Value bounds and Convergence Analysis for Averages of LRP attributions

Supplementary Material

9. Exemplary results on LRP Hyperparameters and faithfulness

Fig. 2 shows that faithfulness degenerates for too high choices of the β hyperparameter in LRP-β and for too low choices of the γ hyperparameter in LRP-γ.

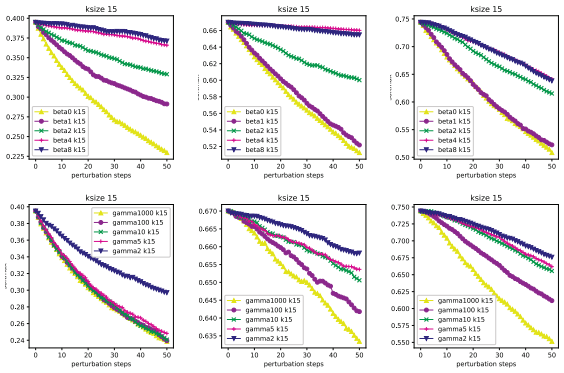

Figure 2. MoRF faithfulness graphs for Res50, EffNetV2-S, SwinV2-T (in that order) for varying β (top row) and unlifted γ (bottom row). Results are averages over 1000 ImageNet test images. This compares suitable against too extreme choices. Lower is better.

This can also be seen in examples of attribution maps as in Fig. 3.

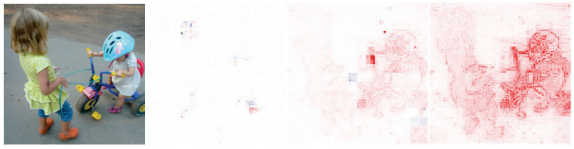

Figure 3. Exemplary Attribution maps for a too low choice, an appropriate choice of the γ hyperparameter, and for the autotuned lifted-γ, namely $\gamma = 10$, $\gamma = 1000$ and lifted-$\gamma = 10$. The network is SwinV2-Tiny, the predicted class is tricycle.

10. Analysis of Singular values

Before we bring our main theoretical result in section 4, we would like to show the results which can be obtained when we analysing LRP from the perspective of singular values of the above attribution matrices $M(g)$.

This provides an easy to obtain insight into the scales of attribution map values across layers and may serve as an initial comparison to the gradient. It will also reveal a limitation of this SVD-based approach.

Theorem 6 (Singular value for the vector of ones) *For any relevance conserving rule of a neural network layer which maps an S-dimensional input onto an R-dimensional output, a singular value of its one-layer transition is given by $\frac{\sqrt{R}}{\sqrt{S}}$, attained for the singular vector $\frac{1}{\sqrt{S}}1_S = \frac{1}{\sqrt{S}}\underbrace{(1,\dots,1)}_{S \text{ times}}^\top$, where R is the output dimension and S the input dimension for the layer in consideration.*

Proof:

$$\frac{1}{\sqrt{S}}1_S^\top M M^\top \frac{1}{\sqrt{S}}1_S = \frac{1}{S}1_R^\top 1_R = \frac{R}{S} \tag{39}$$

The importance of this simple theorem is to show a dependence of the singular values on the output dimensionality R of a layer. This motivates the insight, that observing a term $\sqrt{R}$ in the next theorem 7 is not an artefact of suboptimal proof technique but rather a necessity.

Theorem 7 (Upper bound for singular values for LRP-β) *Let a neural network layer compute a mapping of an S-dimensional input onto an R-dimensional output. For the β-rule we can derive an upper bound on the singular values $\sqrt{R}\sqrt{(1+\beta)^2+\beta^2}$, and as a better readable relaxation $\sqrt{R}(1+\sqrt{2}\beta)$*

The proof for it is in the Supplemental material in Section 13.

Corollary 8 ***Upper bound for singular values for LRP-γ in the limit case***
Let a neural network layer compute a mapping of an S-dimensional input onto an R-dimensional output. In the limit of $\gamma \to \infty$ the upper bound of singular values for LRP-γ is $\sqrt{R}$

This follows from the combination of Known Result 1, which establishes a convergence to LRP-β with $\beta = 0$, the fact that singular values of a real-valued matrix M are the positive eigenvalues of a matrix

$$\begin{pmatrix} 0 & M \\ M^\top & 0 \end{pmatrix} \tag{40}$$

and a continuity result such as Weyl's eigenvalue bound for additive perturbations [53] which can be found in textbooks like [24], and which ensures convergence when taking the limit $\gamma \to \infty$ for the above result $\sqrt{R}(1+\sqrt{2}\beta)$ for the case $\beta = 0$.

10.1. Comparison to the norm of the gradient attribution map

Let us assume that we have Lipschitz continuity for the activation functions σ_i with constant L. Then

$$\begin{aligned}&\|Dg^{(n)} \cdot D\sigma^{(n-1)} \cdot Dg^{(n-1)} \cdot \ldots \cdot D\sigma^{(1)} \cdot Dg^{(1)}(x)\|_2 \\ &\leq L^{n-1}\|Dg^{(n)}\|_2 \ldots \|Dg^{(1)}(x)\|_2 \\ &= L^{n-1}\|W^{(n)}\|_2\|W^{(n-1)}\|_2 \ldots \|W^{(1)}(x)\|_2 \end{aligned} \tag{41}$$

This scales as a function of the norms of the weights of a layer. For β-LRP we see an upper bound which is insensitive to weight norms:

$$\begin{aligned}&\|Mg^{(n)} \cdot M\sigma^{(n-1)} \cdot Mg^{(n-1)} \cdot M\sigma^{(n-2)} \cdot Mg^{(n-2)} \cdot \\ &\ldots \cdot Mg^{(1)}(x)\| \\ &\leq \|Mg^{(n)}\|_2\|Mg^{(n-1)}\|_2 \ldots \|Mg^{(1)}(x)\|_2 \\ &\leq (1+\sqrt{2}\beta)^n \prod_l \sqrt{R_l} \end{aligned} \tag{42}$$

Discussion: There are two observations. Firstly, the independence of the singular values of the LRP-β transition matrices of neural network weights W shows a robustness property of LRP-β and corresponds to an interpretation of LRP-β attributions as an analogy of gradient clipping for modified gradients.

Secondly, equation (42) contains a term $\prod_l \sqrt{R_l}$ which depends on the output dimensions R_l of each layer. This is a typically large quantity. Therefore, this might be of lesser value for deriving concentration inequalities. Therefore, we devise an improved, tighter, bound in the next section, using a different approach.

11. Proof of Theorem 3

Proof:

We considering the LRP-β term

$$\begin{aligned}&Att(g_a^{(r)}, z_b^{(r-1)}) \\ &= (1+\beta)\frac{(w_{ab}z_b)_+}{\sum_{b'}(w_{ab'}z_{b'})_+} - \beta\frac{(w_{ab}z_b)_-}{\sum_{b'}(w_{ab'}z_{b'})_-} \end{aligned} \tag{43}$$

for a layer r computing $g^{(r)}(z) = Wz + c$. To simplify notation, we define

$$p_{ab+} := \frac{(w_{ab}z_b)_+}{\sum_{b'}(w_{ab'}z_{b'})_+} \tag{44}$$

$$p_{ab-} := \frac{(w_{ab}z_b)_-}{\sum_{b'}(w_{ab'}z_{b'})_-} \tag{45}$$

$$Att(g_a^{(r)}, z_b^{(r-1)}) = (1+\beta)p_{ab+} - \beta p_{ab-} \tag{46}$$

We note that $p_{ab+} \in [0,1]$, $p_{ab-} \in [0,1]$. One can observe that $w_{ab}z_b$ is either non-negative or non-positive. Therefore, one of the terms p_{ab+} and p_{ab-} must always be a zero term.

Induction start $t = 1$: We initially the computation of the attribution map at the network output across output components $g_1^{(n)}, \ldots, g_{D_n}^{(n)}$ at the last layer n using a vector q such that $\sum_{u=1}^{D_n} q_u = 1,\ q_u \geq 0$, that is we compute the attribution map for the weighted sum of outputs $\sum_{u=1}^{D_n} q_u g_u^{(n)}$.

The attribution map in the next upstream layer, for the component z_b of the feature map $z^{(n-1)}$, for which we have to prove the bounds, will be

$$\sum_{u=1}^{D_n} q_u Att(g_u^{(n)}, z_b^{(n-1)}) \tag{47}$$

Applying LRP-β to $g_u^{(n)}$ with the weights q_u results in

$$\sum_{u=1}^{D_n} q_u Att(g_u^{(n)}, z_b^{(n-1)}) = \sum_{u=1}^{D_n} q_u((1+\beta)p_{ub+} - \beta p_{ub-}) \tag{48}$$

Using the observation that one of p_{ub+}, p_{ub-} is always zero, we can write it as

$$= \sum_u q_u(\underbrace{(1+\beta)p_{ub+}}_{\geq 0}) + \sum_u q_u(\underbrace{-\beta p_{ub-}}_{\leq 0}) \tag{49}$$

Lets prove the upper bound for

$$\sum_{b:\sum_{u=1} q_u Att(g_u^{(n)}, z_b) > 0} \sum_{u=1}^{D_n} q_u Att(g_u^{(n)}, z_b^{(n-1)}) \tag{50}$$

For this we observe: if b satisfies $\sum_{u=1} q_u Att(g_u^{(n)}, z_b^{(n-1)}) > 0$, there must exist $u : p_{ub+} > 0$ in equation (49) due to $q_u \geq 0$.

Lets define the following true/false logical functions:
$Y_+(b) = \sum_{u=1} q_u Att(g_u^{(n)}, z_b^{(n-1)}) > 0$
$Y_-(b) = \sum_{u=1} q_u Att(g_u^{(n)}, z_b^{(n-1)}) < 0$
Therefore, using $\sum_{b:Y_+(b)}$ to denote those b for which the function evaluates to true:

$$\begin{aligned}
&\sum_{\{b:\sum_{u=1} q_u Att(g_u^{(n)}, z_b)>0\}} \sum_{u=1}^{D_n} q_u Att(g_u^{(n)}, z_b) = \\
&\sum_{b:Y_+(b)} \sum_{u=1}^{D_n} q_u Att(g_u^{(n)}, z_b) \\
&= \sum_{b:Y_+(b)} \sum_u \underbrace{q_u((1+\beta)p_{ub+})}_{\geq 0} + \sum_u \underbrace{q_u(-\beta p_{ub-})}_{\leq 0} \\
&\leq \sum_{b:Y_+(b)} \sum_u q_u((1+\beta)p_{ub+}) + 0 \\
&\leq \sum_b \sum_u q_u((1+\beta)p_{ub+} \\
&= \sum_u q_u(1+\beta) \sum_b p_{ub+} = \sum_u q_u(1+\beta) = (1+\beta)
\end{aligned} \tag{51}$$

For the lower bound we use an analogous argument:

$$\begin{aligned}
&\sum_{\{b:\sum_{u=1} q_u Att(g_u^{(n)}, z_b)<0\}} \sum_{u=1}^{D_n} q_u Att(g_u^{(n)}, z_b) = \\
&\sum_{b:Y_-(b)} \sum_{u=1}^{D_n} q_u Att(g_u^{(n)}, z_b) \\
&= \sum_{b:Y_-(b)} \sum_u \underbrace{q_u((1+\beta)p_{ub+})}_{\geq 0} + \sum_u \underbrace{q_u(-\beta p_{ub-})}_{\leq 0} \\
&\geq \sum_{b:Y_-(b)} 0 + \sum_u q_u(-\beta p_{ub-}) \\
&\geq \sum_b \sum_u q_u(-\beta p_{ub-}) \\
&= \sum_u q_u(-\beta) \sum_b p_{ub-} = \sum_u q_u(-\beta) = -\beta
\end{aligned} \tag{52}$$

Induction step $t-1 \to t$**:** We are given now attribution scores v_u such that

$$v_u = \sum_r q_r Att(g_r^{(n)}, z_u^{(n-(t-1))}) \tag{53}$$

These are the attribution scores for the feature map $z^{(n-(t-1))}$ of layer $n-(t-1)$ backpropagated from the weighted output of the network $\sum_r q_r g_r^{(n)}$.

v_u is the score for the u-th component $z_u^{(n-(t-1))}$ of vector $z^{(n-(t-1))}$.

We can assume that for these attribution scores v_u from the layer $z^{(n-(t-1))}$ downstream we have $\sum_u v_u = 1$, however, we can have now both signs for values v_u.

Note that the set of scores $\{v_u\}$ satisfies the induction assumption as stated above for layer $n-(t-1)$.

It should be noted here that due to equations (13) or (20) we have

$$\begin{aligned}
&\sum_r q_r Att(g_r^{(n)}, z_b^{((n-t))}) \\
&= q^\top M^\top(g^{(n)}) \cdot M^\top(g^{(n-1)}) \\
&\cdot \ldots \cdot M^\top(g^{(n-(t-1))})(z_b^{(n-t)}) \qquad (54) \\
&= (q^\top M^\top(g^{(n)}) \cdot M^\top(g^{(n-1)}) \\
&\cdot \ldots \cdot M^\top(g^{(n-t+2)})(z^{(n-(t-1))})) \cdot M^\top(g^{(n-(t-1))})(z_b^{(n-t)}) \qquad (55) \\
&= \sum_u \left(\sum_r q_r Att(g_r^{(n)}, z_u^{(n-(t-1))}) \right) Att(g_u^{(n-(t-1))}, z_b^{(n-t)}) \qquad (56) \\
&= \sum_u v_u Att(g_u^{(n-(t-1))}, z_b^{(n-t)}) \qquad (57) \\
&= v \cdot M^\top(g^{(n-(t-1))})(z_b^{(n-t)}) \qquad (58)
\end{aligned}$$

Therefore, as a consequence of equation (57), we need to obtain bounds for

$$\sum_{b:\sum_u v_u Att(g_u^{(n-(t-1))}, z_b^{(n-t)})>0} \sum_u v_u Att(g_u^{(n-(t-1))}, z_b^{(n-t)}) \tag{59}$$

$$\sum_{b:\sum_u v_u Att(g_u^{(n-(t-1))}, z_b^{(n-t)})<0} \sum_u v_u Att(g_u^{(n-(t-1))}, z_b^{(n-t)}) \tag{60}$$

Lets define the true/false-valued functions
$Y_+(b) = \sum_u v_u Att(g_u^{(n-(t-1))}, z_b^{(n-t)}) > 0,$
$Y_-(b) = \sum_u v_u Att(g_u^{(n-(t-1))}, z_b^{(n-t)}) < 0$
We will shorten $g_u^{(n-(t-1))}$ to g_u and $z_b^{(n-t)}$ to z_b further below.

