Hierarchical Attention Networks for Document Classification

Zichao Yang el al.

이종진

Seoul National University

ga0408@snu.ac.kr

July 06, 2018

Two characteristics

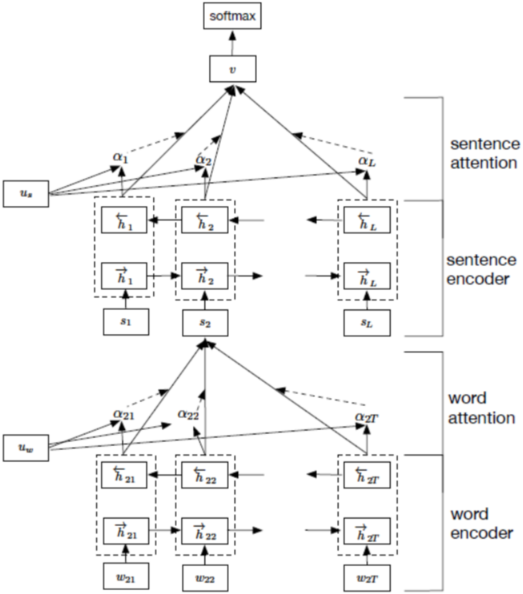
- Hierarchical document structure

  → Word-level GRU and sentence-level GRU

- Informative words and sentences are different in a document

  → Two attention mechanism.

Model


softmax
v
sentence attention
α_1
α_2
α_L
u_s
$\overleftarrow{h}_1$
$\overleftarrow{h}_2$
$\overleftarrow{h}_L$
$\overrightarrow{h}_1$
$\overrightarrow{h}_2$
$\overrightarrow{h}_L$
sentence encoder
s_1
s_2
s_L
word attention
α_{21}
α_{22}
α_{2T}
u_w
$\overleftarrow{h}_{21}$
$\overleftarrow{h}_{22}$
$\overleftarrow{h}_{2T}$
$\overrightarrow{h}_{21}$
$\overrightarrow{h}_{22}$
$\overrightarrow{h}_{2T}$
word encoder
w_{21}
w_{22}
w_{2T}

Word-level

- w_{ijt} : word at time t in jth sentence of ith document
- Word encoder

$$\begin{aligned} x_{ijt} &= W_e w_{ijt} \\ h_{ijt} &= [G\vec{R}U(x_{ijt}), G\overleftarrow{R}U(x_{ijt})] \end{aligned} \tag{1}$$

- Word Attention

$$\begin{aligned} u_{ijt} &= tanh(W_w h_{ijt} + b_w) \\ \alpha_{ijt} &= \frac{exp(u_{ijt}^T w_w)}{\sum exp(u_{ijt}^T w_w)} \\ s_{ij} &= \sum \alpha_{ijt} h_{ijt} \end{aligned} \tag{2}$$

- Sentence encoder

$$h_{ij} = [G\vec{R}U(s_{ij}), G\overleftarrow{R}U(s_{ij})] \tag{3}$$

- Sentence Attention

$$\begin{aligned} u_{ij} &= tanh(W_w h_{ij} + b_w) \\ \alpha_{ijt} &= \frac{exp(u_{ij}^T w_w)}{\sum exp(u_{ij}^T w_w)} \\ d_i &= \sum \alpha_{ij} h_{ij} \end{aligned} \tag{4}$$

- Classification

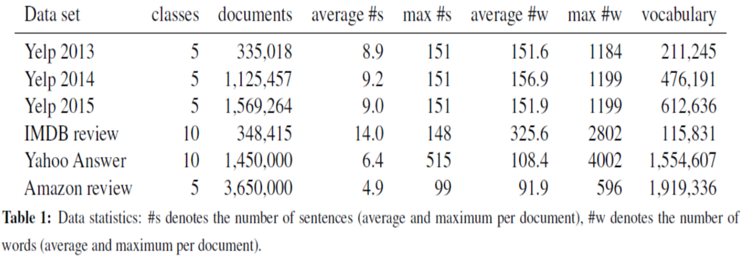
$$\hat{p}_i = softmax(W_c d_i + b_c) \tag{5}$$

- Loss : Negative log likelihood

$$L = -\sum p_i \log \hat{p}_i = softmax(W_c d_i + b_c) \tag{6}$$

Experiments

- Yelp reviews / IMDB reviews / Yahoo answers / Amazon reviews

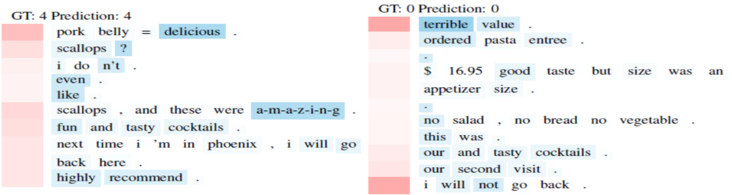


Data set	classes	documents	average #s	max #s	average #w	max #w	vocabulary
Yelp 2013	5	335,018	8.9	151	151.6	1184	211,245
Yelp 2014	5	1,125,457	9.2	151	156.9	1199	476,191
Yelp 2015	5	1,569,264	9.0	151	151.9	1199	612,636
IMDB review	10	348,415	14.0	148	325.6	2802	115,831
Yahoo Answer	10	1,450,000	6.4	515	108.4	4002	1,554,607
Amazon review	5	3,650,000	4.9	99	91.9	596	1,919,336

Table 1: Data statistics: #s denotes the number of sentences (average and maximum per document), #w denotes the number of words (average and maximum per document).

Experiments

- Visualization of attention



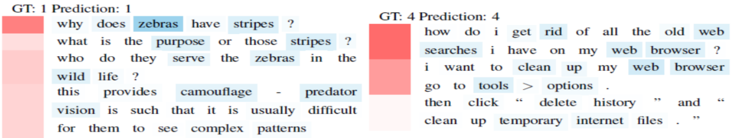
GT: 4 Prediction: 4

pork belly = delicious .
scallops ?
i do n't .
even .
like .
scallops , and these were a-m-a-z-i-n-g .
fun and tasty cocktails .
next time i 'm in phoenix , i will go
back here .
highly recommend .

GT: 0 Prediction: 0

terrible value .
ordered pasta entree .
.
$ 16.95 good taste but size was an
appetizer size .
.
no salad , no bread no vegetable .
this was .
our and tasty cocktails .
our second visit .
i will not go back .

Figure 5: Documents from Yelp 2013. Label 4 means star 5, label 0 means star 1.



GT: 1 Prediction: 1

why does zebras have stripes ?
what is the purpose or those stripes ?
who do they serve the zebras in the
wild life ?
this provides camouflage - predator
vision is such that it is usually difficult
for them to see complex patterns

GT: 4 Prediction: 4

how do i get rid of all the old web
searches i have on my web browser ?
i want to clean up my web browser
go to tools > options .
then click " delete history " and "
clean up temporary internet files . "

Figure 6: Documents from Yahoo Answers. Label 1 denotes Science and Mathematics and label 4 denotes Computers and Internet.