Paper | How Does | Batch Normalization work

2018

So these guys from MIT come in and say Well Batch norm undoubtedly Works but we don't realy know why.

They claim the Benifits of Batch norm come from the smoothing of the Landscape for the optimization problem and not from the ICS reduction even Claming that Batch Norm seems to not do this.

They show sample distributions of the random input to a from batch during training but we see the change in mean & variance with and without batch norm is marginal.

So this prompts them to ask 2 questions.

1. is Batch norm effectiveness acctrully related to the ICS?

2. Does Batch norm actually stabalize the input distributions aka does it actually reduce ICS?

So to test question #1

They add non zero mean non unit variance at the end of each batch norm layer output effectively ensuring there is a ICS. (Vatified by the distibution pots of the noisy network as compared to the other ones)

normal + normal with batch norm

So as far as ICS is tied to mean and variance of input distribution it seems to not matter to performance.

but what if There is a Broader idea of ICS that linked to performance?

Maybe Batch norm reduces this Broader Notion.

So they attack the problem from a new angle. Idea here is that the the gradient tells us How to update a layers weight acording to the other layers but assuming they stay constant. but in reality we change all the

Weights. They believe this is problematic so to study this we look at G_i which is the gradient for layer i & then G'_i is the gradient of layer i but after all previous layer weights have been updated. So we can have a diffrance $\|G' - G\|_2$ so the diffrance describes the change in the optimization landscape for W_i

So they look at this new metric ICS & find that Batch norm even increases it & no Batch norm NN has Low ICS as defined here!!!

So Why does Batch norm Work?

Well they say Because it Smooths out the optimization surface.

This lets us confidantly take Bigger steps.

They then Demonstrat through a series of graphs Proving the Smothness of loss.