# A Limitation

While this paper contains an in-depth analysis of Tanh, it does not propose a function that is superior to existing activation functions. Generally, it is known that Tanh does not perform better than ReLU, but it's hard to find a study that deeply analyzes the reasons for this. To our knowledge, this paper is the most comprehensive analysis of the performance degradation when Tanh is used in a conventional manner. Although it suggests strategies to utilize Tanh more effectively, achieving performance nearly on par with ReLU, it fundamentally does not assert that Tanh surpasses ReLU in terms of effectiveness.

# B Accuracy of the Shifted Tanh According to $\tau$

In this section, we focus on the hyperparameter $\tau$ in the shifted Tanh function. We carry out experiments using VGG16_11, MobileNet, and PreAct-ResNet models trained on CIFAR-10 and CIFAR-100 datasets, testing different values of the parameter $\tau$ (-1.5, -1.2, -1.0, -0.8, and -0.5). The performance results corresponding to each of these $\tau$ values are presented in Table B.1. In these experiments, all hyperparameters, except for $\tau$, are maintained at their best-averaged accuracy settings based on each model trained on the CIFAR dataset. Our aim is to identify the effective $\tau$ value without pushing all inputs into the saturation zone due to excessive asymmetry. The best-performance $\tau$ is identified by averaging the accuracy across different models for each $\tau$ and selecting the one with the highest performance. In our case, -1 is the best.

Table B.1: The results of shifted Tanh models on various $\tau$ values.

| $\tau$ | $-1.5$ | -1.2 | -1.0 | -0.8 | -0.5 |
|---|---|---|---|---|---|
| Accuracy | 80.56 | 83.86 | **84.15** | 84.01 | 83.58 |

# C Metric

## C.1 Saturation and Skewness on Different Distribution Patterns

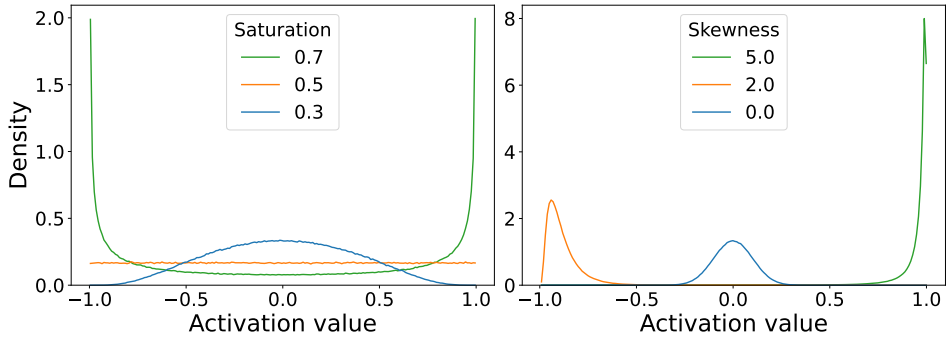


Figure C.1: The Saturation (left) and Skewness (right) values on different distributions.

Our saturation metric is designed to quantify the concentration of elements to the maximum absolute output values. The metric has bounds from 0 to 1, and it equates to 0 when all the elements are 0 in the Tanh case. Conversely, it grows larger as elements tend towards the maximum limit of the output range. For example, the saturation metric is evaluated at 0.5 in a uniform distribution. For saturation values of the Tanh output on the different Gaussian distributions, refer to Figure C.1 (left).

Skewness is a measure that quantifies the asymmetry within a distribution. Symmetric distribution will yield the Skewness of 0, and the Skewness increases with the rise in asymmetry. Notably, we calculate the absolute skewness value in our asymmetry metric, ensuring that the metric remains

unaffected by the direction of skewness. For a better understanding of how this measure applies to various distributions, Figure C.1 (right) provides Skewness on each distribution.

### C.2 Impacts of Mean and Standard Deviation Variations on Tanh Asymmetry

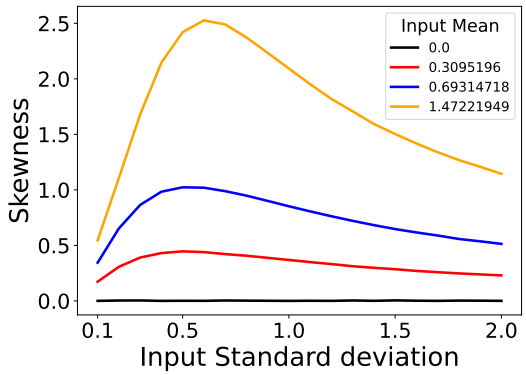


Figure C.2: Skewness value on the different means and standard deviations of the Gaussian distribution for Tanh input.

