A self-attentive sentence embedding

MAP583 - Deep Learning

Jean-Charles Layoun, Inès Multrier, Tom Sander

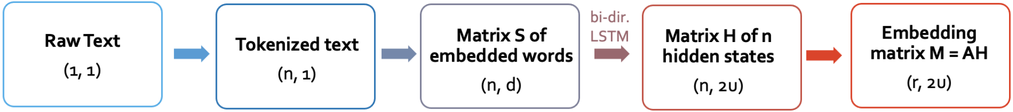
Ecole polytechnique

March 9, 2021

Overview

1. Main ideas of the paper
2. Annotation matrix
3. Tiny learning improvements
4. Our results on both datasets

How is a sentence embedded?



where A is the annotation matrix, a two-layer perceptron defined as follows:

$$A = softmax(W_{s2} tanh(W_{s1} H^T))$$

W_{s1} and W_{s2} are the parameters, of size d_a-by-$2u$ and r-by-d_a respectively.

A new approach

- New: different aspects of the sentence into multiple vector-representations
- It relieves the burden of LSTM to carry on long term dependencies, because the last hidden state access previous steps with the annotation matrix
- The use of multiple hops of attention enables the model to catch the global semantic of the sentence
- Interpretable model

Penalization

- Problem : each vector of weights gives importance to many words
- KL Divergence between vector of weights seen as probability distributions brings instability in the training
- Penalization of the Frobenius norm of $AA^t - I$ erases redundancy and forces each vector of weights to focus on a specific part of the sentence

Annotation Matrix

- We used some of the code from this [Github repository] to build a pipeline that automatically creates heatmaps from the batch with best accuracy global (86%)
- Summing up over all the annotation vectors, normalizing the result gives general overview.

Annotation Matrix

One Annotation matrix for an accurate high prediction

label : 4, prediction 4

what a great place ! love the food and the atmosphere ! i was in last friday and although it was packed due to phoenix comicon , a dbacks game , the phoeinix symphony and first friday ... we got a table in the bar , had attentive service and made it out just in time to make it to the phoenix symphony for our show ! my boyfriend and i went in yesterday to enjoy some drinks and appetizers for the afternoon and had a wonderful time again ! lauren was our server and she was fabulous ! again attentive and made sure we had everything we needed ! i will definitely be back !

Annotation Matrix

One Annotation matrix for an accurate high prediction

label : 4, prediction 4

we love the thumb , come to get our car washed and have breakfast or lunch or dinner , just depending on the time of the day . the brisket sandwich is the best there is (or anything with brisket ! !) . the fries are just the right amount of crisp . staff is great , especially the car wash staff who do an awesome and consistent job . looking forward to enjoying the newly remodeled patio this fall !

Annotation Matrix

One Annotation matrix for very wrong prediction

label : 0, prediction 4

i have to say that i was just appalled by the attendants in the show . this was a very fun loud amazing show and everyone was dancing and clapping and singing along to the show . why my self , partner , and best friend was pin pointed out as being to loud is beyond me . it 's sad to say i will never go back even though the show was amazing . i felt discriminated and humiliated that they can dismiss us from a show for being too loud . i am a 45 year old woman and my partner is 55 and you are really going to kick us out like we were little kids that would n't behave .. i am not al all a negative person but this was way out of line and uncalled for .

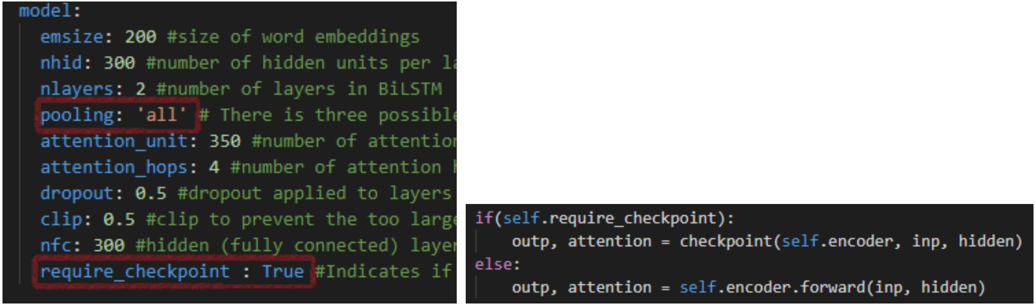
Annotation Matrix

Most of the time, the matrix is not interpretable

label : 1, prediction 0

new in town & this place was recommended to my husband and i by many people . we are both from ca , so we know good mexican food . we have been living in italy for the past few years , where mexican food is non - existent , so we were looking forward to eating here . i got the 1 taco & 1 enchilada plate & my husband got the 1 burrito & 1 enchilada plate . while eating the beans i was thinking to myself how salty they were , & then moved on to the taco which was even more salty . it tasted like i was eating pure salt . i mentioned it to my husband , & he said the same thing . needless to say , we did not finish our food . will not be eating here again .

