Batch Norm | Feb 2015

So if we do Batch norm then Pass to a s.gmoid We could constraine the none linearity to it's Linear region.  this region

So then we add(γ, β) $\gamma \hat{x} + \beta$ normalized

these are Learned & can be addjusted, So We could recover the or.ginal activations prenormalization. [thats all from paper] So now why even bother doing this Work Just to have the possebility to undo it? Well the original outputs are Strongly dependent on the initatization of the weights In the layer Which means that the activations could Be large do to the fact that the weights Were randomly initalized some way. Now we are putting a prior that Says at the Start the activations Will have the Same shape But the mean and sd will be 0 & 1 respectivly & we can change that even going Back to the original activations However now it must be Justified, Doing so (changing β & γ) must be Justified by addiqetly Minimizing the loss. meaning the Magnitude of the activation is no longer Largly determined by some artetrary initial Weights but rather must be Justified by a Minimization of loss

So Why exploding & Vanishing gradient in Basic Recurrent Neural nets?

It's pretty damn simple.

Think of it ~~it~~ Like this in a network with no recurrent connections ~~xxx~~ our gradient is calculated Based only on the values of the current training example

But in a RNN

you get something Like

$$\frac{\partial L}{\partial w_i} = \ldots\ldots\ldots\ldots \cdot O_{t-1}$$

times

activation output at some previous time period

So if that was a huge output or a very small one it will still effect us.

but not only there

you will also get terms like

some activation output at the curr time

$$\frac{dL}{dw_i} = \ldots\ldots\ldots \cdot \overset{\text{tiny}}{O_{o_t}}$$

this looks like normal but

O_{o_t} can be calc like this if

it's the output of a recurrent connection.

$$\mathrm{RELU}(Xw_0 + b + w_i O_{o_{t-1}}) = O_{o_t}$$

So it too is effected by

pass out outputs. all this

means if we had an output

that was HUGE or tiny we

continue to have it effect out

weight & bias updates even after

the particular example has gone.

This is why RNN are more

prone to these issue (exploding vanishing grad)