The mean and variance of input distribution on Tanh affect the asymmetry of Tanh output. The Skewness of Tanh output on various mean and standard deviation can be found in Figure C.2. In the Skewness of varied mean distribution, the Skewness is increased on the mean increase. However, the maximum Skewness does not align with the standard deviation increases. The Skewness decreases not only the small input standard deviation but also the large input standard deviation. Additionally, in the same mean condition, a decrease in standard deviation from the maximum skewness point more rapidly decreases the Skewness than an increase in standard deviation.

### C.3 The Proof of Demonstrating the Sparsity Metric Criteria

This section examines the adequacy of our sparsity metric against the heuristic criteria for sparsity measures defined by Hurley & Rickard [1]. Our sparsity metric meets five of these criteria: Scaling, Rising Tide, Cloning, Bill Gates, and Babies. Without loss of generality, we consider the absolute operation as pre-applied to the metric input, according to Hurley & Rickard [1], and set the number of channels as one for simplicity. Thus, our sparsity metric is represented as $f : \mathbf{x} \mapsto s$, where $\mathbf{x} \in \mathbb{R}^{m+}$ is the vector of absolute values of one channel in BN output, i.e., $[|B_{i,0}|, |B_{i,1}|, ..., |B_{i,m}|]$, and $x^{\max +}$ signifies the maximum value in $\mathbf{x}$, i.e., $x^{\max +} = \max_i x_i$. Below are the five criteria our sparsity metric meets, along with proof for each.

**Proof of Scaling.** Sparsity is scale invariant. That is, for any scalar $\alpha \in \mathbb{R}$, where $\alpha > 0$, the function $f$ satisfies:

$$f(\alpha \mathbf{x}) = f(\mathbf{x}). \tag{1}$$

The invariance of scale means that sparsity considers relative differences rather than absolute magnitudes.

*Proof.* If we scale the vector $\mathbf{x}$ by $\alpha$, we also scale the maximum positive value $x^{\max +}$ by $\alpha$. Therefore, we have:

$$\begin{aligned} f(\alpha \mathbf{x}) &= 1 - \frac{\Sigma_{i=1}^{m} \alpha x_i}{m \alpha x^{\max +}} \\ &= 1 - \frac{\Sigma_{i=1}^{m} x_i}{m x^{\max +}} \\ &= f(\mathbf{x}). \end{aligned}$$

This demonstrates that the sparsity of the vector $\mathbf{x}$ remains the same when the vector is scaled by any positive scalar $\alpha$. □

**Proof of Rising Tide.** Adding a constant to each element decreases sparsity. Formally, for any scalar $\alpha \in \mathbb{R}$, where $\alpha > 0$, the function $f$ satisfies:

$$f(\alpha + \mathbf{x}) < f(\mathbf{x}).$$

We exclude the case, as mentioned in [1], where all elements of $\mathbf{x}$ are the same.

This property also indicates that sparsity increases as more values approach zero.

*Proof.* We begin by noting the sparsity of the vector $\mathbf{x} + \alpha$:

$$f(\mathbf{x} + \alpha) = 1 - \frac{\Sigma_{i=1}^{m} x_i + m\alpha}{m x^{\max +} + m\alpha}.$$

Assume for contradiction that $f(\mathbf{x} + \alpha) \geq f(\mathbf{x})$. This leads to the following inequalities:

$$m x^{\max +} > \Sigma_{i=1}^{m} x_i \implies \frac{\Sigma_{i=1}^{m} x_i + m\alpha}{m x^{\max +} + m\alpha} > \frac{\Sigma_{i=1}^{m} x_i}{m x^{\max +}}, \tag{2}$$

$$1 - \frac{\Sigma_{i=1}^{m} x_i + m\alpha}{m x^{\max +} + m\alpha} < 1 - \frac{\Sigma_{i=1}^{m} x_i}{m x^{\max +}}, \tag{3}$$