Lets consider the upper bound first. For the upper bound

we can separate terms by their signs to obtain

$$\sum_{b:Y_+(b)} \sum_u v_u Att(g_u, z_b) \tag{61}$$

$$= \sum_{b:Y_+(b)} \sum_u v_u(1+\beta)p_{ub+} + \sum_u v_u(-\beta)p_{ub-} \tag{62}$$

$$= \sum_{b:Y_+(b)} \sum_{u:v_u>0} v_u(1+\beta)p_{ub+} + \sum_{u:v_u>0} v_u(-\beta)p_{ub-} \tag{63}$$

$$+ \sum_{b:Y_+(b)} \sum_{u:v_u<0} v_u(1+\beta)p_{ub+} + \sum_{u:v_u<0} v_u(-\beta)p_{ub-} \tag{64}$$

$$\leq \sum_{b:Y_+(b)} \sum_{u:v_u>0} v_u(1+\beta)p_{ub+} + \sum_{u:v_u>0} 0 \tag{65}$$

$$+ \sum_{b:Y_+(b)} \sum_{u:v_u<0} 0 + \sum_{u:v_u<0} v_u(-\beta)p_{ub-} \tag{66}$$

$$= \sum_{b:Y_+(b)} \sum_{u:v_u>0} v_u(1+\beta)p_{ub+} + \sum_{u:v_u<0} v_u(-\beta)p_{ub-} \tag{67}$$

The above inequality comes from checking the signs of the terms.

In the following $\sum_b f(b)$ denotes the sum over all b, while $\sum_{b:Y_+(b)} f(b)$ is the sum over the subset of input indices b for which $Y_+(b)$ evaluates to true.

All terms in the last statement are non-negative (note $p_{ub+} \in [0,1],\ p_{ub-} \in [0,1]$).

Therefore we can upper bound

$$\sum_{b:Y_+(b)} \sum_{u:v_u>0} v_u(1+\beta)p_{ub+} + \sum_{u:v_u<0} v_u(-\beta)p_{ub-} \tag{68}$$

$$\leq \sum_b \sum_{u:v_u>0} v_u(1+\beta)p_{ub+} + \sum_{u:v_u<0} v_u(-\beta)p_{ub-} \tag{69}$$

$$= \sum_{u:v_u>0} v_u(1+\beta) \sum_b p_{ub+} + \sum_{u:v_u<0} v_u(-\beta) \sum_b p_{ub-} \tag{70}$$

$$= \sum_{u:v_u>0} v_u(1+\beta) + \sum_{u:v_u<0} v_u(-\beta) \tag{71}$$

Now $\sum_{v_u>0} v_u$ are the positive scores from the next downstream layer $n-(t-1)$. They satisfy according to the induction assumption

$$\sum_{v_u>0} v_u \leq +2^{t-2}(1+\beta)^{t-1} \tag{72}$$

Analogously, $\sum_{v_u<0} v_u$ are the negative scores from the next downstream layer $n-(t-1)$. They satisfy according to the induction assumption

$$\sum_{v_u<0} v_u \geq -2^{t-2}\beta(1+\beta)^{t-2}$$

$$\Leftrightarrow \sum_{v_u<0} (-v_u) \leq +2^{t-2}\beta(1+\beta)^{t-2} \tag{73}$$

Plugging these inequalities in, results in

$$\sum_{v_u>0} v_u(1+\beta) + \sum_{v_u<0} (-v_u)\beta$$

$$\leq + 2^{t-2}(1+\beta)^{t-1}(1+\beta) + 2^{t-2}\beta^2(1+\beta)^{t-2}$$

$$\leq + 2^{t-2}(1+\beta)^t + 2^{t-2}(1+\beta)^t \tag{74}$$

$$= 2^{t-1}(1+\beta)^t \tag{75}$$

For the lower bound we can use an analogous reasoning: We can look at the signs to obtain

$$\sum_{b:Y_-(b)} \sum_u v_u Att(g_u, z_b) \tag{76}$$

$$= \sum_{b:Y_-(b)} \sum_u v_u(1+\beta)p_{ub+} + \sum_u v_u(-\beta)p_{ub-} \tag{77}$$

$$= \sum_{b:Y_-(b)} \sum_{u:v_u>0} v_u(1+\beta)p_{ub+} + \sum_{u:v_u>0} v_u(-\beta)p_{ub-} \tag{78}$$

$$+ \sum_{b:Y_-(b)} \sum_{u:v_u<0} v_u(1+\beta)p_{ub+} + \sum_{u:v_u<0} v_u(-\beta)p_{ub-} \tag{79}$$

$$\geq \sum_{b:Y_-(b)} \sum_{u:v_u>0} 0 + \sum_{u:v_u>0} v_u(-\beta)p_{ub-} \tag{80}$$

$$+ \sum_{b:Y_-(b)} \sum_{u:v_u<0} v_u(1+\beta)p_{ub+} + \sum_{u:v_u<0} 0 \tag{81}$$

$$= \sum_{b:Y_-(b)} \sum_{u:v_u>0} v_u(-\beta)p_{ub-} + \sum_{u:v_u<0} v_u(1+\beta)p_{ub+} \tag{82}$$

All terms are non-positive (note $p_{ub+} \in [0,1],\ p_{ub-} \in [0,1]$).

Therefore we can lower bound

$$\sum_{b:Y_-(b)} \sum_{u:v_u>0} v_u(-\beta)p_{ub-} + \sum_{u:v_u<0} v_u(1+\beta)p_{ub+} \tag{83}$$

$$\geq \sum_b \sum_{u:v_u>0} v_u(-\beta)p_{ub-} + \sum_{u:v_u<0} v_u(1+\beta)p_{ub+} \tag{84}$$

$$= \sum_{u:v_u>0} v_u(-\beta) \sum_b p_{ub-} + \sum_{u:v_u<0} v_u(1+\beta) \sum_b p_{ub+} \tag{85}$$

$$= \sum_{u:v_u>0} v_u(-\beta) + \sum_{u:v_u<0} v_u(1+\beta) \tag{86}$$

By induction assumption we have bounds as follows:

$$\sum_{v_u>0} v_u \leq 2^{t-2}(1+\beta)^{t-1} \tag{87}$$

$$\sum_{v_u<0} v_u \geq -2^{t-2}\beta(1+\beta)^{t-2} \tag{88}$$

Plugging them in yields:

$$\sum_{v_u>0} v_u(-\beta) + \sum_{v_u<0} v_u(1+\beta) \tag{89}$$

$$\geq 2^{t-2}(1+\beta)^{t-1}(-\beta) - 2^{t-2}\beta(1+\beta)^{t-2}(1+\beta) \tag{90}$$

$$= -2^{t-2}\beta(1+\beta)^{t-1} - 2^{t-2}\beta(1+\beta)^{t-1} \tag{91}$$

$$= -2^{t-1}\beta(1+\beta)^{t-1} \tag{92}$$

This concludes the upper and the lower bound in the induction step.

12. Proof of Theorem 4

We considering the LRP-γ term

$$Att(g_a, z_b) = \frac{w_{ab}z_b + \gamma(w_{ab}z_b)_+}{\sum_{b'} w_{ab'}z_{b'} + \gamma(w_{ab}z_{b'})_+} \tag{93}$$

Let us omit layer indices again, and define as notation

$$y_{ab} = w_{ab}z_b \tag{94}$$

Induction start $t = 1$: We initially the computation of the attribution map at the network output across output components $g_1^{(n)}, \ldots, g_{D_n}^{(n)}$ at the last layer n using a vector q such that $\sum_{u=1}^{D_n} q_u = 1,\ q_u \geq 0$, that is we compute the attribution map for the weighted sum of outputs $\sum_{u=1}^{D_n} q_u g_u^{(n)}$.

The attribution map in the next upstream layer, for which we have to prove the bounds, will be

$$\sum_{u=1}^{D_n} q_u Att(g_u^{(n)}, z_b) \tag{95}$$

Applying LRP-γ to $g_u^{(n)}$ with weights q_u results in

$$\sum_{u=1}^{D_n} q_u Att(g_u, z_b) \tag{96}$$

$$= \sum_{u=1}^{D_n} q_u \frac{y_{ub} + \gamma(y_{ub})_+}{\sum_{b'} y_{ub'} + \gamma(y_{ub'})_+} \tag{97}$$

$$= \sum_{u=1}^{D_n} q_u \frac{\gamma^{-1}y_{ub} + (y_{ub})_+}{\sum_{b'} \gamma^{-1}y_{ub'} + (y_{ub'})_+} \tag{98}$$

$$= \sum_{u:y_{ub}>0} q_u \frac{\gamma^{-1}y_{ub} + (y_{ub})_+}{\sum_{b'} \gamma^{-1}y_{ub'} + (y_{ub'})_+} + \sum_{u:y_{ub}<0} q_u \frac{\gamma^{-1}y_{ub} + (y_{ub})_+}{\sum_{b'} \gamma^{-1}y_{ub'} + (y_{ub'})_+} \tag{99}$$

$$= \sum_{u:y_{ub}>0} q_u \frac{(1+\gamma^{-1})(y_{ub})_+}{\sum_{b'} \gamma^{-1}y_{ub'} + (y_{ub'})_+} + \sum_{u:y_{ub}<0} q_u \frac{\gamma^{-1}y_{ub}}{\sum_{b'} \gamma^{-1}y_{ub'} + (y_{ub'})_+} \tag{100}$$

If $\sum_b (y_{ub})_+ = 0$ it means that all $y_{ub} < 0$, then we get:

$$\sum_{b:Y_+(b)} \sum_{u=1}^{D_n} q_u Att(g_u, z_b)$$

$$= 0 + \sum_{b:Y_+(b)} \sum_{u:y_{ub}<0} q_u \frac{\gamma^{-1} y_{ub}}{\sum_{b'} \gamma^{-1} y_{ub'} + 0} \tag{101}$$

$$= \sum_{b:Y_+(b)} \sum_{u:y_{ub}<0} q_u \frac{y_{ub}}{\sum_{b'} y_{ub'}}$$

$$= \sum_{b:Y_+(b)} \sum_{u:y_{ub}<0} q_u \frac{(y_{ub})_-}{\sum_{b'} (y_{ub'})_-} \tag{102}$$

$$\leq \sum_b \sum_u q_u \frac{y_{ub}}{\sum_{b'} y_{ub'}} \tag{103}$$

$$= \sum_{u:} q_u \sum_b \frac{y_{ub}}{\sum_{b'} y_{ub'}} = \sum_u q_u = 1 \tag{104}$$

If $\sum_b (y_{ub})_+ > 0$, then we require by the assumption of the lemma (in the lemma we have set $\alpha = \gamma^{-1/2}$ as seen further below, while here we execute it for a general $\alpha \in (0, 1)$)

$$\gamma^{-1} \sum_{b:y_{ub}<0} y_{ub} > -\alpha \sum_{b:y_{ub}>0} (y_{ub})_+ \tag{105}$$

$$\Leftrightarrow \sum_{b:y_{ub}<0} \gamma^{-1} y_{ub} + \sum_{b:y_{ub}>0} (1+\gamma^{-1})(y_{ub})_+$$

$$> -\alpha \sum_{b:y_{ub}>0} (y_{ub})_+ + (1+\gamma^{-1}) \sum_{b:y_{ub}>0} (y_{ub})_+ \tag{106}$$

$$\Leftrightarrow \sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+$$

$$> (1+\gamma^{-1}-\alpha) \sum_{b'} (y_{ub'})_+ \tag{107}$$

The left hand side holds due to

$$\sum_{b:y_{ub}<0} \gamma^{-1} y_{ub} + \sum_{b:y_{ub}>0} (1+\gamma^{-1})(y_{ub})_+$$

$$= \sum_{b:y_{ub}<0} \gamma^{-1} y_{ub} + (y_{ub})_+ + \sum_{b:y_{ub}>0} \gamma^{-1} y_{ub} + (y_{ub})_+$$

$$= \sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+ \tag{108}$$

Then, an upper bound will be

$$\sum_{b:Y_+(b)} \sum_{u=1}^{D_n} q_u Att(g_u, z_b) \tag{109}$$

$$\leq \sum_{b:Y_+(b)} \sum_{u:y_{ub}>0} q_u \frac{(1+\gamma^{-1})(y_{ub})_+}{\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+} \tag{110}$$

$$\leq \sum_{b:Y_+(b)} \sum_{u:y_{ub}>0} q_u \frac{(1+\gamma^{-1})(y_{ub})_+}{(1+\gamma^{-1}-\alpha) \sum_{b'} (y_{ub'})_+} \tag{111}$$

$$= \frac{1+\gamma^{-1}}{1+\gamma^{-1}-\alpha} \sum_{b:Y_+(b)} \sum_{u:y_{ub}>0} q_u \frac{(y_{ub})_+}{\sum_{b'} (y_{ub'})_+} \tag{112}$$

Next we use the trick, that we can drop the conditioning on $u : y_{ub} > 0$ because the terms in the upper bound would be simply zero if $y_{ub} < 0$. After that we can sum over all input dimensions b because all terms have the same positive sign or are zero.

$$\leq \frac{1+\gamma^{-1}}{1+\gamma^{-1}-\alpha} \sum_{b:Y_+(b)} \sum_u q_u \frac{(y_{ub})_+}{\sum_{b'} (y_{ub'})_+}$$

$$\leq \frac{1+\gamma^{-1}}{1+\gamma^{-1}-\alpha} \sum_b \sum_u q_u \frac{(y_{ub})_+}{\sum_{b'} (y_{ub'})_+} \tag{113}$$

$$= \frac{1+\gamma^{-1}}{1+\gamma^{-1}-\alpha} \sum_u q_u \frac{\sum_b (y_{ub})_+}{\sum_{b'} (y_{ub'})_+} \tag{114}$$

$$= \frac{1+\gamma^{-1}}{1+\gamma^{-1}-\alpha} \sum_u q_u \tag{115}$$

$$= \frac{1+\gamma^{-1}}{1+\gamma^{-1}-\alpha} \tag{116}$$

This proves an upper bound in the induction step of $\frac{1+\gamma^{-1}}{1+\gamma^{-1}-\alpha}$.

