P-TELU: Parametric Tan Hyperbolic Linear Unit Activation for Deep Neural Networks

Rahul Duggal & Anubha Gupta,

SBILab, Dept. of E.C.E, IIIT Delhi, India

Motivation

This poster proposes a new activation function that leverages two specific advantages of previous functions

1. Optimal parameter tunability: like the RELU function
2. Better noise robustness: similar to ELU function

Past Activation Functions

1. RELU [1]:

$$f(z) = \begin{cases} z, & z > 0 \\ 0, & z \le 0 \end{cases}$$

- 6 times faster convergence than tanh.
- Unit gradient in the positive plane obviates exploding/vanishing gradients.
- Zero gradient in the negative plane leads to de-activated neurons.

2. L-RELU [2]:

$$f(z) = \begin{cases} z, & z > 0 \\ \alpha z, & z \le 0 \end{cases},$$

where α *is a small positive constant*

- Small gradient in negative plane solves the problem of de-activated neurons.
- optimal α needs to be set by hand.

3. P-RELU [3]:

$$f(z) = \begin{cases} z, & z > 0 \\ \alpha z, & z \le 0 \end{cases},$$

where α *learnt from the data distribution*

- The α parameter of L-RELU is allowed to be learnt from the data.

4. ELU [4]:

$$f(z) = \begin{cases} z, & z > 0 \\ \alpha(e^z - 1), & z \le 0 \end{cases},$$

where α *is a small positive constant*

- The exponential function in ELU provides a noise robust deactivation state for large negative inputs
- optimal α needs to be set by hand.

P-TELU Activation Function

The forward Pass

$$f(z) = \begin{cases} z, & z > 0 \\ \alpha.\tanh(\beta.z), & \alpha, \beta \ge 0 \end{cases},$$

where α and β are learnt from the data distribution.

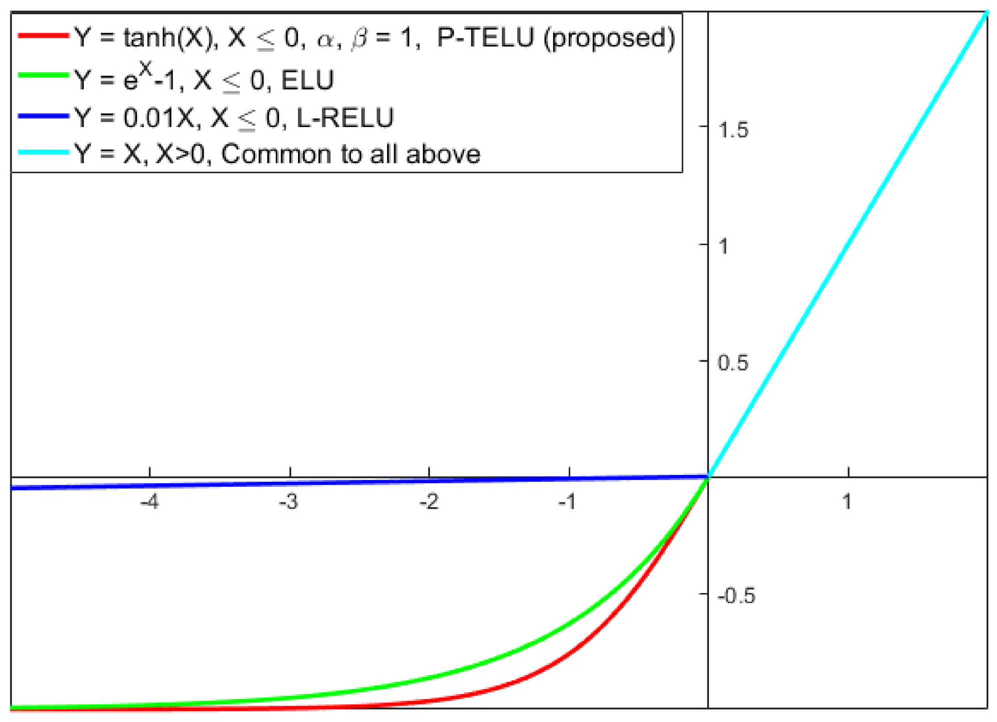

Fig 1: Graphical comparison of various activation functions

The backward Pass

Derivative of the loss w.r.t α of the j^{th} neuron in layer 'i'.

$$\frac{dL}{d\alpha_{i,j}} = \frac{dL}{df_{i,j}} \times \frac{df_{i,j}}{d\alpha_{i,j}} \quad (1)$$

Output of the j^{th} neuron in layer 'i'.

$$\frac{dL}{d\beta_{i,j}} = \frac{dL}{df_{i,j}} \times \frac{df_{i,j}}{d\beta_{i,j}} \quad (2)$$