$$\tag{4}$$

which contradicts the initial assumption, thus implying that

$$f(\alpha + \mathbf{x}) < f(\mathbf{x}),$$

as desired. □

**Proof of Cloning.** Sparsity is invariant under cloning.

That is, for a vector $\mathbf{x}$, the function $f$ satisfies:

$$f(\mathbf{x}) = f(\mathbf{x}||\mathbf{x}) = f(\mathbf{x}||\mathbf{x}||\mathbf{x}) = \ldots = f(\mathbf{x}||\mathbf{x}||\ldots||\mathbf{x}),$$

where $||$ denotes concatenation, such that $\mathbf{x}||\mathbf{x} = [x_1, x_2, \ldots, x_m, x_1, x_2, \ldots, x_m]$.

The principle of cloning implies that sparsity remains the same even if a set of values is replicated.

*Proof.* Define a function $g : \mathbf{x}, \delta \mapsto \mathbf{y}$, where $\delta \in \mathbb{Z}^+$ and $\mathbf{y} \in \mathbb{R}^{m\delta}$ is the result of concatenating $\mathbf{x}$ $\delta$ times. We then have:

$$\begin{aligned} f(g(\mathbf{x}, \delta)) &= 1 - \frac{\delta \Sigma_{i=1}^{m} x_i}{\delta m x^{\max +}} \\ &= 1 - \frac{\Sigma_{i=1}^{m} x_i}{m x^{\max +}} \\ &= f(\mathbf{x}). \end{aligned}$$

This shows that the sparsity of $\mathbf{x}$ remains the same even when it is concatenated with itself $\delta$ times. □

**Proof of Bill Gates.** As one individual element becomes infinitely large, the sparsity increases. Formally, for every $i$, there exists $\rho > 0$, such that for all $\alpha > 0$:

$$f([x_1, ..., x_i + \rho + \alpha, ..., x_m]) > f([x_1, ..., x_i + \rho, ..., x_m])$$

The implication is that if a single value grows without bounds, the sparsity also increases indefinitely.

*Proof.* We first choose a sufficiently large $\rho$ such that $\forall i, x_i + \rho > x^{\max +}$.

Assume for contradiction that $S([x_1, ..., x_i + \rho + \alpha, ..., x_m]) \leq S([x_1, ..., x_i + \rho, ..., x_m])$.

This implies:

$$
\begin{aligned}
1 - \frac{\Sigma_{k=1}^{m} x_k + \rho + \alpha}{m(x_i + \rho + \alpha)} &\leq 1 - \frac{\Sigma_{k=1}^{m} x_k + \rho}{m(x_i + \rho)} \\
\Leftrightarrow \frac{\Sigma_{k=1}^{m} x_k + \rho + \alpha}{m(x_i + \rho + \alpha)} &\geq \frac{\Sigma_{k=1}^{m} x_k + \rho}{m(x_i + \rho)} \\
\Leftrightarrow \frac{\Sigma_{k \neq i} x_k + x_i + \rho + \alpha}{x_i + \rho + \alpha} &\geq \frac{\Sigma_{k \neq i} x_k + x_i + \rho}{x_i + \rho} \\
\Leftrightarrow \frac{\Sigma_{k \neq i} x_k}{x_i + \rho + \alpha} &\geq \frac{\Sigma_{k \neq i} x_k}{x_i + \rho} \\
\Leftrightarrow \frac{1}{x_i + \rho + \alpha} &\ngeq \frac{1}{x_i + \rho},
\end{aligned}
$$

Leading to a contradiction. Therefore,

$$
f([x_1, ..., x_i + \rho + \alpha, ..., x_m]) > f([x_1, ..., x_i + \rho, ..., x_m]).
$$

□

**Proof of Babies.** Adding a new element of zero increases sparsity. Formally, for a vector $\mathbf{x}$, the function $f$ satisfies:

$$
f(\mathbf{x}||0) > f(\mathbf{x}) \tag{5}
$$

Concatenating zero elements to the existing values increases the relative difference to the other values, which increases sparsity.