For a lower bound we use

$$\gamma^{-1} \sum_{b:y_{ub}<0} y_{ub} > -\alpha \sum_{b:y_{ub}>0} (y_{ub})_+ \tag{117}$$

$$\Leftrightarrow 0 \leq -\gamma^{-1} \alpha^{-1} \sum_{b:y_{ub}<0} y_{ub} < \sum_{b:y_{ub}>0} (y_{ub})_+$$

$$= \sum_{b'} (y_{ub'})_+ \tag{118}$$

so that

$$\sum_{b:Y_-(b)} \sum_{u=1}^{D_n} q_u Att(g_u, z_b) \tag{119}$$

$$\geq 0 + \sum_{u:y_{ub}<0} q_u \frac{\gamma^{-1} y_{ub}}{\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+} \tag{120}$$

$$\geq \sum_{b:Y_-(b)} \sum_{u:y_{ub}<0} q_u \frac{\gamma^{-1} y_{ub}}{\sum_{b':y_{ub'}<0} \gamma^{-1}(1-\alpha^{-1}) y_{ub'}} \tag{121}$$

$$=(1-\alpha^{-1})^{-1} \sum_{b:Y_-(b)} \sum_{u:y_{ub}<0} q_u \frac{y_{ub}}{\sum_{b':y_{ub'}<0} y_{ub'}} \tag{122}$$

$$=(1-\alpha^{-1})^{-1} \sum_{b:Y_-(b)} \sum_{u:y_{ub}<0} q_u \frac{(y_{ub})_-}{\sum_{b':y_{ub'}<0} (y_{ub'})_-} \tag{123}$$

$$=(1-\alpha^{-1})^{-1} \sum_{b:Y_-(b)} \sum_{u} q_u \frac{(y_{ub})_-}{\sum_{b'} (y_{ub'})_-} \tag{124}$$

$$\geq (1-\alpha^{-1})^{-1} \sum_{b} \sum_{u} q_u \frac{(y_{ub})_-}{\sum_{b'} (y_{ub'})_-} \tag{125}$$

$$=(1-\alpha^{-1})^{-1} \sum_{u} q_u = \frac{\alpha}{\alpha - 1} \tag{126}$$

We obtain
for the sum of positive attributions

$$\sum_{b:Y_+(b)} \sum_{u=1}^{D_n} q_u Att(g_u^{(n)}, z_b^{(n-1)}) \leq \frac{1+\gamma^{-1}}{1+\gamma^{-1}-\alpha} \tag{127}$$

for the sum of negative attributions as

$$\sum_{b:Y_-(b)} \sum_{u=1}^{D_n} q_u Att(g_u^{(n)}, z_b^{(n-1)}) \geq \frac{\alpha}{\alpha - 1} \tag{128}$$

To simplify, set $\alpha := \gamma^{-1/2}$, resulting in a requirement of

$$\gamma^{-1/2} \sum_{b:y_{ub}<0} (-1) y_{ub} < \sum_{b:y_{ub}>0} (y_{ub})_+ \tag{129}$$

then we get for the sum of positive attributions

$$\leq \frac{1+\gamma^{-1}}{1+\gamma^{-1}-\gamma^{-1/2}} = \frac{1+\gamma}{1+\gamma-\gamma^{1/2}} \tag{130}$$

for the sum of negative attributions as

$$\geq \frac{1}{1-\gamma^{1/2}} \tag{131}$$

Induction step $t-1 \to t$**:** To start with, by our assumption of the lemma, we have set γ such that for all activations $y_{ub} = w_{ub} z_b$ we have

$$\gamma^{-1} \sum_{b:y_{ub}<0} y_{ub} > -\gamma^{-1/2} \sum_{b:y_{ub}>0} (y_{ub})_+ \tag{132}$$

We are given now attribution scores v_u such that

$$v_u = \sum_r q_r Att(g_r^{(n)}, z_u^{(n-(t-1))}) \tag{133}$$

These are the attribution scores for the feature map $z^{(n-(t-1))}$ of layer $n-(t-1)$ backpropagated from the weighted output of the network $\sum_r q_r g_r^{(n)}$.

Note that the set of scores $\{v_u\}$ satisfies the induction assumption as stated above for layer $n-(t-1)$, that is

$$\sum_{u:v_u<0} v_u \geq 2^{t-2} \frac{1}{1-\gamma^{1/2}} b(\gamma)^{t-2} \tag{134}$$

$$\sum_{u:v_u>0} v_u \leq 2^{t-2} \frac{1+\gamma}{1+\gamma-\gamma^{1/2}} b(\gamma)^{t-2} \tag{135}$$

Take note that for $\gamma > 1$: $\frac{1}{1-\gamma^{1/2}} < 0$

It should be noted here that due to equations (13) or (20) we have

$$\sum_r q_r Att(g_r^{(n)}, z_b^{(n-t)}) = \tag{136}$$

$$q^\top M^\top(g^{(n)}) \cdot M^\top(g^{(n-1)}) \tag{137}$$

$$\cdot \ldots \cdot M^\top(g^{(n-(t-1))})(z_b^{(n-t)}) \tag{138}$$

$$= \Big(q^\top M^\top(g^{(n)}) \cdot M^\top(g^{(n-1)}) \cdot \tag{139}$$

$$\ldots \cdot M^\top(g^{(n-t+2)})(z^{(n-(t-1))}) \Big) \tag{140}$$

$$\cdot M^\top(g^{(n-(t-1))})(z_b^{(n-t)}) \tag{141}$$

$$= \sum_u \left(\sum_r q_r Att(g_r^{(n)}, z_u^{(n-(t-1))}) \right) Att(g_u^{(n-(t-1))}, z_b^{(n-t)}) \tag{142}$$

$$= \sum_u v_u Att(g_u^{(n-(t-1))}, z_b^{(n-t)}) \tag{143}$$

$$= v \cdot M^\top(g^{(n-(t-1))})(z_b^{(n-t)}) \tag{144}$$

Therefore, as a consequence of equation (143), we need to obtain bounds for

$$\sum_{b:\sum_u v_u Att(g_u^{(n-(t-1))}, z_b^{(n-t)})>0} \sum_u v_u Att(g_u^{(n-(t-1))}, z_b^{(n-t)}) \tag{145}$$

$$\sum_{b:\sum_u v_u Att(g_u^{(n-(t-1))}, z_b^{(n-t)})<0} \sum_u v_u Att(g_u^{(n-(t-1))}, z_b^{(n-t)}) \tag{146}$$

Lets define the true/false-valued functions
$Y_+(b) = \sum_u v_u Att(g_u^{(n-(t-1))}, z_b^{(n-t)}) > 0,$
$Y_-(b) = \sum_u v_u Att(g_u^{(n-(t-1))}, z_b^{(n-t)}) < 0$

We will shorten $g_u^{(n-(t-1))}$ to g_u and $z_b^{(n-t)}$ to z_b further below.

Let $b(\gamma) := \max(-\frac{1}{1-\gamma^{1/2}}, \frac{1+\gamma}{1+\gamma-\gamma^{1/2}})$

Applying LRP-γ to g_u with weights v_u results in

$$\sum_{u=1} v_u Att(g_u, z_b) = \sum_{u=1} v_u \frac{y_{ub} + \gamma(y_{ub})_+}{\sum_{b'} y_{ub'} + \gamma(y_{ub'})_+} \tag{147}$$

$$= \sum_{u=1} v_u \frac{\gamma^{-1} y_{ub} + (y_{ub})_+}{\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+} \tag{148}$$

$$= \sum_{u:y_{ub}>0} v_u \frac{\gamma^{-1} y_{ub} + (y_{ub})_+}{\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+} \tag{149}$$

$$= \sum_{u:y_{ub}>0} v_u \frac{(1+\gamma^{-1})(y_{ub})_+}{\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+} + \sum_{u:y_{ub}<0} v_u \frac{\gamma^{-1} y_{ub}}{\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+} \tag{150}$$

Now we have to split this further according to signs of v_u:

$$= \sum_{u:v_u>0,y_{ub}>0} v_u \frac{(1+\gamma^{-1})(y_{ub})_+}{\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+} \tag{151}$$

$$+ \sum_{u:v_u>0,y_{ub}<0} v_u \frac{\gamma^{-1} y_{ub}}{\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+} \tag{152}$$

$$+ \sum_{u:v_u<0,y_{ub}>0} v_u \frac{(1+\gamma^{-1})(y_{ub})_+}{\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+} \tag{153}$$

$$+ \sum_{u:v_u<0,y_{ub}<0} v_u \frac{\gamma^{-1} y_{ub}}{\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+} \tag{154}$$

For the upper bound we can derive from that:

$$\sum_{b:Y_+(b)} \sum_{u=1} v_u Att(g_u, z_b)$$

$$= \sum_{b:Y_+(b)} \sum_{u:v_u>0,y_{ub}>0} v_u \frac{(1+\gamma^{-1})(y_{ub})_+}{\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+}$$

$$+ \sum_{u:v_u>0,y_{ub}<0} v_u \frac{\gamma^{-1} y_{ub}}{\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+}$$

$$+ \sum_{b:Y_+(b)} \sum_{u:v_u<0,y_{ub}>0} v_u \frac{(1+\gamma^{-1})(y_{ub})_+}{\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+}$$

$$+ \sum_{u:v_u<0,y_{ub}<0} v_u \frac{\gamma^{-1} y_{ub}}{\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+} \tag{155}$$

$$\leq \sum_{b:Y_+(b)} \sum_{u:v_u>0,y_{ub}>0} v_u \frac{(1+\gamma^{-1})(y_{ub})_+}{\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+} + 0 \tag{156}$$

$$+ \sum_{b:Y_+(b)} 0 + \sum_{u:v_u<0,y_{ub}<0} v_u \frac{\gamma^{-1} y_{ub}}{\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+} \tag{157}$$

For the upper line eq. 156 we will use from our requirement

$$-\gamma^{-1/2} \sum_{b:y_{ub}<0} y_{ub} < \sum_{b:y_{ub}>0} (y_{ub})_+ \tag{158}$$

$$\Leftrightarrow \gamma^{-1} \sum_{b:y_{ub}<0} y_{ub} > -\gamma^{-1/2} \sum_{b:y_{ub}>0} (y_{ub})_+ \tag{159}$$

and therefore:

$$\Leftrightarrow \sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+ = \tag{160}$$

$$\sum_{b':y_{ub'}<0} \gamma^{-1} y_{ub'} + (y_{ub'})_+ + \sum_{b':y_{ub'}>0} \gamma^{-1} y_{ub'} + (y_{ub'})_+ \tag{161}$$

$$= \sum_{b':y_{ub'}<0} \gamma^{-1} y_{ub'} + 0 + \sum_{b':y_{ub'}>0} \gamma^{-1} (y_{ub'})_+ + (y_{ub'})_+ \tag{162}$$

$$= \sum_{b':y_{ub'}<0} \gamma^{-1} y_{ub'} + (1+\gamma^{-1}) \sum_{b':y_{ub'}>0} (y_{ub'})_+ \tag{163}$$

$$> -\gamma^{-1/2} \sum_{b:y_{ub}>0} (y_{ub})_+ + (1+\gamma^{-1}) \sum_{b':y_{ub'}>0} (y_{ub'})_+ \tag{164}$$

$$= (1+\gamma^{-1}-\gamma^{-1/2}) \sum_{b':y_{ub'}>0} (y_{ub'})_+ \tag{165}$$

$$\Leftrightarrow \sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+ > (1+\gamma^{-1}-\gamma^{-1/2}) \sum_{b'} (y_{ub'})_+ \tag{166}$$

For the lower line eq. 157, we employ

$$\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+ \tag{167}$$

$$= \sum_{y_{ub'}<0} \gamma^{-1} y_{ub'} + (y_{ub'})_+ + \sum_{y_{ub'}>0} \gamma^{-1} y_{ub'} + (y_{ub'})_+ \tag{168}$$

$$= \sum_{y_{ub'}<0} \gamma^{-1} y_{ub'} + 0 + \sum_{y_{ub'}>0} (1+\gamma^{-1})(y_{ub'})_+ \tag{169}$$

$$= \gamma^{-1} \sum_{y_{ub'}<0} y_{ub'} + (1+\gamma^{-1}) \sum_{y_{ub'}>0} (y_{ub'})_+ \tag{170}$$

$$> \gamma^{-1} \sum_{y_{ub'}<0} y_{ub'} + (1+\gamma^{-1})(-1)\gamma^{-1/2} \sum_{y_{ub'}<0} y_{ub'} \tag{171}$$

$$= (\gamma^{-1} - (1+\gamma^{-1})\gamma^{-1/2}) \sum_{y_{ub'}<0} y_{ub'} \tag{172}$$

Also note that all terms are non-negative, so that replacing positive terms in the divisor by smaller positive ones yields an upper bound. We obtain:

$$\sum_{b:Y_+(b)} \sum_{u=1} v_u Att(g_u, z_b) \tag{173}$$

$$\leq \sum_{b:Y_+(b)} \sum_{u:v_u>0, y_{ub}>0} v_u \frac{(1+\gamma^{-1})(y_{ub})_+}{\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+} + 0 \tag{174}$$

$$+ \sum_{b:Y_+(b)} 0 + \sum_{u:v_u<0, y_{ub}<0} v_u \frac{\gamma^{-1} y_{ub}}{\sum_{b'} \gamma^{-1} y_{ub'} + (y_{ub'})_+} \tag{175}$$

$$\leq \sum_{b:Y_+(b)} \sum_{u:v_u>0, y_{ub}>0} v_u \cdot \tag{176}$$

$$\cdot \frac{(1+\gamma^{-1})(y_{ub})_+}{(1+\gamma^{-1}-\gamma^{-1/2}) \sum_{b'} (y_{ub'})_+} \tag{177}$$

$$+ \sum_{b:Y_+(b)} \sum_{u:v_u<0, y_{ub}<0} v_u \cdot \tag{178}$$

$$\cdot \frac{\gamma^{-1} y_{ub}}{(\gamma^{-1} - (1+\gamma^{-1})\gamma^{-1/2}) \sum_{y_{ub'}<0} y_{ub'}} \tag{179}$$

$$= \sum_{b:Y_+(b)} \sum_{u:v_u>0, y_{ub}>0} v_u \frac{(1+\gamma^{-1})(y_{ub})_+}{(1+\gamma^{-1}-\gamma^{-1/2}) \sum_{b'} (y_{ub'})_+} \tag{180}$$

$$+ \sum_{b:Y_+(b)} \sum_{u:v_u<0, y_{ub}<0} v_u \cdot \tag{181}$$

$$\cdot \frac{\gamma^{-1} (y_{ub})_-}{(\gamma^{-1} - (1+\gamma^{-1})\gamma^{-1/2}) \sum_{y_{ub'}<0} (y_{ub'})_-} \tag{182}$$

Next we use the trick that for $y_{ub} < 0$ we have $y_{ub} = (y_{ub})_-$, however terms $(y_{ub})_-$ can be summed over all b because for those where $y_{ub} > 0$ it would be just zero: $(y_{ub})_- = 0$.