Architecture Modifications



```
model:
  emsize: 200 #size of word embeddings
  nhid: 300 #number of hidden units per la
  nlayers: 2 #number of layers in BiLSTM
  pooling: 'all' # There is three possible
  attention_unit: 350 #number of attention
  attention_hops: 4 #number of attention
  dropout: 0.5 #dropout applied to layers
  clip: 0.5 #clip to prevent the too large
  nfc: 300 #hidden (fully connected) layer
  require_checkpoint : True #Indicates if
```

(a) Config File

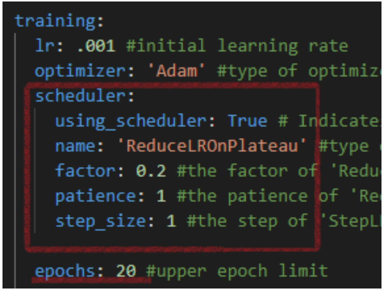
```
if(self.require_checkpoint):
    outp, attention = checkpoint(self.encoder, inp, hidden)
else:
    outp, attention = self.encoder.forward(inp, hidden)
```

(b) Checkpoint

Figure: Model modifications at each level.

- Adding a checkpoint lets us run our model on the big dataset.

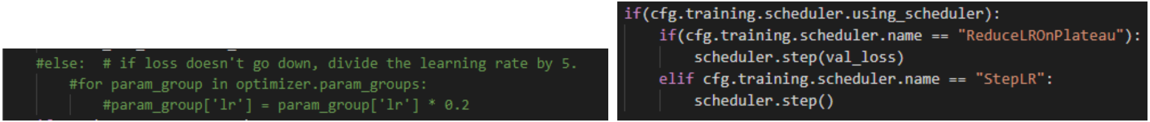
Training Modifications



```
training:
  lr: .001 #initial learning rate
  optimizer: 'Adam' #type of optimiz
  scheduler:
    using_scheduler: True # Indicate
    name: 'ReduceLROnPlateau' #type
    factor: 0.2 #the factor of 'Redu
    patience: 1 #the patience of 'Re
    step_size: 1 #the step of 'StepL

  epochs: 20 #upper epoch limit
```

Figure: Training modifications for configuration file.



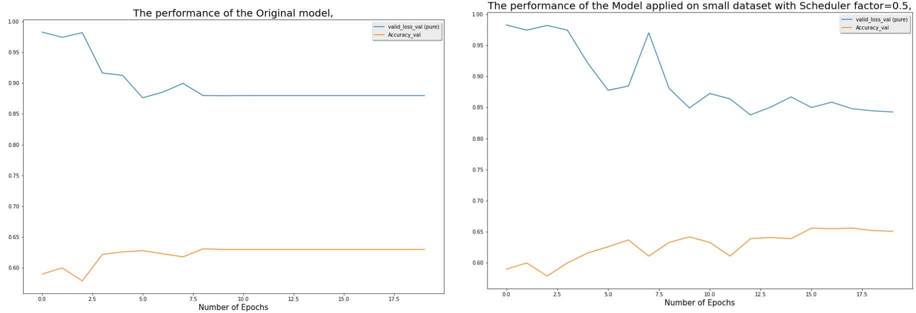
```
    #else:  # if loss doesn't go down, divide the learning rate by 5.
        #for param_group in optimizer.param_groups:
            #param_group['lr'] = param_group['lr'] * 0.2
```

```
if(cfg.training.scheduler.using_scheduler):
    if(cfg.training.scheduler.name == "ReduceLROnPlateau"):
        scheduler.step(val_loss)
    elif cfg.training.scheduler.name == "StepLR":
        scheduler.step()
```

(a) Old Scheduler

(b) New Implementation

Learning curve comparison



(a) Original Implementation

(b) New Implementation

Figure: Comparison of learning curves.

- We can see that for the original implementation it is useless to train for more than 5 epochs. Indeed, after only three epochs the learning rate is already at 2.10^{-4}.

Results on Small Data

State	Max Accuracy (%)	Min Loss
Before Modifs	63.1%	0.8761
After Modifs	65.6%	0.8380

Table: Performances of our best models before and after our modifications.

Absence of Results on Large Dataset

Difficulties of working with a lot of data:

- Tokenizing and splitting the data took us way more time than what we expected.
- We were limited by our hardware: training our model on this Dataset was not possible on our graphic cards. $\Rightarrow$ use of checkpoint.
- In order to make use of this big dataset, we should increase the number of parameters of our model.

Thank you!

References

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou and Yoshua Bengio
A Structured Self-attentive Sentence Embedding
ICLR 2017.

Yang, Jie and Zhang, Yue
NCRF++: An Open-source Neural Sequence Labeling Toolkit
Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.