*Proof.*

$$
1 - \frac{\Sigma_{k=1}^{m} x_k}{m+1} > 1 - \frac{\Sigma_{k=1}^{m} x_k}{m}
$$

We start by subtracting one from both sides of the inequality:

$$
-\frac{\Sigma_{k=1}^{m} x_k}{m+1} > -\frac{\Sigma_{k=1}^{m} x_k}{m}.
$$

The negative sign can be removed by reversing the inequality:

$$
\frac{\Sigma_{k=1}^{m} x_k}{m+1} < \frac{\Sigma_{k=1}^{m} x_k}{m}.
$$

Multiplying both sides of the inequality by $m(m+1)$ (since $m$ is a positive integer, $m+1$ is also positive, and the inequality sign will not change), we get:

$$
m\Sigma_{k=1}^{m} x_k < (m+1)\Sigma_{k=1}^{m} x_k.
$$

Subtracting $m\Sigma_{k=1}^{m} x_k$ from both sides of the inequality, we obtain:

$$
0 < \Sigma_{k=1}^{m} x_k.
$$

Since $x_k$ is not a negative value, the sum is also non-negative, which verifies the inequality. Therefore, the original inequality is proven. □

# D Deciding Model Configurations for Investigation

## D.1 Investigation of Depth Impact on Accuracy in VGG16 Models

Table E.2: Experimental results of shortened VGG16 models with the Swap order for CIFAR-100. The number of removed convolution layers in the VGG16_n model is the difference between 16 and n.

| | **VGG16** | **VGG16_15** | **VGG16_14** | **VGG16_13** | **VGG16_12** | **VGG16_11** | **VGG16_10** | **VGG16_9** | **VGG16_8** |
|---|---|---|---|---|---|---|---|---|---|
| Accuracy | 72.17 | 73.02 | 73.48 | 73.85 | 73.76 | **73.92** | 72.57 | 70.91 | 70.69 |

Table E.3: Experimental comparison between the official VGG11 and our VGG16_11 models trained by CIFAR-100.

| **Models** | **Order** | |
|---|---|---|
| | Convention | Swap |
| VGG11 | 64.55 | 69.94 |
| VGG16_11 | 69.5 | 74.11 |

To identify a model for focused analysis using the CIFAR dataset, we examined various VGG16 variants. This examination progressively removes convolution layers from the end towards the front of VGG16. The result can be seen in Table E.2. We observed an increase in accuracy until peaking at the VGG16_11 model, followed by a decline. Although a VGG11 model has already been proposed in Simonyan & Zisserman [2], the validation accuracy of VGG16_11 is significantly higher than VGG11 on CIFAR-100. The results can be seen in Table E.3

# E Training Hyperparameters

We sweep the learning rate and weight decay hyperparameter. The learning rate was 0.1 and 0.01. For CIFAR and Tiny-ImageNet datasets, we trained models with a batch size of 128, and the learning rate was reduced by one-tenth at 100 and 150 of the total 200 epochs, and we swept four weight decay of 0.005, 0.001, 0.0005, and 0.0001. For ImageNet datasets, we trained models with a batch size of 256, and the learning rate was reduced by one-tenth at 30 and 60 of the total 100 epochs, and we swept three weight decay of 0.001, 0.0005, and 0.0001. We chose the best averaged-accuracy model for the three random seeds and averaged the values of these three models for all measurements for analysis. Because of the computation issue, we only use one seed for the ImageNet dataset with early stopping. The hyperparameters of VGG, MobileNet, and PreAct-ResNet are shown in E.1, E.2, E.3, respectively.

### E.1 VGG

Table F.1: Hyperparameters of VGG11 with Tanh

| | **Convention** | **Swap** | **NWDBN** |
|---|---|---|---|
| Training Epochs | 200 | 200 | 200 |
| Learning Rate | 0.1 | 0.1 | 0.1 |
| Learning Rate Drop | 100, 150 | 100, 150 | 100, 150 |
| Weight Decay | 0.0005 | 0.0005 | 0.0005 |
| Batch Size | 128 | 128 | 128 |

Table F.2: Hyperparameters of VGG16 with Tanh

| | **Convention** | | | | **Swap** | | | |
|---|---|---|---|---|---|---|---|---|
| | **CIFAR-10** | **CIFAR-100** | **Tiny ImageNet** | **ImageNet** | **CIFAR-10** | **CIFAR-100** | **Tiny ImageNet** | **ImageNet** |
| Training Epochs | 200 | 200 | 200 | 90 | 200 | 200 | 200 | 90 |
| Learning Rate | 0.1 | 0.01 | 0.01 | 0.01 | 0.01 | 0.1 | 0.01 | 0.01 |
| Learning Rate Drop | 100, 150 | 100, 150 | 100, 150 | 30, 60 | 100, 150 | 100, 150 | 100, 150 | 30, 60 |
| Weight Decay | 0.0001 | 0.0005 | 0.001 | 0.0001 | 0.001 | 0.0005 | 0.001 | 0.001 |
| Batch Size | 128 | 128 | 128 | 256 | 128 | 128 | 128 | 256 |