The same idea holds for $y_{ub} > 0$ and $(y_{ub})_+$.

Therefore we can replace $\sum_{u:v_u<0, y_{ub}<0}$ by $\sum_{u:v_u<0}$ and $\sum_{u:v_u>0, y_{ub}>0}$ by $\sum_{u:v_u>0}$:

$$= \sum_{b:Y_+(b)} \sum_{u:v_u>0} v_u \frac{(1+\gamma^{-1})(y_{ub})_+}{(1+\gamma^{-1}-\gamma^{-1/2}) \sum_{b'} (y_{ub'})_+} \tag{183}$$

$$+ \sum_{b:Y_+(b)} \sum_{u:v_u<0} v_u \frac{\gamma^{-1} (y_{ub})_-}{(\gamma^{-1} - (1+\gamma^{-1})\gamma^{-1/2}) \sum_{b'} (y_{ub'})_-} \tag{184}$$

Now all terms are non-negative [note that $\gamma > 1$, so $1 - \gamma^{-1/2} > 0$ and that

$(\gamma^{-1} - (1+\gamma^{-1})\gamma^{-1/2}) = \gamma^{-1}(1 - (1+\gamma^{-1})\gamma^{1/2}) < 0$ which multiplies with $v_u < 0$ to something non-negative]

so that we can upper bound by increasing the sum from $\sum_{b:Y_+(b)}$ to $\sum_b$:

$$\leq \sum_{b} \sum_{u:v_u>0} v_u \frac{(1+\gamma^{-1})(y_{ub})_+}{(1+\gamma^{-1}-\gamma^{-1/2}) \sum_{b'} (y_{ub'})_+} \tag{185}$$

$$+ \sum_{b} \sum_{u:v_u<0} v_u \frac{\gamma^{-1} (y_{ub})_-}{(\gamma^{-1} - (1+\gamma^{-1})\gamma^{-1/2}) \sum_{b'} (y_{ub'})_-} \tag{186}$$

$$= \sum_{u:v_u>0} v_u \frac{1+\gamma^{-1}}{1+\gamma^{-1}-\gamma^{-1/2}} \sum_{b} \frac{(y_{ub})_+}{\sum_{b'} (y_{ub'})_+} \tag{187}$$

$$+ \sum_{u:v_u<0} v_u \frac{\gamma^{-1}}{(\gamma^{-1} - (1+\gamma^{-1})\gamma^{-1/2})} \sum_{b} \frac{(y_{ub})_-}{\sum_{b'} (y_{ub'})_-} \tag{188}$$

$$= \sum_{u:v_u>0} v_u \frac{1+\gamma^{-1}}{1+\gamma^{-1}-\gamma^{-1/2}} \tag{189}$$

$$+ \sum_{u:v_u<0} v_u \frac{\gamma^{-1}}{(\gamma^{-1} - (1+\gamma^{-1})\gamma^{-1/2})} \tag{190}$$

$$= \sum_{u:v_u>0} v_u \frac{1+\gamma}{1+\gamma-\gamma^{+1/2}} \tag{191}$$

$$+ \sum_{u:v_u<0} v_u \underbrace{\frac{1}{(1-(1+\gamma^{-1})\gamma^{+1/2})}}_{<0} \tag{192}$$

Now we can plug in the induction assumption

$$\leq 2^{t-2}\frac{1+\gamma}{1+\gamma-\gamma^{1/2}}b(\gamma)^{t-2}\frac{1+\gamma}{1+\gamma-\gamma^{1/2}} \tag{193}$$
$$+2^{t-2}\frac{1}{1-\gamma^{1/2}}b(\gamma)^{t-2}\frac{1}{(1-(1+\gamma^{-1})\gamma^{1/2})} \tag{194}$$
$$=2^{t-2}\frac{1+\gamma}{1+\gamma-\gamma^{1/2}}b(\gamma)^{t-2}\frac{1+\gamma}{1+\gamma-\gamma^{1/2}} \tag{195}$$
$$+2^{t-2}\frac{1}{\gamma^{1/2}-1}b(\gamma)^{t-2}\frac{-1}{(1-(1+\gamma^{-1})\gamma^{1/2})} \tag{196}$$
$$\tag{197}$$

Finally note

$$\frac{-1}{(1-(1+\gamma^{-1})\gamma^{1/2})} \tag{198}$$
$$=-\frac{\gamma^{1/2}}{(\gamma^{1/2}-(1+\gamma^{-1})\gamma)}=-\frac{\gamma^{1/2}}{(\gamma^{1/2}-(1+\gamma))} \tag{199}$$
$$=\frac{\gamma^{1/2}}{1+\gamma-\gamma^{1/2}}\leq\frac{1+\gamma}{1+\gamma-\gamma^{1/2}} \tag{200}$$

Therefore:

$$\sum_{b:Y_+(b)}\sum_{u=1}v_u Att(g_u,z_b) \tag{201}$$
$$\leq 2^{t-2}\frac{1+\gamma}{1+\gamma-\gamma^{1/2}}b(\gamma)^{t-2}\frac{1+\gamma}{1+\gamma-\gamma^{1/2}} \tag{202}$$
$$+2^{t-2}\frac{1}{\gamma^{1/2}-1}b(\gamma)^{t-2}\frac{-1}{(1-(1+\gamma^{-1})\gamma^{1/2})} \tag{203}$$
$$\leq 2^{t-2}\frac{1+\gamma}{1+\gamma-\gamma^{1/2}}b(\gamma)^{t-2}b(\gamma) \tag{204}$$
$$+2^{t-2}b(\gamma)b(\gamma)^{t-2}\frac{1+\gamma}{1+\gamma-\gamma^{1/2}} \tag{205}$$
$$=2^{t-1}\frac{1+\gamma}{1+\gamma-\gamma^{1/2}}b(\gamma)^{t-1} \tag{206}$$

which proves the induction claim for the positive upper bound.

For the lower bound we can derive in similar spirit:

$$\sum_{b:Y_-(b)}\sum_{u=1}v_u Att(g_u,z_b) \tag{207}$$
$$=\sum_{b:Y_-(b)}\sum_{u:v_u>0,y_{ub}>0}v_u\frac{(1+\gamma^{-1})(y_{ub})_+}{\sum_{b'}\gamma^{-1}y_{ub'}+(y_{ub'})_+} \tag{208}$$
$$+\sum_{u:v_u>0,y_{ub}<0}v_u\frac{\gamma^{-1}y_{ub}}{\sum_{b'}\gamma^{-1}y_{ub'}+(y_{ub'})_+} \tag{209}$$
$$\tag{210}$$
$$+\sum_{b:Y_-(b)}\sum_{u:v_u<0,y_{ub}>0}v_u\frac{(1+\gamma^{-1})(y_{ub})_+}{\sum_{b'}\gamma^{-1}y_{ub'}+(y_{ub'})_+} \tag{211}$$
$$+\sum_{u:v_u<0,y_{ub}<0}v_u\frac{\gamma^{-1}y_{ub}}{\sum_{b'}\gamma^{-1}y_{ub'}+(y_{ub'})_+} \tag{212}$$
$$\geq\sum_{b:Y_-(b)}0+\sum_{u:v_u>0,y_{ub}<0}v_u\frac{\gamma^{-1}y_{ub}}{\sum_{b'}\gamma^{-1}y_{ub'}+(y_{ub'})_+} \tag{213}$$
$$+\sum_{b:Y_-(b)}\sum_{u:v_u<0,y_{ub}>0}v_u\frac{(1+\gamma^{-1})(y_{ub})_+}{\sum_{b'}\gamma^{-1}y_{ub'}+(y_{ub'})_+}+0 \tag{214}$$

We will use two inequalities derived in equations (166) and (172).

For the upper line (213) we will use

$$\sum_{b'}\gamma^{-1}y_{ub'}+(y_{ub'})_+>(\gamma^{-1}-(1+\gamma^{-1})\gamma^{-1/2})\sum_{y_{ub'}<0}y_{ub'} \tag{215}$$

which works because in (213) we have $v_u > 0$ and $\gamma^{-1}y_{ub}<0$.

For the lower line (214) we will use

$$\sum_{b'}\gamma^{-1}y_{ub'}+(y_{ub'})_+>(1+\gamma^{-1}-\gamma^{-1/2})\sum_{b'}(y_{ub'})_+ \tag{216}$$

which works because in (214) we have $v_u < 0$ and $(1+\gamma^{-1})(y_{ub})_+ > 0$.

Note that the terms in (213) and (214) are non-positive, so that replacing positive terms in the divisor by smaller

positive ones yields a lower bound. Therefore:

$$\sum_{b:Y_-(b)}\sum_{u=1} v_u Att(g_u, z_b) \tag{217}$$

$$\geq \sum_{b:Y_-(b)}\sum_{u:v_u>0,y_{ub}<0} v_u \frac{\gamma^{-1}y_{ub}}{\sum_{b'}\gamma^{-1}y_{ub'} + (y_{ub'})_+} \tag{218}$$

$$+ \sum_{b:Y_-(b)}\sum_{u:v_u<0,y_{ub}>0} v_u \frac{(1+\gamma^{-1})(y_{ub})_+}{\sum_{b'}\gamma^{-1}y_{ub'} + (y_{ub'})_+} \tag{219}$$

$$\geq \sum_{b:Y_-(b)}\sum_{u:v_u>0,y_{ub}<0} v_u \cdot \tag{220}$$

$$\cdot \frac{\gamma^{-1}y_{ub}}{(\gamma^{-1}-(1+\gamma^{-1})\gamma^{-1/2})\sum_{y_{ub'}<0} y_{ub'}} \tag{221}$$

$$+ \sum_{b:Y_-(b)}\sum_{u:v_u<0,y_{ub}>0} v_u \cdot \tag{222}$$

$$\cdot \frac{(1+\gamma^{-1})(y_{ub})_+}{(1+\gamma^{-1}-\gamma^{-1/2})\sum_{b'}(y_{ub'})_+} \tag{223}$$

$$= \frac{\gamma^{-1}}{(\gamma^{-1}-(1+\gamma^{-1})\gamma^{-1/2})} \sum_{b:Y_-(b)}\sum_{u:v_u>0,y_{ub}<0} v_u \cdot \tag{224}$$

$$\cdot \frac{y_{ub}}{\sum_{y_{ub'}<0} y_{ub'}} \tag{225}$$

$$+ \frac{1+\gamma^{-1}}{1+\gamma^{-1}-\gamma^{-1/2}} \sum_{b:Y_-(b)}\sum_{u:v_u<0,y_{ub}>0} v_u \frac{(y_{ub})_+}{\sum_{b'}(y_{ub'})_+} \tag{226}$$

$$= \frac{\gamma^{-1}}{(\gamma^{-1}-(1+\gamma^{-1})\gamma^{-1/2})} \sum_{b:Y_-(b)}\sum_{u:v_u>0,y_{ub}<0} v_u \cdot \tag{227}$$

$$\cdot \frac{(y_{ub})_-}{\sum_{y_{ub'}<0}(y_{ub'})_-} \tag{228}$$

$$+ \frac{1+\gamma^{-1}}{1+\gamma^{-1}-\gamma^{-1/2}} \sum_{b:Y_-(b)}\sum_{u:v_u<0,y_{ub}>0} v_u \frac{(y_{ub})_+}{\sum_{b'}(y_{ub'})_+} \tag{229}$$

Now we use the same trick as for the positive upper bound which allows us to drop the conditioning in the $\sum_{u:v_u<0,y_{ub}>0}$ and $\sum_{u:v_u>0,y_{ub}<0}$ on the sign of y_{ub} - because the for the additional terms $(y_{ub})_+ = 0$ and $(y_{ub})_- = 0$ respectively:

$$= \frac{\gamma^{-1}}{(\gamma^{-1}-(1+\gamma^{-1})\gamma^{-1/2})} \sum_{b:Y_-(b)}\sum_{u:v_u>0} v_u \frac{(y_{ub})_-}{\sum_{b'}(y_{ub'})_-} \tag{230}$$

$$+ \frac{1+\gamma^{-1}}{1+\gamma^{-1}-\gamma^{-1/2}} \sum_{b:Y_-(b)}\sum_{u:v_u<0} v_u \frac{(y_{ub})_+}{\sum_{b'}(y_{ub'})_+} \tag{231}$$

$$\geq \frac{\gamma^{-1}}{(\gamma^{-1}-(1+\gamma^{-1})\gamma^{-1/2})} \sum_{b}\sum_{u:v_u>0} v_u \frac{(y_{ub})_-}{\sum_{b'}(y_{ub'})_-} \tag{232}$$

$$+ \frac{1+\gamma^{-1}}{1+\gamma^{-1}-\gamma^{-1/2}} \sum_{b}\sum_{u:v_u<0} v_u \frac{(y_{ub})_+}{\sum_{b'}(y_{ub'})_+} \tag{233}$$

$$= \frac{\gamma^{-1}}{(\gamma^{-1}-(1+\gamma^{-1})\gamma^{-1/2})} \sum_{u:v_u>0} v_u \frac{\sum_b (y_{ub})_-}{\sum_{b'}(y_{ub'})_-} \tag{234}$$

$$+ \frac{1+\gamma^{-1}}{1+\gamma^{-1}-\gamma^{-1/2}} \sum_{u:v_u<0} v_u \frac{\sum_b (y_{ub})_+}{\sum_{b'}(y_{ub'})_+} \tag{235}$$

$$= \frac{\gamma^{-1}}{(\gamma^{-1}-(1+\gamma^{-1})\gamma^{-1/2})} \sum_{u:v_u>0} v_u + \tag{236}$$

$$+ \frac{1+\gamma^{-1}}{1+\gamma^{-1}-\gamma^{-1/2}} \sum_{u:v_u<0} v_u \tag{237}$$

$$= \frac{1}{(1-(1+\gamma^{-1})\gamma^{1/2})} \sum_{u:v_u>0} v_u + \tag{238}$$

$$+ \frac{1+\gamma}{1+\gamma-\gamma^{1/2}} \sum_{u:v_u<0} v_u \tag{239}$$

Here we can plug in again the induction assumption to obtain

$$\geq \frac{1}{(1-(1+\gamma^{-1})\gamma^{1/2})} 2^{t-2} \frac{1+\gamma}{1+\gamma-\gamma^{1/2}} b(\gamma)^{t-2} + \tag{240}$$

$$+ \frac{1+\gamma}{1+\gamma-\gamma^{1/2}} 2^{t-2} \frac{1}{1-\gamma^{1/2}} b(\gamma)^{t-2} \tag{241}$$

$$\geq \frac{1}{(1-\gamma^{1/2})} 2^{t-2} \frac{1+\gamma}{1+\gamma-\gamma^{1/2}} b(\gamma)^{t-2} + \tag{242}$$

$$+ \frac{1+\gamma}{1+\gamma-\gamma^{1/2}} 2^{t-2} \frac{1}{1-\gamma^{1/2}} b(\gamma)^{t-2} \tag{243}$$

$$= \frac{1}{(1-\gamma^{1/2})} 2^{t-1} b(\gamma)^{t-2} \frac{1+\gamma}{1+\gamma-\gamma^{1/2}} \tag{244}$$

$$\geq \frac{1}{(1-\gamma^{1/2})} 2^{t-1} b(\gamma)^{t-2} b(\gamma) \tag{245}$$

$$= \frac{1}{(1-\gamma^{1/2})} 2^{t-1} b(\gamma)^{t-1} \tag{246}$$

13. Proof of Theorem 7

Proof: We know that for a vector v such that $\|v\|_2 = 1$, to be a singular vector for value $c \geq 0$ implies that there exists another unit norm vector $\|u\|_2 = 1$ such that:

$$M^\top v = cu \Rightarrow v^\top M M^\top v = c^2 \|u\|_2^2 = c^2 \tag{247}$$

Now:

$$\begin{aligned} v^\top M M^\top v &= \sum_k (v \cdot M[:,k])(M^\top[k,:] \cdot v) \\ &= \sum_k (v \cdot M[:,k])^2 \end{aligned} \tag{248}$$

