HW4

Problem 1 "Barack Obama" example. Consider a MLP with one hidden layer. Let x be the input (Barack), h be the hidden layer, and y be the output (Obama). Suppose both x and y are words in the same dictionary, where x is the current word, and y is the next word. Both x and y are one-hot vectors. Let $h = W_{\text{embed}}x$, $s = W_{\text{unembed}}h$, $p = \text{softmax}(s)$, and $y \sim p$, i.e., $p(y_c = 1|s) = p_c$.

(1) What are the dimensionalities of W_{embed} and W_{unembed}? Interpret the meaning of the columns of W_{embed}.

(2) Let $J = \log p(y|s)$, show that $\partial J/\partial s = y - p = e$. Calculate $\partial J/\partial h$, $\partial J/\partial W_{\text{embed}}$, and $\partial J/\partial W_{\text{unembed}}$ via chain rule. In your calculation, you can first pretend all the vectors and matrices are scalars (one-dimensional numbers), and then guess the forms of the general results.

(3) Draw a diagram of network with multiple layers of latent vectors.

Problem 2 "I love machine learning" example. Suppose we observe "I love machine" and we want to predict the next word. Let x_t be one-hot vectors, x_1 = "I", x_2 = "love", x_3 = "machine". Let h_t be the hidden vectors, $h_0 = 0$, $h_1 = \tanh(W_{\text{embed}}x_1 + W_{\text{recurrent}}h_0)$, $h_2 = \tanh(W_{\text{embed}}x_2 + W_{\text{recurrent}}h_1)$, $h_3 = \tanh(W_{\text{embed}}x_3 + W_{\text{recurrent}}h_2)$, and $s = W_{\text{unembed}}h_3$, $p = \text{softmax}(s)$, x_4 is sampled according to p.

(1) Draw a diagram to illustrate the model.

(2) Let $J = \log p(x_4|s)$. Calculate $\partial J/\partial W_{\text{embed}}$, $\partial J/\partial_{\text{unembed}}$, and $\partial J/\partial W_{\text{recurrent}}$. In your calculation, you can first pretend all the vectors and matrices are scalars (one-dimensional numbers), and then guess the forms of the general results.

(3) Draw a diagram of network with multiple recurrent layers of latent vectors.

Problem 3 Residual stream. For the "I love machine learning" example, consider the residual parametrization: $h_t = h_{t-1} + \tanh(W_{\text{recurrent}}h_{t-1} + W_{\text{embed}}x_t)$. Let $J = \log p(x_4|s)$, where $s = W_{\text{unembed}}h_3$. Starting from $\partial J/\partial h_3$, calculate $\partial J/\partial h_1$. Explain that it alleviates the gradient vanishing problem. Please explain that in both Problem 2 and Problem 3, $\partial J/\partial h_1$ can be computed only after $\partial J/\partial h_2$ is available, which in turn can be computed only after $\partial J/\partial h_3$ is available.

Problem 4 Please play with the PyTorch code provided by the following webpage:

`https://machinelearningmastery.com/text-generation-with-lstm-in-pytorch/`

Please write a brief explanation of the code and show your results. You can explore the code by varying the design parameters.