Table F.3: Hyperparameters of VGG16 with the shifted Tanh

| | **Convention** | | | | **Swap** | | | |
|---|---|---|---|---|---|---|---|---|
| | **CIFAR-10** | **CIFAR-100** | **Tiny ImageNet** | **ImageNet** | **CIFAR-10** | **CIFAR-100** | **Tiny ImageNet** | **ImageNet** |
| Training Epochs | 200 | 200 | 200 | 90 | 200 | 200 | 200 | 90 |
| Learning Rate | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| Learning Rate Drop | 100, 150 | 100, 150 | 100, 150 | 30, 60 | 100, 150 | 100, 150 | 100, 150 | 30, 60 |
| Weight Decay | 0.0001 | 0.0005 | 0.0001 | 0.0001 | 0.0005 | 0.0005 | 0.0005 | 0.0001 |
| Batch Size | 128 | 128 | 128 | 256 | 128 | 128 | 128 | 256 |

Table F.4: Hyperparameters of VGG16 with the ReLU

| | **Convention** | | | | **Swap** | | | |
|---|---|---|---|---|---|---|---|---|
| | **CIFAR-10** | **CIFAR-100** | **Tiny ImageNet** | **ImageNet** | **CIFAR-10** | **CIFAR-100** | **Tiny ImageNet** | **ImageNet** |
| Training Epochs | 200 | 200 | 200 | 90 | 200 | 200 | 200 | 90 |
| Learning Rate | 0.01 | 0.01 | 0.1 | 0.1 | 0.01 | 0.01 | 0.01 | 0.01 |
| Learning Rate Drop | 100, 150 | 100, 150 | 100, 150 | 30, 60 | 100, 150 | 100, 150 | 100, 150 | 30, 60 |
| Weight Decay | 0.001 | 0.005 | 0.0001 | 0.0001 | 0.001 | 0.005 | 0.001 | 0.0005 |
| Batch Size | 128 | 128 | 128 | 256 | 128 | 128 | 128 | 256 |

## E.2 MobileNet

Table F.5: Hyperparameters for MobileNet with Tanh

| | Convention | | | | Swap | | | |
|---|---|---|---|---|---|---|---|---|
| | **CIFAR-10** | **CIFAR-100** | **Tiny ImageNet** | **ImageNet** | **CIFAR-10** | **CIFAR-100** | **Tiny ImageNet** | **ImageNet** |
| Training Epochs | 200 | 200 | 200 | 90 | 200 | 200 | 200 | 90 |
| Learning Rate | 0.1 | 0.1 | 0.01 | 0.1 | 0.1 | 0.1 | 0.01 | 0.1 |
| Learning Rate Drop | 100, 150 | 100, 150 | 100, 150 | 30, 60 | 100, 150 | 100, 150 | 100, 150 | 30, 60 |
| Weight Decay | 0.0001 | 0.0005 | 0.005 | 0.0001 | 0.0001 | 0.0005 | 0.005 | 0.0001 |
| Batch Size | 128 | 128 | 128 | 256 | 128 | 128 | 128 | 256 |

Table F.6: Hyperparameters for MobileNet with the Shifted Tanh

| | Convention | | | | Swap | | | |
|---|---|---|---|---|---|---|---|---|
| | **CIFAR-10** | **CIFAR-100** | **Tiny ImageNet** | **ImageNet** | **CIFAR-10** | **CIFAR-100** | **Tiny ImageNet** | **ImageNet** |
| Training Epochs | 200 | 200 | 200 | 90 | 200 | 200 | 200 | 90 |
| Learning Rate | 0.1 | 0.1 | 0.01 | 0.1 | 0.1 | 0.1 | 0.01 | 0.01 |
| Learning Rate Drop | 100, 150 | 100, 150 | 100, 150 | 30, 60 | 100, 150 | 100, 150 | 100, 150 | 30, 60 |
| Weight Decay | 0.0001 | 0.0005 | 0.005 | 0.0001 | 0.0005 | 0.0005 | 0.005 | 0.0005 |
| Batch Size | 128 | 128 | 128 | 256 | 128 | 128 | 128 | 256 |

