Pattern Recognition and Machine Learning: Homework 12, Zhengzuo Liu

Backprop Through a Simple RNN

Answer:

$$t = w(wu_1 + u_2) + u_3$$
$$y = w(w(wu_1 + u_2) + u_3)$$
$$\frac{dy}{dw} = 3u_1w^2 + 2u_2w + u_3$$
$$\frac{\partial y}{\partial p} = w^3$$

Problem 2

Answer:

Let us look at the relationship between gradients of adjacent loss.

$$\frac{\partial L}{\partial C_{T-1}} = \frac{\partial L}{\partial C_T}\frac{\partial C_T}{\partial C_{T-1}}$$
$$\begin{aligned}\frac{\partial C_T}{\partial C_{T-1}} &= \frac{\partial}{\partial C_{T-1}}[f_T \odot C_{T-1} + i_T \odot \tilde{C}_T] \\ &= A_t + B_t + C_t + D_t\end{aligned}$$

Where

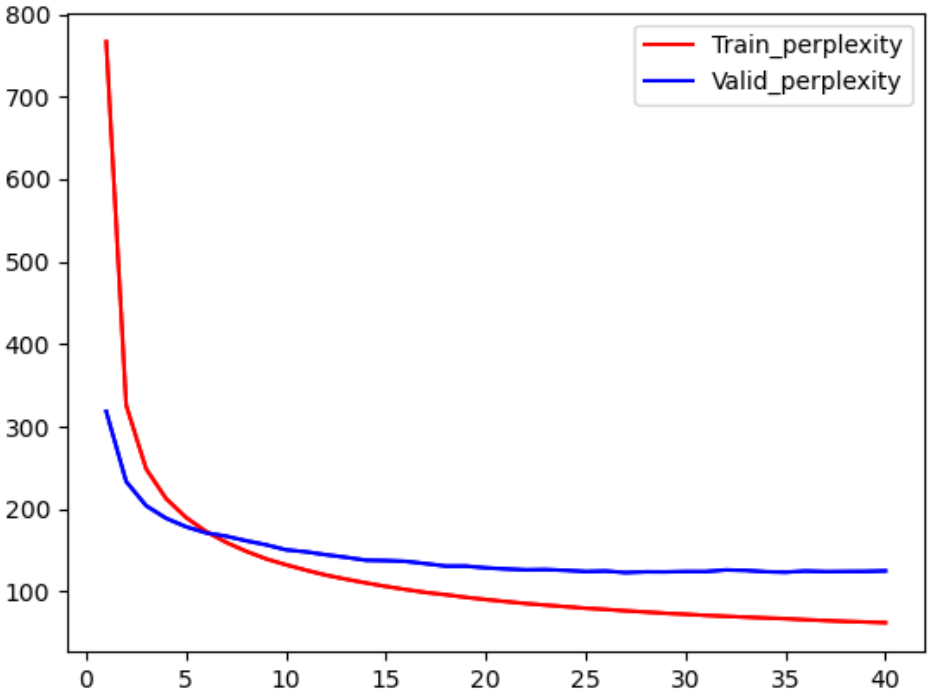
$$\begin{aligned}A_t &= \sigma'(x_T U^f + h_{T-1}W^f + b^f) \cdot W^f \cdot o_{T-1} \odot tanh'(C_{T-1}) \cdot C_{T-1} \\ B_t &= f_T \\ C_t &= \sigma'(x_T U^i + h_{T-1}W^i + b^i) \cdot W^i \cdot o_{T-1} \odot tanh'(C_{T-1}) \cdot \tilde{C}_T \\ D_t &= tanh'(x_T U^g + h_{T-1}W^g + b^g) \cdot W^g \cdot o_{T-1} \odot tanh'(C_{T-1}) \cdot i_T\end{aligned}$$

As seen in the equations above, the forget gate's vector (allows decision for forgetting) along with the additive structure (gains more balance) make it unlikely for the whole gradient to vanish very quickly.

Problem 3

Answer:

The training and validation curves are as follows:



Here's why we need a source mask: In a transformer model, we use source masks to prevent attention to future tokens. This is crucial for tasks like language modeling where are predicting future tokens and we don't want our model to peek into the future.

Problem 4

Preprocessing in text data need to include the step of tokenization, which is the process of breaking down text into smaller pieces (often words, but could be sentences, phrases, or even individual characters). Preprocessing in image data resizing, which is to resize images to a standard dimension so that the model receives consistently sized input.