WEEKLY PARTICIPATION 10: NORMALIZATION AND RESIDUAL CONNECTIONS

Normalization[1] is useful for reducing the time to fit models, and residual connections are useful for that reason as well as because they incorporate a desirable inductive bias. They are often used in conjunction.

Consider the choice of normalizing post-summation or pre-summation, i.e. taking

$$o^\ell = \mathrm{BN}\left(\sigma(\boldsymbol{W}^\ell o^{\ell-1} + \boldsymbol{b}^\ell) + o^{\ell-1}\right) \quad \text{or} \quad o^\ell = \mathrm{BN}\left(\sigma(\boldsymbol{W}^\ell o^{\ell-1} + \boldsymbol{b}^\ell)\right) + o^{\ell-1}.$$

Explain why post-summation normalization may be more attractive from an optimization perspective, and why pre-summation normalization may be more attractive from the perspective of inductive bias.

[1] Layer or batch normalization.