Learning Personas from Dialogues with Attentive Memory Networks: Supplementary Material

Eric Chu *
MIT Media Lab
echu@mit.edu

Prashanth Vijayaraghavan*
MIT Media Lab
pralav@mit.edu

Deb Roy
MIT Media Lab
dkroy@media.mit.edu

A Read-Write Memory Losses

A.1 Memory Ranking Loss J_{MR}

We want the model to learn to make efficient matches with the memory keys in order to facilitate look up on past data. To do this, we find a positive and negative neighbor after computing the k nearest neighbors $(n_1, ..., n_k)$ by finding the smallest index p such that $V[p] = P$ and n such that $V[n] \neq P$ respectively. We define the memory ranking loss as:

$$J_{MR} = max(0, s(z, K[n]) - s(z, K[p]) + \alpha_{MR}) \quad (1)$$

where α_{MR} is a learnable margin parameter and $s(\cdot, \cdot)$ denotes the similarity between persona embeddings (z), key representations of positive ($K[p]$) and negative ($K[n]$) neighbors. The above equation is consistent with the memory loss defined in the original work (Kaiser and Nachum, 2017).

A.2 Memory Classification Loss J_{MCE}

The Read-Write memory returns $\hat{z}_M$ and values $v_M = V[n_i]\ \forall i \in \{n_1, .., n_k\}$. The probability of the given input dialogues belonging to a particular persona category P is computed using values v_M returned from the memory via:

$$q^M = softmax(f_{P_M}(v_M)) \quad (2)$$

where $f_{P_M} : \mathbb{R}^k \mapsto \mathbb{R}^{N_P}$ is a fully-connected layer. We replace the q_j with q_j^M in Equation 12 and calculate the categorical cross entropy to get J_{MCE}.

B Experimental details

The vocabulary size was set to 20000. We used a GRU hidden size of 200, a word embedding size of 300, and the word embedding lookup was initialized with GLoVe (Pennington et al., 2014). For the Read-Write memory module, we used k=8 when calculating the nearest neighbors and a memory size of 150. Our models were trained using Adam (Kingma and Ba, 2014).

References

Łukasz Kaiser and Ofir Nachum. 2017. Aurko roy, and samy bengio. *Learning to remember rare events. arXiv preprint*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

*The first two authors contributed equally to this work