March 2020

Desighning NETWORK Design Spaces

SO these guys come in and say hey up until now we have been looking at a model space then finding the best singular model ether Manually or automaticly Within that Space. But they say lets take it a step/layer above And find a space which describes a Subset of all models Where these models on avg Better But also "simpler, work well and generalize across settings"

They Start With an rather unconstrained space and progressivly constraing it "while maintaing or improving" the error distribution produced by the Models.

The least Constrained Space is called AnyNET and is as follows: (note look at paper for good drawings)

The Body is where they will Be defining the model + the Bulk of the Work Will Be done. Stem = Stide2 3×3 conv the 32 channels + the head is AvG Pooling + a FC layer There are 4 Stages in the Body Where each Stage i has b_i (Blocks) w_i (width) + other Block params.

Since each network has 4 Stages & each Stage has 4 degrees of freedom in total there are 16 degrees of freedom.

each Stage has d_i (Blocks) w_i (width), b_i (Bottle neck ratio), and g_i (group width)

$d_i \leq 16$ $\quad w_i \leq 1024$ (and divisable by 8)

$b_i = \{1, 2, 4\}$ $\quad g_i \in \{1, 2, \ldots 32\}$

So above is the AnyNet design space with 10^{18} possible model configs.

Step one they set all $b_i = 6$ So it's the same across all Stages of a model. They find no increas in error But now the design Space is Simpler.

Step two is to set $g_i = g$ as above & they find same result

Step 3 they find Pattern after Step 2 where increasing width over the Stages results in Better models So they test Anynet_D where AnyNet_C is after Step 2 & AnyNet_D is after Step 2 & only models where $w_{i+1} \geq w_i$ & find it Significantly Better distribution of error

Step 4 they find that similar as with step 3 if now we increase depth $d_{i+1} \geq d_i$ the models are Better.

So after all these reductions our design space went from 10^{18} posobility to 10^{7}

So then they come up with the final Design space Described as such:

ReyNET generated from: d, w_0, w_a, w_m

$d < 64$ $\quad w_0, w_a \leq 256$ But we Have to quantize $\quad$ Control the scaling of width

$1.5 \leq w_m \leq 3$ $\quad u_j = w_0 + w_a \cdot j$ via w_m

The original tests we have Been reading about are all done in the Low epochs & low compute range.
So Now they compare in higher compute higher epoch + 5 stages
the ordering is always ReyNetx > AnyNet$_E$ > AnyNet$_A$

Post Step 4 $\quad$ Pre Step 1

They then have further observations that the common $b < 1$ & $g = 1$ are not as good as $b = 1$ & $g \geq 1$

also

they also found optimal depth = 20 Blocks (interesting deeper not allways Better!)

& a width multiple of 2.5 close to the common one of 2

So Now lets compare ResNET model to other models.

RegNET Models tend to have lower Flops But maintain or Better results the ResNet

In general the RegNETS Matched or did Better than state of the art ResNET

and at low flop Efficent net Better But at Higher flops RegNET Better. & is much faster in the Higher flop regions

$d = 4$
Say: $w_0 = 32$
$w_a = 8$
$w_m = 2$

On Quantization: So we get Powers of 2

$32 = u_0 = 32 + 8 \cdot 0$	$\log\left(\frac{32}{32}\right) = 0$	Round $= 0$	$w_0 = 32 \cdot 2^0 = 32$
$40 = u_1 = 32 + 8 \cdot 1$	$\log(40/32) = .32$	$\Rightarrow\ = 0$	$w_1 = 32 \cdot 2^0 = 32$
$48 = u_2 = 32 + 8 \cdot 2$	$\log(48/32) = .58$	$= 1$	$w_2 = 32 \cdot 2^1 = 64$
$56 = u_3 = 32 + 8 \cdot 3$	$\log(56/32) = .81$	$= 1$	$w_3 = 32 \cdot 2^1 = 64$

Now 2 Stages

Stage 1	Stage 2
2 Blocks, 32 width	2 Blocks, 64 width