We can maximize an inner product by choosing $v = \frac{M[:,k]}{\|M[:,k]\|}$. This yields here an upper bound because we have k vectors $M[:,k]$ but only one vector v.

$$\begin{aligned} &\sup_{v:\|v\|_2=1} v^\top M M^\top v \\ &= \sup_{v:\|v\|_2=1} \sum_k (v \cdot M[:,k])^2 \\ &\leq \sum_k (\frac{1}{\|M[:,k]\|} M[:,k] \cdot M[:,k])^2 \\ &= \sum_k \|M[:,k]\|_2^2 \end{aligned} \tag{249}$$

Next we consider the specific shape of $M[:,k]$ under LRP-β. $M[s,k]$ contains in every sum exclusively either a term $(1+\beta)(w_{ks} \cdot z_s)_+ / C_{k+}$ or a term $-\beta(w_{ks} \cdot z_s)_- / C_{k-}$. Both cannot be present at the same time.

We aim to compute $\|M[:,k]\|_2^2$. This norm is invariant to reordering the components of the vector $M[:,k]$. We can assume without loss of generality after ordering the terms according to the sign of $w_{ks} \cdot z_s$ that

$$\begin{aligned} M[:,k] = \big((1+\beta)p_{1,+}, \ldots, (1+\beta)p_{t,+}, \\ -\beta p_{t+1,-}, \ldots, -\beta p_{S,-}\big) \end{aligned} \tag{250}$$

where $\sum_{i=1}^t p_{i,+} = 1$ and $\sum_{i=t+1}^S p_{i,-} = 1$.

This is due to the fact that in LRP-β the positive entries $\frac{(w_{ab}z_b)_+}{\sum_{b'}(w_{ab'}z_{b'})_+}$ and the negative entries $\frac{(w_{ab}z_b)_-}{\sum_{b'}(w_{ab'}z_{b'})_-}$ are separately normalized to sum up to 1 for both signs. Now

$$\begin{aligned} \|M[:,k]\|_2^2 &= ((1+\beta)^2 \sum_{i=1}^t p_{i,+}^2 + \beta^2 \sum_{i=t+1}^S p_{i,-}^2 \\ &\leq (1+\beta)^2 \sum_{i=1}^t p_{i,+} + \beta^2 \sum_{i=t+1}^S p_{i,-} \\ &= (1+\beta)^2 + \beta^2 \end{aligned} \tag{251}$$

To obtain an upper bound on the largest singular value, we have to consider

$$\sup_{v:\|v\|_2=1} v^\top M M^\top v \leq \sum_{k=1}^R \|M[:,k]\|_2^2 \leq R((1+\beta)^2 + \beta^2) \tag{252}$$

Taking the square root results in

$$\sqrt{R}\sqrt{(1+\beta)^2 + \beta^2} \tag{253}$$

A more interpretable form can be derived by

$$\begin{aligned} &\sqrt{R}\sqrt{(1+\beta)^2 + \beta^2} = \sqrt{R}\sqrt{1 + 2\beta + 2\beta^2} \\ &\leq \sqrt{R}\sqrt{1 + 2\sqrt{2}\beta + (\sqrt{2}\beta)^2} = \sqrt{R}(1 + \sqrt{2}\beta) \end{aligned} \tag{254}$$

14. Convergence Statistics for the gradient against LRP-β.

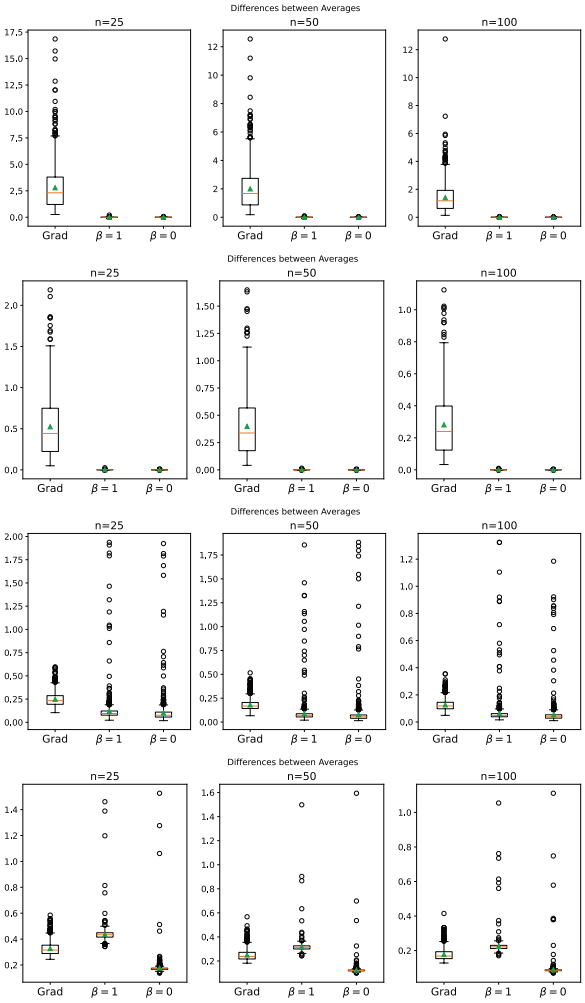

Figure 1: Convergence statistics for EfficientNet-V2-S. Lower is better. First row: no normalization, photometric augmentation. Second row: no normalization, noise augmentation. Third row: ℓ_2-normalization, photometric augmentation. Fourth row: ℓ_2-normalization, noise augmentation.

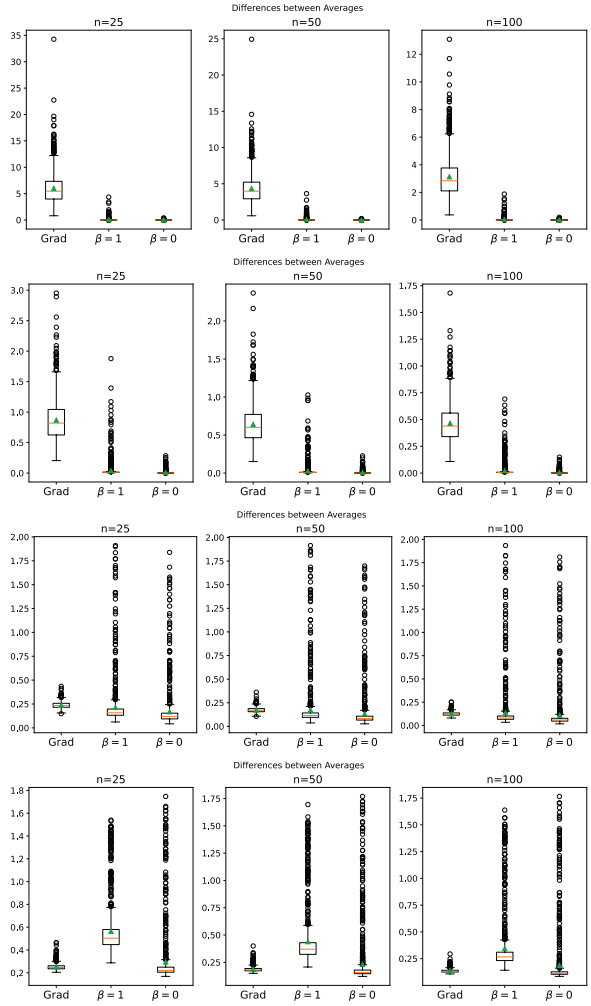

Figure 2: Convergence statistics for ResNet-50. Lower is better. First row: no normalization, photometric augmentation. Second row: no normalization, noise augmentation. third row: ℓ_2-normalization, photometric augmentation. Fourth row: ℓ_2-normalization, noise augmentation.

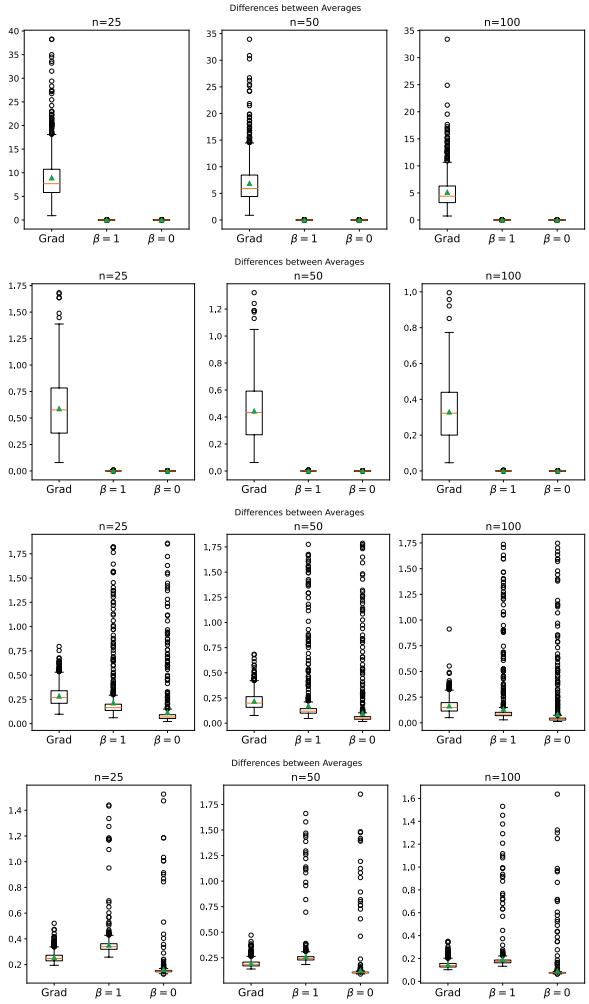

Figure 3: Convergence statistics for SwinTransformer-V2-Tiny. Lower is better. First row: no normalization, photometric augmentation. Second row: no normalization, noise augmentation. third row: ℓ_2-normalization, photometric augmentation. Fourth row: ℓ_2-normalization, noise augmentation.

15. Convergence Statistics for the gradient times input against LRP-β.

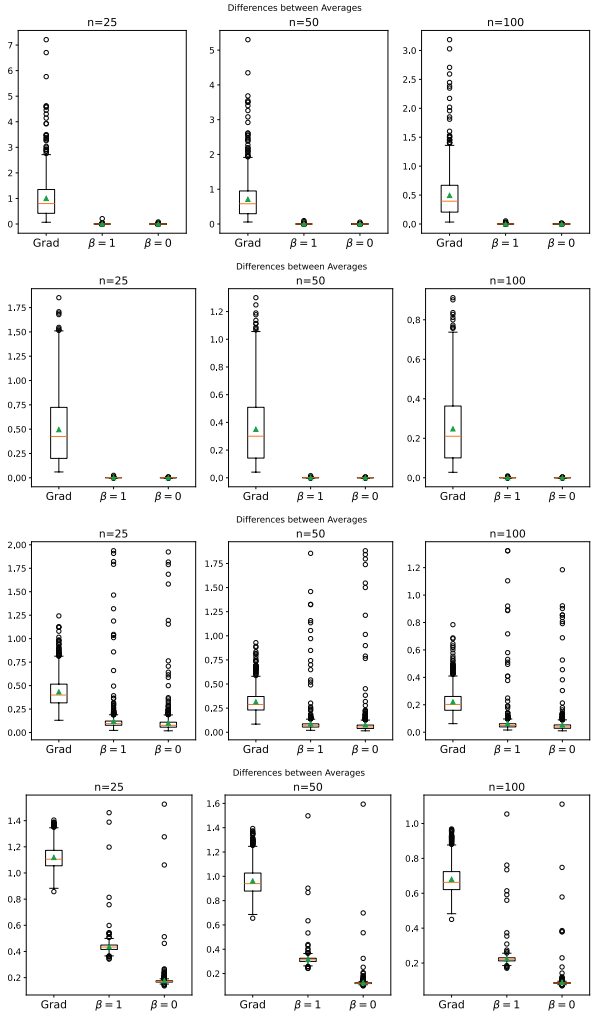

Figure 4: Convergence statistics for EfficientNet-V2-S. Lower is better. First row: no normalization, photometric augmentation. Second row: no normalization, noise augmentation. Third row: ℓ_2-normalization, photometric augmentation. Fourth row: ℓ_2-normalization, noise augmentation.