Net Input to the j^{th} neuron in layer 'i'

$$\frac{df_{i,j}}{d\alpha_{i,j}} = \begin{cases} 0, & z_{i,j} > 0 \\ tanh(\beta_{i,j}.z), & z_{i,j} \ge 0 \end{cases} \quad (3)$$

$$\frac{df_{i,j}}{d\beta_{i,j}} = \begin{cases} 0, & z_{i,j} > 0 \\ \alpha \times z_{i,j} \times tanh(\beta_{i,j}.z), & z_{i,j} \ge 0 \end{cases} \quad (4)$$

Experiments and Results

CNN models considered

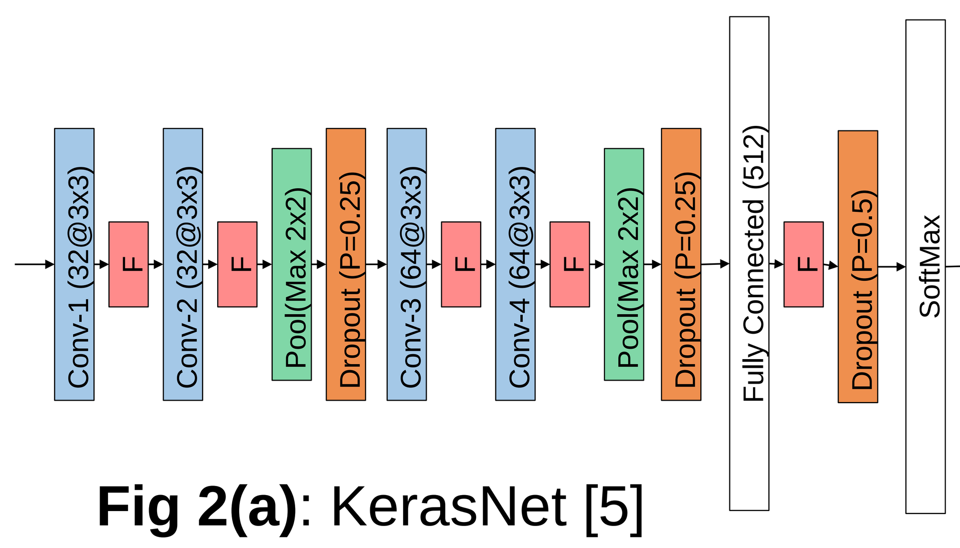

Fig 2(a): KerasNet [5]

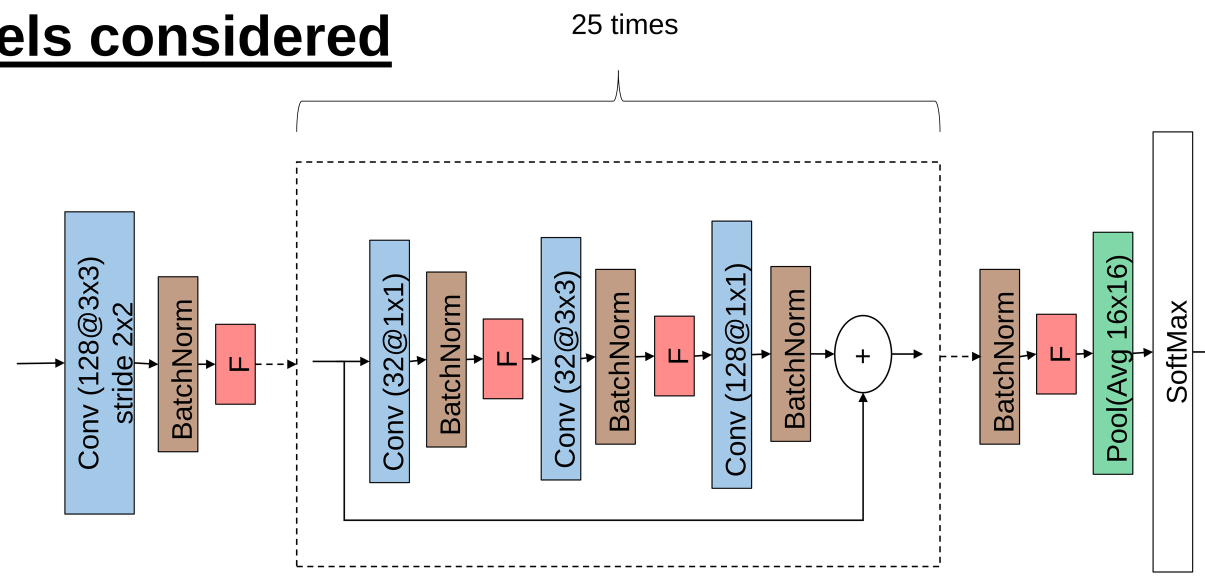

Fig 2(b): ResNet-76 [6]

Performance of KerasNet

Function	CIFAR-10	CIFAR-100
RELU	85.45	56.56
ELU	85.84	57.97
P-RELU	86.26	58.75
P-TELU	86.5	59.76

