SS-4290: Computer Science Final Year Project Weekly Report

Week 7 (Due 16th September 2021)

Muhammad Syauqi Waiz bin Haji Sufri

17B9082

Objectives:

<u>Understand Memo architecture</u>

Memo architecture are inspired from End-to-End Memory Network. There are few differences between them:

1. The attention function where the input sentences and questions are passed through include DropOut which is to avoid overfitting during learning and LayerNorm which is similar to batch normalization.
2. MEMO utilize multi head attention with the key, value and query. Attention function are performed in parallel for each of these.

Firstly, input sentences are embedded with common embedding:

"Ali go office" – key/value >>> Vali + Vgo + Voffice

"Ali go kitchen" – key/value >>> Vali +Vgo + Vkitchen

"Ali drop milk" – key/value >>> Vali + Vdrop + Vmilk

All input sentences converted to vector and will go to the memory slots

Where is milk? – query (A: kitchen) >>> Vwhere + Vis + Vmilk *

* '?' Is not included because it is a pretrained vector/not necessary since it does not go to the memory

Key and Value are basically the words in an input sentence. Note that it can be the same.

Query is the question that will match with the key.

However, in MEMO, each of the word vector are combined with its one hot encoding:

e.g. Ali go office

Ali: [1 0 0]

go: [0 1 0]

office: [0 0 1]

This allows capturing flexibly in any part of the input sentence. These combinations will be stored in the memory slots. So, when a query comes in, they will perform an attention function.

An attention function can be described as mapping a query and a set of key-value pairs to an output. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. Below is how Multi-Head Attention work:
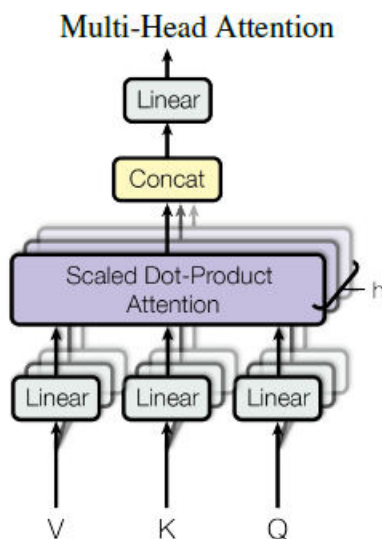
Also, LayerNorm and DropOut are performed here, where the output will be the answer to the query.

End