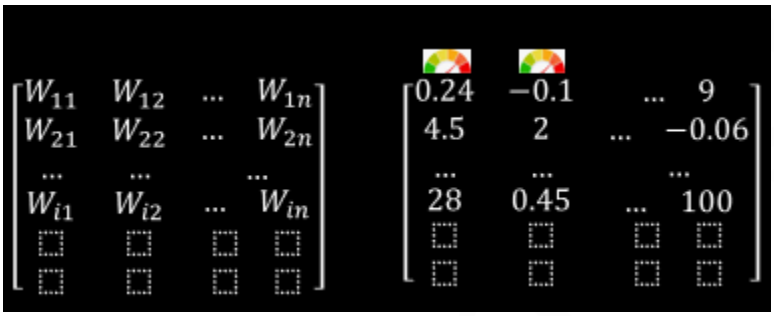
2025-12-10

The first parameter matrix we can handle is actually the last, because we work backwards. So the essential question is to 'tune' these parameters: all the elements in the W matrix. You start with random numbers, then you nudge each dial little by little, watch the response, to find out the best setting. But the question is: how to tune? Dial more or less, and how much? Yes we can solely rely on chain rule going backwards to work out the math, work out the formula telling us how fast the end performance changes with each setting, but then the chance for intuition might be lost. If you look at the formula, it's actually following quite some nice pattern: $\frac{\partial L}{\partial W_{ij}} = p_i h_j$ (non true token), and : $\frac{\partial L}{\partial W_{ij}} = p_i h_j$ (non true token), $\frac{\partial L}{\partial W_{yj}} = (p_y - 1)h_j$ (for the true token). p_i is the probability for each token, h_j is the value for each meaning component. So their magnitudes impact the sensitivity. First focus on the directional: is it getting greater or smaller? Because this derivative tells us how the performance changes with the setting, quantitatively: does it get better or worse? If it is positive, it is getting worse, if it is negative, it is getting better. Why? That's because of the particular way we value the performance: how big the error is from the true value, of course the smaller the better.



2025-12-12

So $\frac{\partial L}{\partial W_{ij}}$ tells us how fast the performance deteriorate with the parameter (setting), the bigger it is, the worse it is. So you want this value to be small. $p_i h_j$ is the formula to compute it. p_i is the probability the model assigns, and obviously since we are talking about a wrong token, the bigger this value, the bigger the deterioration rate. h_j is the j^{th} meaning component. Hmm, p_i is the probability predicted for i^{th} token, h_j is the j^{th} meaning component, isn't that strange? Why their magnitudes will be related to the performance at all? And what is W_{ij}? Sure it is a parameter, but which one? --- the i^{th} token, j^{th} meaning component. So this is not a parameter actually, a parameter is something you use to control others. But this is purely just the token itself, broken down into 'meaning components'!