Preprocessing for seq2seq model

2

I am trying to build a seq2seq model , I tried to follow Tensorflow official tutorial but there is no preprocessing steps mentioned. I tried to search on web , every tutorial start from model , There is no preprocessing steps info.

I need some info about preprocessing steps involved in seq2seq :

If i have a dataset like this : (after encoding with index2word vocabulary)

```
encoder [1, 2, 1, 3, 4] decoder [2, 3, 4]
encoder [2, 3, 4, 1] decoder [11, 3, 4, 5, 1, 22, 45, 1, 3, 42, 32, 65]
encoder [4, 5, 3, 11, 23, 1, 33, 44, 1, 3] decoder [4, 2, 3, 5]
encoder [44, 55] decoder [5, 6, 3, 2, 4, 22, 42, 11, 34]
encoder [1] decoder [55, 6, 3, 2, 4, 5, 6, 7, 7]
encoder [4, 2, 3, 4, 5] decoder [6, 5, 3, 5, 6, 7, 8, 2, 4, 5]
encoder [44, 2, 1, 22, 5, 3, 2] decoder [6, 5, 3, 4, 5, 6, 7]
encoder [55, 3, 1, 5, 1] decoder [5, 3, 2, 3, 4, 5]
encoder [14] decoder [5, 6, 7]
```

If i take 5 as batch size then first batch:

```
encoder [1, 2, 1, 3, 4] decoder [2, 3, 4]
encoder [2, 3, 4, 1] decoder [11, 3, 4, 5, 1, 22, 45, 1, 3, 42, 32, 65]
encoder [4, 5, 3, 11, 23, 1, 33, 44, 1, 3] decoder [4, 2, 3, 5]
encoder [44, 55] decoder [5, 6, 3, 2, 4, 22, 42, 11, 34]
encoder [1] decoder [55, 6, 3, 2, 4, 5, 6, 7, 7]
```

Now after reading many articles i found there are four special tokens which you have to use for encoding data :

> `<PAD>` : During training, we'll need to feed our examples to the network in batches.
>
> `<EOS>` : This is another necessity of batching as well, but more on the decoder side. It allows us to tell the decoder where a sentence ends, and it allows the decoder to indicate the same thing in its outputs as well.
>
> `<UNK>` : replace unknown with .
>
> `<GO>` : This is the input to the first time step of the decoder to let the decoder know when to start generating output.

Now if i take my batch example then i have question after padding :

should encoder batch should be same size to decoder batch ?

If my padded encoder data batch looks like:

```
encoder_input=[[1, 2, 1, 3, 4],
[2, 3, 4, 1],
[4, 5, 3, 11, 23, 1, 33, 44, 1, 3],
[44, 55],
[1]]

#after padding ( max time stamp is 10 )

encoder_padded=[[1, 2, 1, 3, 4, 0, 0, 0, 0, 0],
                [2, 3, 4, 1, 0, 0, 0, 0, 0, 0],
                [4, 5, 3, 11, 23, 1, 33, 44, 1, 3],
                [44, 55, 0, 0, 0, 0, 0, 0, 0, 0],
                [1, 0, 0, 0, 0, 0, 0, 0, 0, 0]]
```

Now should i pad my decoder sequence length to same size ? (max 10 ?) or should i pad with decoder max sequence (max 12) like this:

```
decoder_input=[[2, 3, 4],
               [11, 3, 4, 5, 1, 22, 45, 1, 3, 42, 32, 65],
               [4, 2, 3, 5],
               [5, 6, 3, 2, 4, 22, 42, 11, 34],
               [55, 6, 3, 2, 4, 5, 6, 7, 7]]

#after padding ( decoder batch max length is 12)

decoder_padded=[[2, 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0],
               [11, 3, 4, 5, 1, 22, 45, 1, 3, 42, 32, 65],
               [4, 2, 3, 5, 0, 0, 0, 0, 0, 0, 0, 0],
               [5, 6, 3, 2, 4, 22, 42, 11, 0, 0, 0, 0],
               [55, 6, 3, 2, 4, 5, 6, 7, 7, 0, 0, 0]]
```

and how my last preprocessed data should look like :

```
encoder_input  = ['hello','how','are','you','<PAD>','<PAD>','<PAD'>]

decoder_output = ['<GO>','i','am','fine','<EOS>','<PAD>','<PAD>']
```

is this correct format ?

tensorflow machine-learning deep-learning rnn seq2seq

asked Jun 28 '18 at 19:16

Aaditya Ura

5,167 2 17 41

1 Answer

I hope this is useful.

0

> should encoder batch should be same size to decoder batch ?

No, decoder calculations follow the encoder, so the respective data will be fed to the network at separate times. The example you showed is correct.

One small correction in the last example what you mention as decoder_output should be decoder_input. For that pair of input, target label you should have:

```
encoder_input  = ['hello','how','are','you','<PAD>','<PAD>','<PAD'>]
decoder_input  = ['<GO>','i','am','fine','<EOS>','<PAD>','<PAD>']
target_label   = ['i','am','fine','<EOS>','<PAD>','<PAD>']
```