Resumo: notas de aula – Stanford CS231n – Backpropagation intuitions

Backpropagation é uma forma de computar gradientes de expressões pela aplicação recursiva da regra da cadeia.

Dada uma função de perda (L) cujas entradas são os dados de treinamento, ou seja, as entradas (X) e saídas (Y) da rede, bem como os parâmetros da rede, ou seja, pesos (W) e os vieses (b). Em geral, em Machine Learning, consideramos os dados de treinamento como dados e fixos, e os pesos e vieses como variáveis que podemos controlar. Calculamos os gradientes pra os parâmetros de forma a fazermos uma modificação em parâmetros.

O gradiente é o vetor de derivadas parciais, cada qual indicando a taxa de mudança da função com respeito a uma variável na região infinitesimal em torno de um ponto particular. Isso pode ser visto como a aproximação de uma reta cuja inclinação é a derivada.

Uma operação formada pelo produto de duas variáveis – f(x,y) = x.y – possui como derivada parcial em uma variável o valor da outra variável naquele ponto. Uma operação de adição de duas variáveis – f(x,y) = x + y – possui derivadas parciais iguais a um. Uma operação de máximo – f(x,y) = max(x, y) – possui derivada um para a variável que assumir o valor máximo no ponto e zero para a outra variável.

Em expressões compostas por múltiplas operações, por exemplo f(x,y,z) = (x + y) . z, a regra da cadeia nos traz um forma de encadear as expressões de gradiente por meio de multiplicação das derivadas dessas operações, ou gradientes locais. Esse encadeamento pode ser visualizado por meio de diagramas de circuito, computando valores das funções na ida (forward pass) e valores de gradientes na volta (backward pass). O interessante desse processamento é que o cálculo do gradiente local independe dos detalhes do circuito completo ao qual ele pertence.

Para a computação do backward pass, é interessante armazenar os valores das variáveis e funções calculadas no forward pass, como fica claro no exemplo das derivadas da operação multiplicativa. Além disso, como as variáveis se repetem múltiplas vezes no circuito, é interessante calcular backpropagation com operações acumulativas dos gradientes ao invés de meramente reescrever os valores.

Nem sempre backpropagation traz os melhores resultados possíveis. Operações multiplicativas entre entradas de valores muito pequenos com outras de valores muito grandes geram gradientes pequenos em entradas com valores grandes e gradientes grandes em entradas com valores pequenos. O pré-processamento pode desempenhar um papel em uma otimização mais eficiente, assim como a escolha da melhor taxa de aprendizado.