Table 1: Comparison of functions

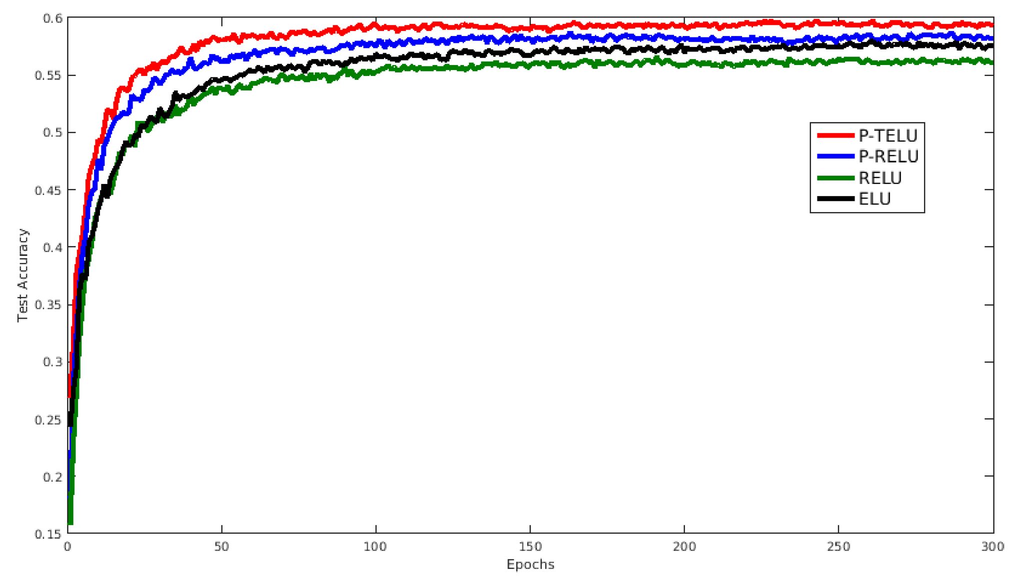

Fig 3: Test accuracy on CIFAR-100

Performance of ResNet-76

Function	CIFAR-10	CIFAR-100
RELU	90.77	70.06
ELU	91.26	69.05
P-RELU	90.99	69.05
P-TELU, with $\alpha, \beta \ge 0$	91.52	70.13
P-TELU, with $\alpha, \beta \ge 0.01$	91.16	70.63

Table 2: Comparison of functions

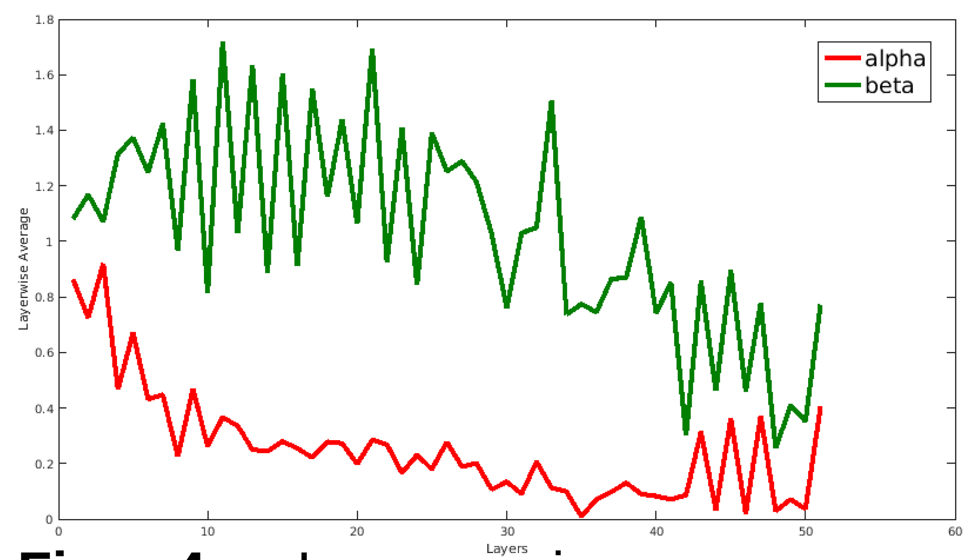

Fig 4: Layer wise average of α, β ($\alpha, \beta \ge 0$) calculated on a ResNet-76 model fitted with P-TELU

References

[1] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

[2] Maas, Andrew L., Awni Y. Hannun, and Andrew Y. Ng. "Rectifier nonlinearities improve neural network acoustic models." *Proc. ICML*. Vol. 30. No. 1. 2013.

[3] He, Kaiming, et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." *Proceedings of the IEEE international conference on computer vision*. 2015.

[4] Clevert, Djork-Arné, Thomas Unterthiner, and Sepp Hochreiter. "Fast and accurate deep network learning by exponential linear units (elus)." *arXiv preprint arXiv:1511.07289* (2015).

[5] Veličković, Petar, et al. "X-CNN: Cross-modal convolutional neural networks for sparse datasets." *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on*. IEEE, 2016.

[6] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

Acknowledgement

Authors gratefully acknowledge the research funding support (Grant Number: 1(7)/2014-ME&HI) from the Ministry of Communication and IT, Govt of India for this research work.