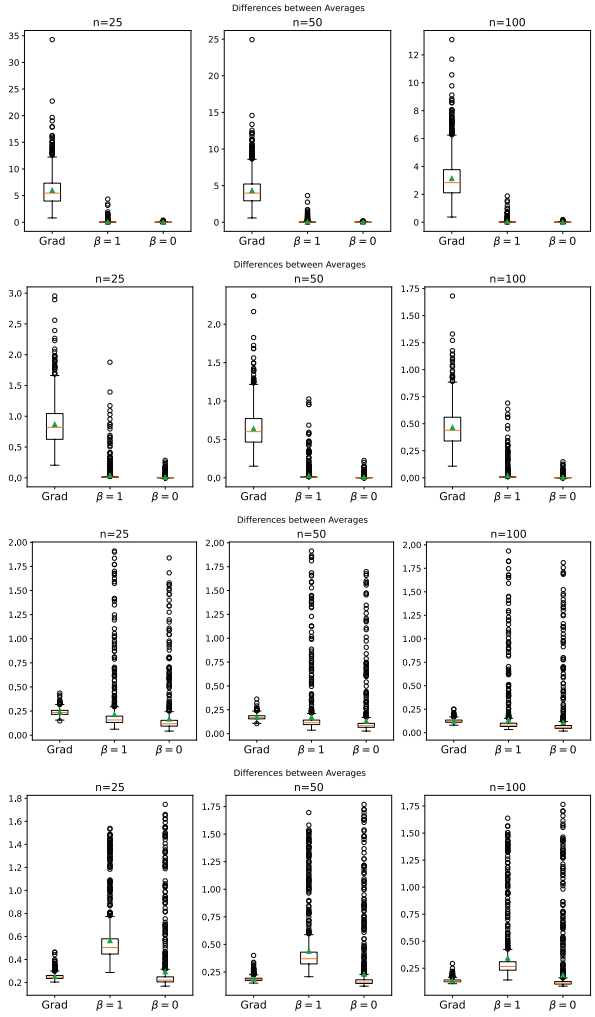

Figure 5: Convergence statistics for ResNet-50. Lower is better. First row: no normalization, photometric augmentation. Second row: no normalization, noise augmentation. third row: ℓ_2-normalization, photometric augmentation. Fourth row: ℓ_2-normalization, noise augmentation.

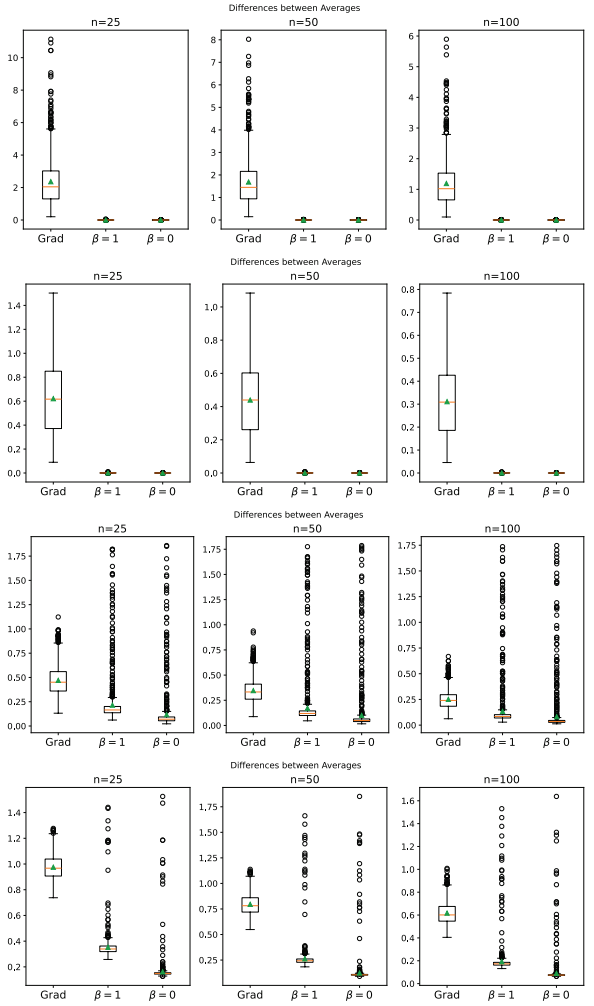

Figure 6: Convergence statistics for SwinTransformer-V2-Tiny. Lower is better. First row: no normalization, photometric augmentation. Second row: no normalization, noise augmentation. third row: ℓ_2-normalization, photometric augmentation. Fourth row: ℓ_2-normalization, noise augmentation.

16. Convergence Statistics for the gradient against LRP-γ

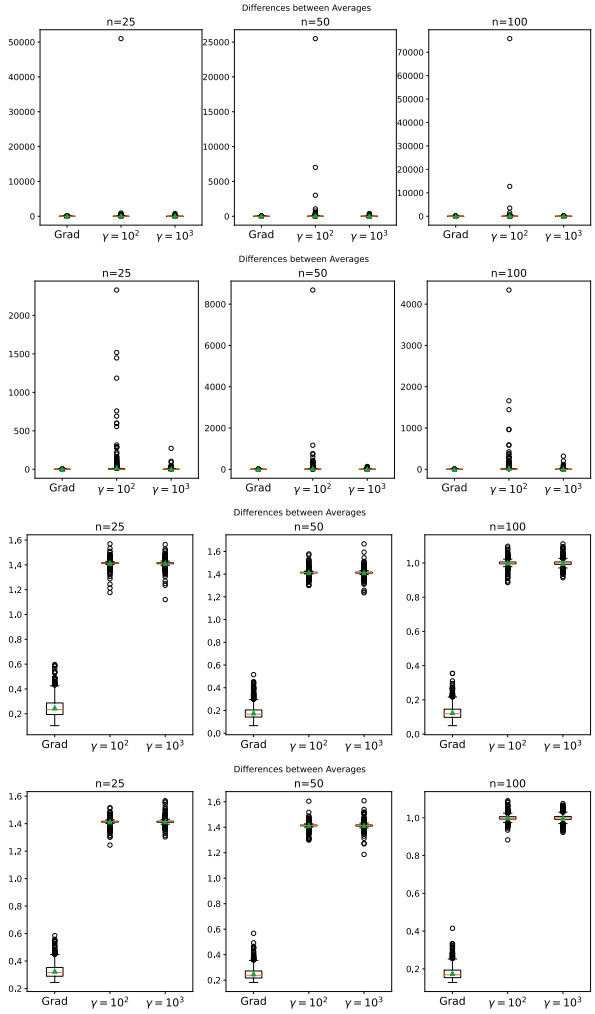

Figure 7: Convergence statistics for EfficientNet-V2-S. Lower is better. First row: no normalization, photometric augmentation. Second row: no normalization, noise augmentation. Third row: ℓ_2-normalization, photometric augmentation. Fourth row: ℓ_2-normalization, noise augmentation.

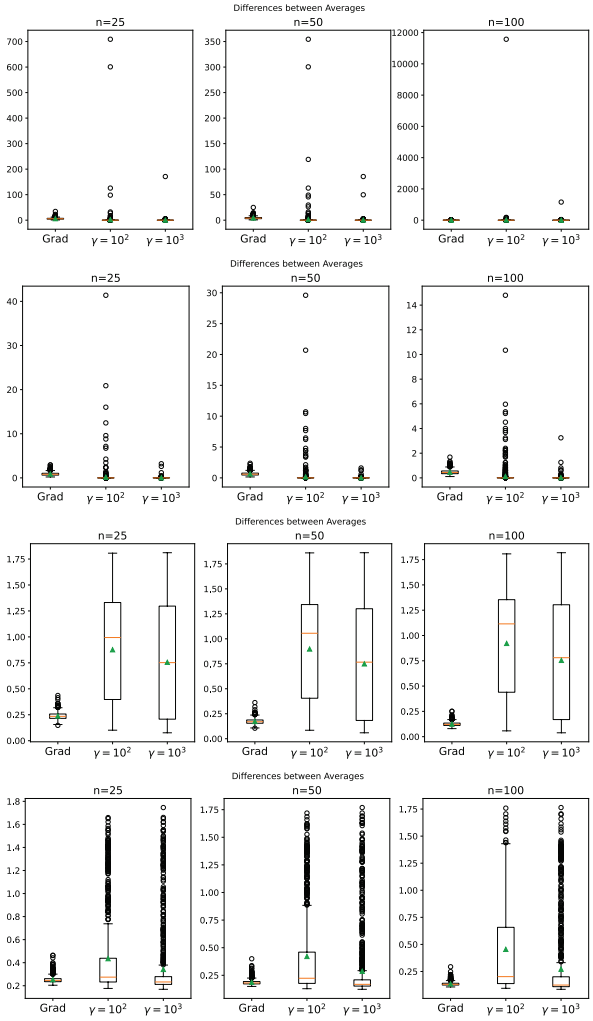

Figure 8: Convergence statistics for ResNet-50. Lower is better. First row: no normalization, photometric augmentation. Second row: no normalization, noise augmentation. Third row: ℓ_2-normalization, photometric augmentation. Fourth row: ℓ_2-normalization, noise augmentation.

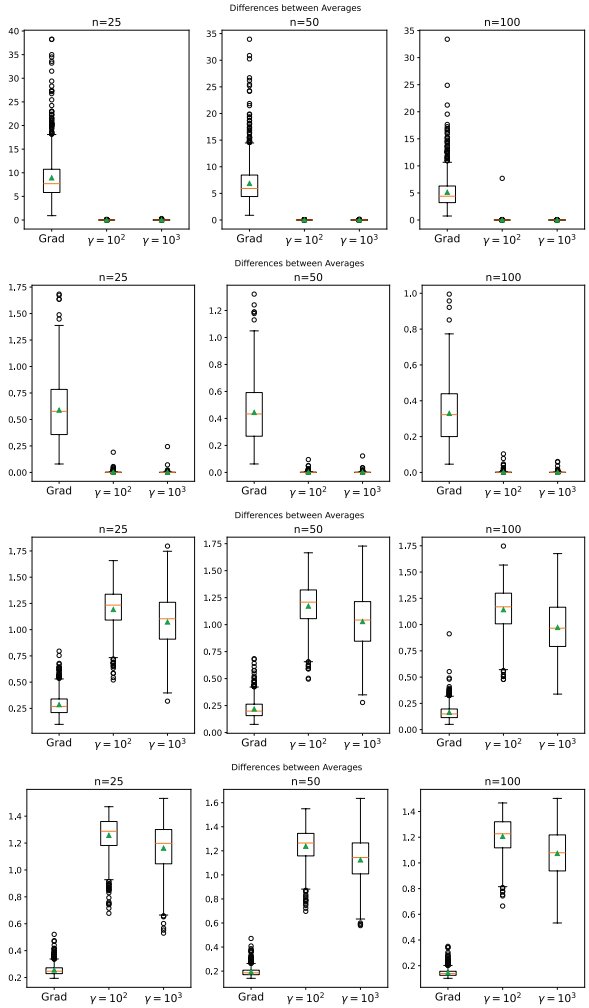

Figure 9: Convergence statistics for SwinTransformer-V2-Tiny. Lower is better. First row: no normalization, photometric augmentation. Second row: no normalization, noise augmentation. third row: ℓ_2-normalization, photometric augmentation. Fourth row: ℓ_2-normalization, noise augmentation.

17. Convergence Statistics for the gradient times input against LRP-γ.

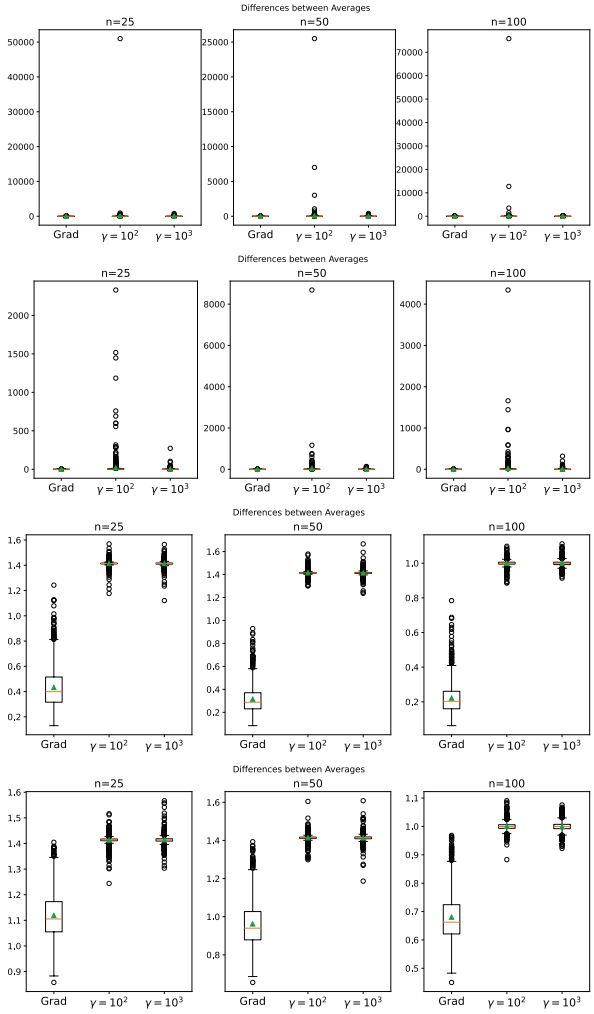

Figure 10: Convergence statistics for EfficientNet-V2-S. Lower is better. First row: no normalization, photometric augmentation. Second row: no normalization, noise augmentation. Third row: ℓ_2-normalization, photometric augmentation. Fourth row: ℓ_2-normalization, noise augmentation.

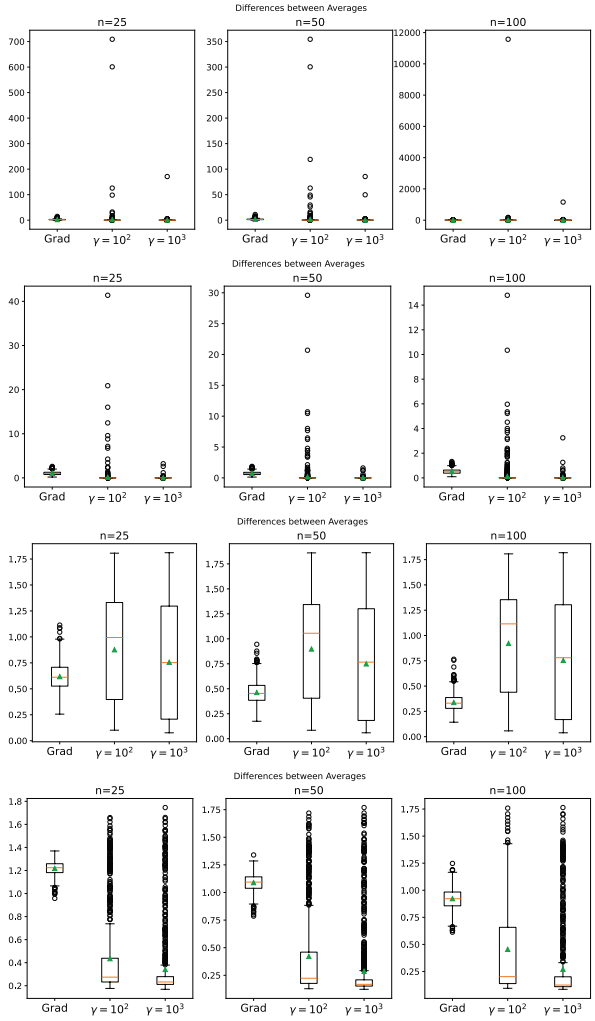

Figure 11: Convergence statistics for ResNet-50. Lower is better. First row: no normalization, photometric augmentation. Second row: no normalization, noise augmentation. Third row: ℓ_2-normalization, photometric augmentation. Fourth row: ℓ_2-normalization, noise augmentation.