Table F.7: Hyperparameters for MobileNet with ReLU

| | Convention | | | | Swap | | | |
|---|---|---|---|---|---|---|---|---|
| | **CIFAR-10** | **CIFAR-100** | **Tiny ImageNet** | **ImageNet** | **CIFAR-10** | **CIFAR-100** | **Tiny ImageNet** | **ImageNet** |
| Training Epochs | 200 | 200 | 200 | 90 | 200 | 200 | 200 | 90 |
| Learning Rate | 0.01 | 0.01 | 0.1 | 0.01 | 0.01 | 0.01 | 0.1 | 0.1 |
| Learning Rate Drop | 100, 150 | 100, 150 | 100, 150 | 30, 60 | 100, 150 | 100, 150 | 100, 150 | 30, 60 |
| Weight Decay | 0.001 | 0.005 | 0.0001 | 0.0001 | 0.001 | 0.005 | 0.001 | 0.0001 |
| Batch Size | 128 | 128 | 128 | 256 | 128 | 128 | 128 | 256 |

### E.3 PreAct-ResNet

Table F.8: Hyperparameters for PreAct-ResNet-18/34 with Tanh. PreAct-ResNet-18 is for the CIFAR dataset, and PreAct-ResNet-34 is for the Tiny ImageNet dataset.

| | **Convention** | | | **Swap** | | |
|---|---|---|---|---|---|---|
| | **CIFAR-10** | **CIFAR-100** | **Tiny ImageNet** | **CIFAR-10** | **CIFAR-100** | **Tiny ImageNet** |
| Training Epochs | 200 | 200 | 200 | 200 | 200 | 200 |
| Learning Rate | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| Learning Rate Drop | 100, 150 | 100, 150 | 100, 150 | 100, 150 | 100, 150 | 100, 150 |
| Weight Decay | 0.0001 | 0.0001 | 0.0001 | 0.0005 | 0.0005 | 0.0005 |
| Batch Size | 128 | 128 | 128 | 128 | 128 | 128 |

Table F.9: Hyperparameters for PreAct-ResNet-18/34 with the Shifted Tanh. PreAct-ResNet-18 is for the CIFAR dataset, and PreAct-ResNet-34 is for the Tiny ImageNet dataset.

| | **Convention** | | | **Swap** | | |
|---|---|---|---|---|---|---|
| | **CIFAR-10** | **CIFAR-100** | **Tiny ImageNet** | **CIFAR-10** | **CIFAR-100** | **Tiny ImageNet** |
| Training Epochs | 200 | 200 | 200 | 200 | 200 | 200 |
| Learning Rate | 0.1 | 0.1 | 0.01 | 0.1 | 0.1 | 0.1 |
| Learning Rate Drop | 100, 150 | 100, 150 | 100, 150 | 100, 150 | 100, 150 | 100, 150 |
| Weight Decay | 0.0001 | 0.0005 | 0.001 | 0.0005 | 0.0005 | 0.0005 |
| Batch Size | 128 | 128 | 128 | 128 | 128 | 128 |

Table F.10: Hyperparameters for PreAct-ResNet-18/34 with ReLU. PreAct-ResNet-18 is for the CIFAR dataset, and PreAct-ResNet-34 is for the Tiny ImageNet dataset.

| | **Convention** | | | **Swap** | | |
|---|---|---|---|---|---|---|
| | **CIFAR-10** | **CIFAR-100** | **Tiny ImageNet** | **CIFAR-10** | **CIFAR-100** | **Tiny ImageNet** |
| Training Epochs | 200 | 200 | 200 | 200 | 200 | 200 |
| Learning Rate | 0.01 | 0.01 | 0.1 | 0.01 | 0.01 | 0.1 |
| Learning Rate Drop | 100, 150 | 100, 150 | 100, 150 | 100, 150 | 100, 150 | 100, 150 |
| Weight Decay | 0.005 | 0.005 | 0.0005 | 0.005 | 0.005 | 0.0005 |
| Batch Size | 128 | 128 | 128 | 128 | 128 | 128 |

# References

[1] Hurley, N. and Rickard, S. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741, 2009.

[2] Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.