25 k	76	1 m	137 k	225	57 m
Russian (RU)	62 k	62	1 m	497 k	111	25 m
Arabic (AR)	86 k	132	4 m	–	–	–

Results

	PPL	Size
LSTM-Word-Small	97.6	5 m
LSTM-Char-Small	92.3	5 m
LSTM-Word-Large	85.4	20 m
LSTM-Char-Large	78.9	19 m
KN-5 (Mikolov et al. 2012)	141.2	2 m
RNN† (Mikolov et al. 2012)	124.7	6 m
RNN-LDA† (Mikolov et al. 2012)	113.7	7 m
genCNN† (Wang et al. 2015)	116.4	8 m
FOFE-FNNLM† (Zhang et al. 2015)	108.0	6 m
Deep RNN (Pascanu et al. 2013)	107.5	6 m
Sum-Prod Net† (Cheng et al. 2014)	100.0	5 m
LSTM-1† (Zaremba et al. 2014)	82.7	20 m
LSTM-2† (Zaremba et al. 2014)	78.4	52 m

Thank you for your attention!!