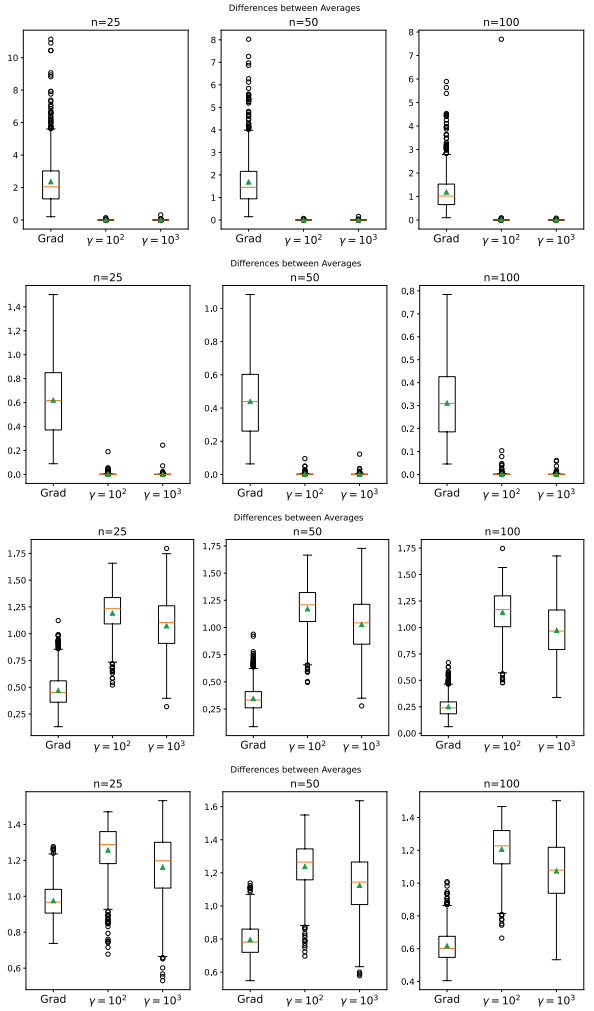

Figure 12: Convergence statistics for SwinTransformer-V2-Tiny. Lower is better. First row: no normalization, photometric augmentation. Second row: no normalization, noise augmentation. third row: ℓ_2-normalization, photometric augmentation. Fourth row: ℓ_2-normalization, noise augmentation.

18. Convergence Statistics for the gradient against LRP-γ-lifted.

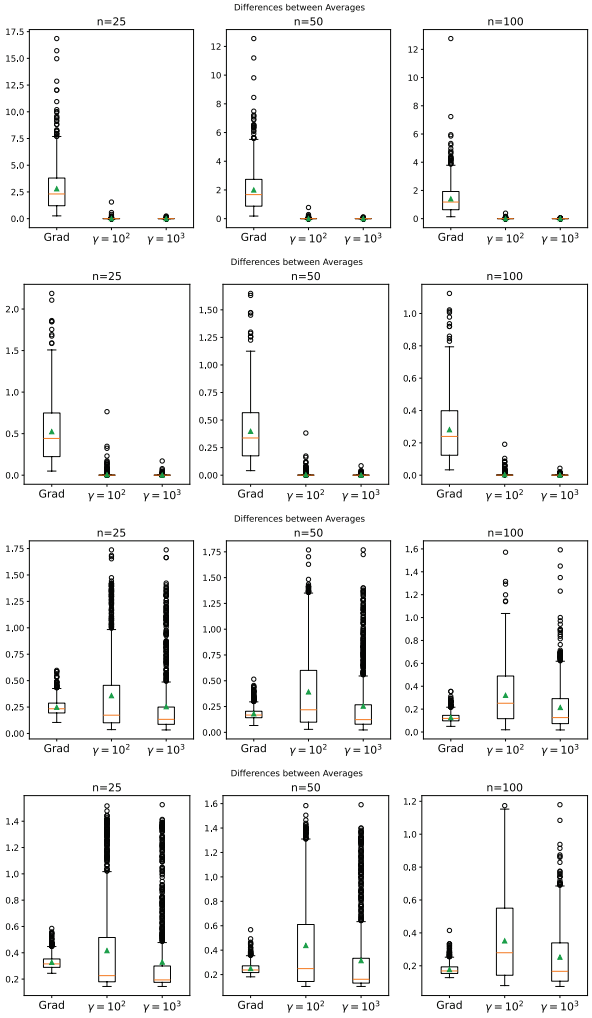

Figure 13: Convergence statistics for EfficientNet-V2-S. Lower is better. First row: no normalization, photometric augmentation. Second row: no normalization, noise augmentation. Third row: ℓ_2-normalization, photometric augmentation. Fourth row: ℓ_2-normalization, noise augmentation.

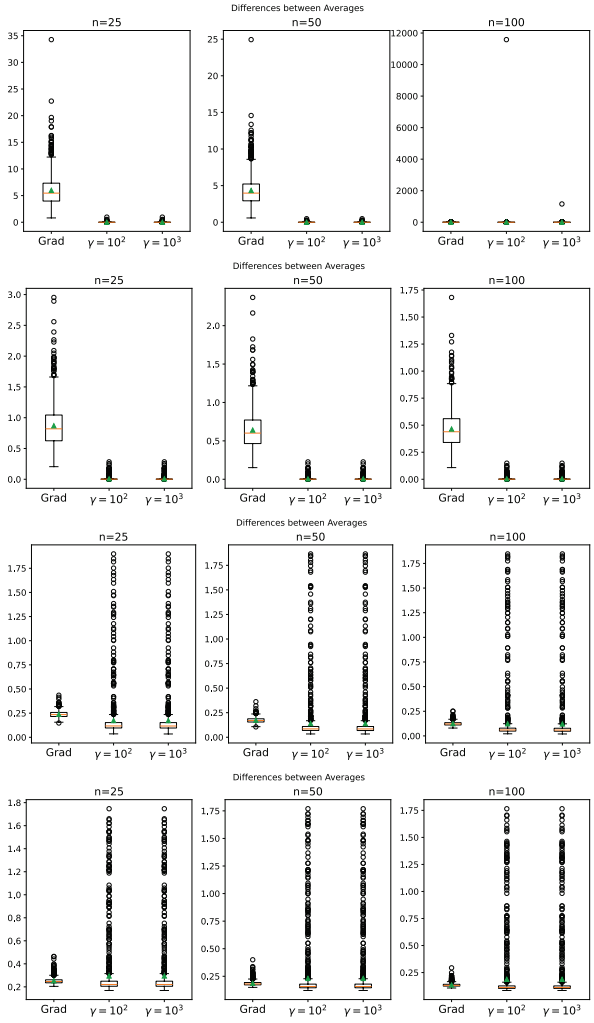

Figure 14: Convergence statistics for ResNet-50. Lower is better. First row: no normalization, photometric augmentation. Second row: no normalization, noise augmentation. Third row: ℓ_2-normalization, photometric augmentation. Fourth row: ℓ_2-normalization, noise augmentation.

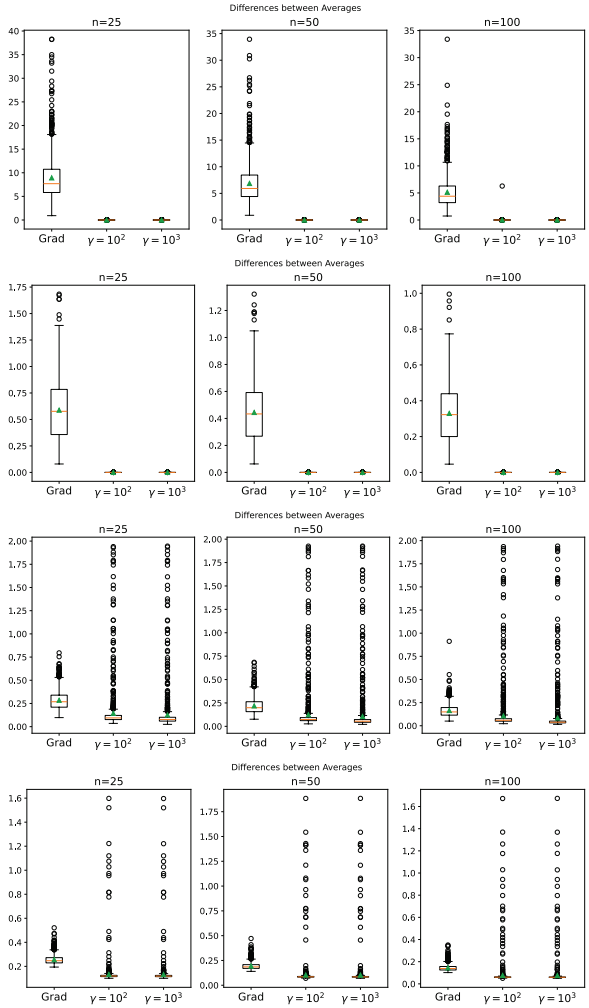

Figure 15: Convergence statistics for SwinTransformer-V2-Tiny. Lower is better. First row: no normalization, photometric augmentation. Second row: no normalization, noise augmentation. third row: ℓ_2-normalization, photometric augmentation. Fourth row: ℓ_2-normalization, noise augmentation.

19. Convergence Statistics for the gradient times input against LRP-γ-lifted.

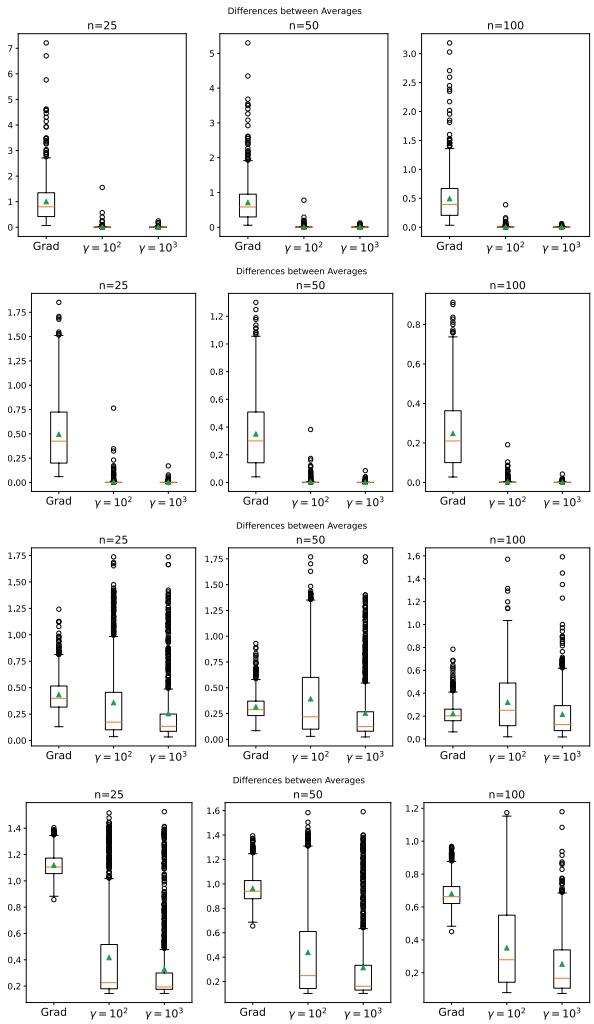

Figure 16: Convergence statistics for EfficientNet-V2-S. Lower is better. First row: no normalization, photometric augmentation. Second row: no normalization, noise augmentation. Third row: ℓ_2-normalization, photometric augmentation. Fourth row: ℓ_2-normalization, noise augmentation.

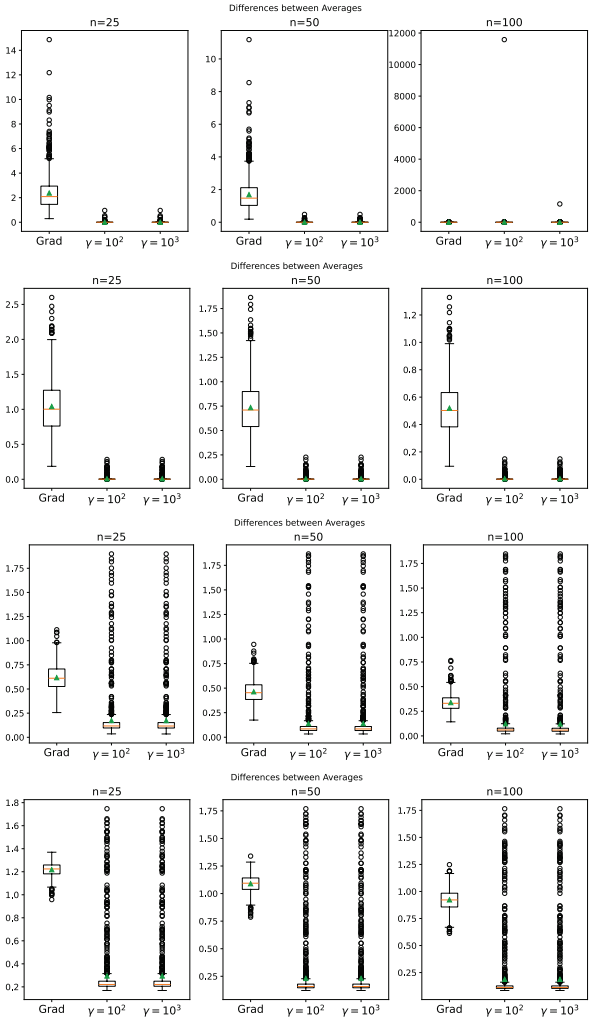

Figure 17: Convergence statistics for ResNet-50. Lower is better. First row: no normalization, photometric augmentation. Second row: no normalization, noise augmentation. Third row: ℓ_2-normalization, photometric augmentation. Fourth row: ℓ_2-normalization, noise augmentation.

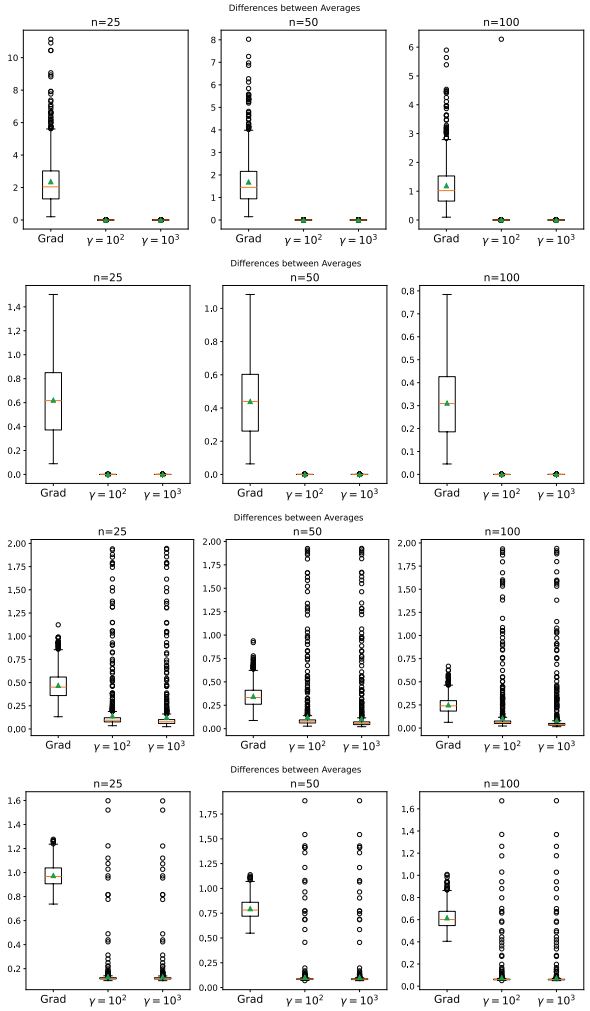

Figure 18: Convergence statistics for SwinTransformer-V2-Tiny. Lower is better. First row: no normalization, photometric augmentation. Second row: no normalization, noise augmentation. third row: ℓ_2-normalization, photometric augmentation. Fourth row: ℓ_2-normalization, noise augmentation.

20. Ratio of Medians of statistics between a Gradient-variant and an LRP-variant, Unnormalized case, covered by the theoretical results

Table 3. EffNet-V2-S no normalization: Gradient attribution compared to LRP-β.

Augmentation	Sample size	p-value	ratio
	Grad vs LRP-$\beta = 0$		
Gaussian	$m = 25$	$2E-165$	1549.4
	$m = 50$	$2E-165$	1665.5
	$m = 100$	$2E-165$	1672.8
photometric	$m = 25$	$2E-165$	15616.3
	$m = 50$	$2E-165$	15958.3
	$m = 100$	$2E-165$	15741.1
	Grad vs LRP-$\beta = 1$		
Gaussian	$m = 25$	$2E-165$	473.1
	$m = 50$	$2E-165$	510.5
	$m = 100$	$2E-165$	512.9
photometric	$m = 25$	$2E-165$	8053.8
	$m = 50$	$2E-165$	8178.5
	$m = 100$	$2E-165$	8104.3

Table 4. ResNet-50 no normalization: Gradient attribution compared to LRP-β.

Augmentation	Sample size	p-value	ratio
	Grad vs LRP-$\beta = 0$		
Gaussian	$m = 25$	$2E-165$	433.5
	$m = 50$	$2E-165$	446.7
	$m = 100$	$2E-165$	462.7
photometric	$m = 25$	$2E-165$	3597.1
	$m = 50$	$2E-165$	3732.0
	$m = 100$	$2E-165$	3760.5
	Grad vs LRP-$\beta = 1$		
Gaussian	$m = 25$	$1E-162$	68.1
	$m = 50$	$6E-163$	70.9
	$m = 100$	$2E-163$	73.9
photometric	$m = 25$	$2E-165$	314.3
	$m = 50$	$2E-165$	318.3
	$m = 100$	$2E-165$	321.7

Table 5. SwinTransformer-V2-Tiny no normalization: Gradient attribution compared to LRP-β.

Augmentation	Sample size	p-value	ratio
	Grad vs LRP-$\beta = 0$		
Gaussian	$m = 25$	$2E-165$	10757.4
	$m = 50$	$2E-165$	11410.3
	$m = 100$	$2E-165$	12004.2
photometric	$m = 25$	$2E-165$	215298.1
	$m = 50$	$2E-165$	239405.3
	$m = 100$	$2E-165$	249866.1
	Grad vs LRP-$\beta = 1$		
Gaussian	$m = 25$	$2E-165$	1469.6
	$m = 50$	$2E-165$	1565.7
	$m = 100$	$2E-165$	1646.0
photometric	$m = 25$	$2E-165$	23303.2
	$m = 50$	$2E-165$	24862.0
	$m = 100$	$2E-165$	26216.2

Table 6. EffNet-V2-S no normalization: Gradient $\times$ Input attribution compared to LRP-β.

Augmentation	Sample size	p-value	ratio
	Grad vs LRP-$\beta = 0$		
Gaussian	$m = 25$	$2E-165$	1486.9
	$m = 50$	$2E-165$	1481.6
	$m = 100$	$2E-165$	1463.6
photometric	$m = 25$	$2E-165$	5391.5
	$m = 50$	$2E-165$	5552.4
	$m = 100$	$2E-165$	5268.6
	Grad vs LRP-$\beta = 1$		
Gaussian	$m = 25$	$2E-165$	454.0
	$m = 50$	$2E-165$	454.1
	$m = 100$	$2E-165$	448.7
photometric	$m = 25$	$2E-165$	2780.6
	$m = 50$	$2E-165$	2845.5
	$m = 100$	$2E-165$	2712.5

Table 7. ResNet-50 no normalization: Gradient $\times$ Input attribution compared to LRP-β.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\beta = 0$			
Gaussian	$m = 25$	$2E-165$	530.0
	$m = 50$	$2E-165$	527.7
	$m = 100$	$2E-165$	529.6
photometric	$m = 25$	$2E-165$	1375.7
	$m = 50$	$2E-165$	1385.1
	$m = 100$	$2E-165$	1387.7
Grad vs LRP-$\beta = 1$			
Gaussian	$m = 25$	$5E-164$	83.2
	$m = 50$	$8E-164$	83.8
	$m = 100$	$5E-164$	84.6
photometric	$m = 25$	$3E-163$	120.2
	$m = 50$	$3E-163$	118.1
	$m = 100$	$5E-164$	118.7

Table 8. SwinTransformer-V2-Tiny no normalization: Gradient $\times$ Input attribution compared to LRP-β.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\beta = 0$			
Gaussian	$m = 25$	$2E-165$	11514.4
	$m = 50$	$2E-165$	11561.4
	$m = 100$	$2E-165$	11487.3
photometric	$m = 25$	$2E-165$	57204.5
	$m = 50$	$2E-165$	58607.9
	$m = 100$	$2E-165$	57993.5
Grad vs LRP-$\beta = 1$			
Gaussian	$m = 25$	$2E-165$	1573.0
	$m = 50$	$2E-165$	1586.4
	$m = 100$	$2E-165$	1575.1
photometric	$m = 25$	$2E-165$	6191.6
	$m = 50$	$2E-165$	6086.4
	$m = 100$	$2E-165$	6084.7

Table 9. EffNet-V2-S no normalization: Gradient attribution compared to LRP-γ.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	1E+00	0.9
	$m = 50$	1E+00	0.6
	$m = 100$	1E+00	0.4
photometric	$m = 25$	5E-73	4.7
	$m = 50$	2E-48	3.4
	$m = 100$	1E-22	2.2
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	1E+00	0.2
	$m = 50$	1E+00	0.1
	$m = 100$	1E+00	0.1
photometric	$m = 25$	1E+00	1.3
	$m = 50$	1E+00	0.7
	$m = 100$	1E+00	0.4

Table 10. ResNet-50 no normalization: Gradient attribution compared to LRP-γ.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	7E-163	405.2
	$m = 50$	3E-162	413.4
	$m = 100$	5E-162	407.1
photometric	$m = 25$	4E-164	388.8
	$m = 50$	7E-163	294.7
	$m = 100$	8E-158	196.6
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	8E-149	334.1
	$m = 50$	5E-139	308.1
	$m = 100$	1E-131	255.1
photometric	$m = 25$	1E-154	239.3
	$m = 50$	6E-149	148.7
	$m = 100$	7E-143	100.2

Table 11. SwinTransformer-V2-Tiny no normalization: Gradient attribution compared to LRP-γ.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	2E-165	925.6
	$m = 50$	2E-165	805.8
	$m = 100$	2E-165	731.2
photometric	$m = 25$	2E-165	14504.4
	$m = 50$	2E-165	13453.4
	$m = 100$	2E-165	11272.4
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	2E-165	641.7
	$m = 50$	2E-165	547.4
	$m = 100$	2E-165	470.3
photometric	$m = 25$	2E-165	9385.9
	$m = 50$	2E-165	8099.6
	$m = 100$	7E-165	6977.9

Table 12. EffNet-V2-S no normalization: Gradient attribution compared to LRP-γ-lifted.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	2E-165	1296.8
	$m = 50$	2E-165	1210.6
	$m = 100$	2E-165	878.0
photometric	$m = 25$	2E-165	9144.9
	$m = 50$	2E-165	7073.2
	$m = 100$	2E-165	4832.1
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	1E-164	1141.2
	$m = 50$	5E-165	801.5
	$m = 100$	3E-165	470.2
photometric	$m = 25$	3E-165	6881.6
	$m = 50$	2E-165	4072.2
	$m = 100$	2E-165	2385.1

Table 13. Resnet50 no normalization: Gradient attribution compared to LRP-γ-lifted.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	2E-165	433.5
	$m = 50$	2E-165	446.7
	$m = 100$	2E-165	462.8
photometric	$m = 25$	2E-165	3562.0
	$m = 50$	2E-165	3610.7
	$m = 100$	3E-164	3727.5
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	2E-165	433.6
	$m = 50$	2E-165	446.5
	$m = 100$	2E-165	463.0
photometric	$m = 25$	2E-165	3541.5
	$m = 50$	2E-165	3581.0
	$m = 100$	3E-164	3667.5

Table 14. SwinTransformer-V2-Tiny no normalization: Gradient attribution compared to LRP-γ-lifted.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	2E-165	20507.5
	$m = 50$	2E-165	21722.4
	$m = 100$	2E-165	22806.0
photometric	$m = 25$	2E-165	340111.4
	$m = 50$	2E-165	368805.0
	$m = 100$	2E-165	391209.2
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	2E-165	20412.8
	$m = 50$	2E-165	21541.6
	$m = 100$	2E-165	22493.0
photometric	$m = 25$	2E-165	274937.7
	$m = 50$	2E-165	278034.6
	$m = 100$	3E-165	257857.2

Table 15. EfficientNet-V2-S no normalization: Gradient × Input attribution compared to LRP-γ.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	1E+00	0.8
	$m = 50$	1E+00	0.6
	$m = 100$	1E+00	0.4
photometric	$m = 25$	5E-06	1.6
	$m = 50$	1E+00	1.2
	$m = 100$	1E+00	0.7
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	1E+00	0.2
	$m = 50$	1E+00	0.1
	$m = 100$	1E+00	0.1
photometric	$m = 25$	1E+00	0.4
	$m = 50$	1E+00	0.2
	$m = 100$	1E+00	0.1

Table 16. ResNet-50 no normalization: Gradient × Input attribution compared to LRP-γ.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	6E-163	495.4
	$m = 50$	8E-163	488.4
	$m = 100$	3E-162	465.9
photometric	$m = 25$	2E-157	148.7
	$m = 50$	1E-159	109.4
	$m = 100$	1E-150	72.5
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	1E-149	408.5
	$m = 50$	4E-141	363.9
	$m = 100$	1E-133	292.0
photometric	$m = 25$	5E-142	91.5
	$m = 50$	1E-132	55.2
	$m = 100$	2E-120	37.0

Table 17. SwinTransformer-V2-Tiny no normalization: Gradient × Input attribution compared to LRP-γ.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	2E-165	990.7
	$m = 50$	2E-165	816.5
	$m = 100$	2E-165	699.7
photometric	$m = 25$	2E-165	3853.8
	$m = 50$	2E-165	3293.5
	$m = 100$	2E-165	2616.3
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	2E-165	686.9
	$m = 50$	2E-165	554.6
	$m = 100$	2E-165	450.0
photometric	$m = 25$	2E-165	2493.8
	$m = 50$	2E-165	1982.8
	$m = 100$	3E-164	1619.6

Table 18. EfficientNet-V2-S no normalization: Gradient × Input attribution compared to LRP-γ-lifted.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	2E-165	1244.5
	$m = 50$	2E-165	1077.0
	$m = 100$	2E-165	768.2
photometric	$m = 25$	2E-165	3157.3
	$m = 50$	2E-165	2461.0
	$m = 100$	2E-165	1617.3
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	1E-164	1095.2
	$m = 50$	7E-165	713.0
	$m = 100$	4E-165	411.4
photometric	$m = 25$	1E-164	2375.9
	$m = 50$	7E-165	1416.8
	$m = 100$	4E-165	798.3

Table 19. ResNet-50 no normalization: Gradient $\times$ Input attribution compared to LRP-γ-lifted.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	2E-165	530.1
	$m = 50$	2E-165	527.7
	$m = 100$	2E-165	529.7
photometric	$m = 25$	2E-165	1362.3
	$m = 50$	2E-165	1340.1
	$m = 100$	3E-164	1375.6
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	2E-165	530.1
	$m = 50$	2E-165	527.5
	$m = 100$	2E-165	529.9
photometric	$m = 25$	2E-165	1354.4
	$m = 50$	2E-165	1329.0
	$m = 100$	3E-164	1353.4

Table 20. SwinTransformer-V2-Tiny no normalization: Gradient $\times$ Input attribution compared to LRP-γ-lifted.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	2E-165	21950.6
	$m = 50$	2E-165	22010.1
	$m = 100$	2E-165	21824.0
photometric	$m = 25$	2E-165	90367.3
	$m = 50$	2E-165	90285.7
	$m = 100$	2E-165	90799.1
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	2E-165	21849.3
	$m = 50$	2E-165	21826.9
	$m = 100$	2E-165	21524.5
photometric	$m = 25$	2E-165	73050.7
	$m = 50$	2E-165	68064.5
	$m = 100$	3E-164	59848.2

21. Ratio of Medians of statistics between a Gradient-variant and an LRP-variant, ℓ_2-normalized case, not covered by the theoretical results

Table 21. EfficientNet-V2-S with ℓ_2-normalization: Gradient attribution compared to LRP-β.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\beta = 0$			
Gaussian	$m = 25$	$2E-160$	1.8
	$m = 50$	$2E-161$	2.0
	$m = 100$	$3E-157$	2.0
photometric	$m = 25$	$5E-146$	3.1
	$m = 50$	$7E-145$	3.1
	$m = 100$	$5E-146$	3.0
Grad vs LRP-$\beta = 1$			
Gaussian	$m = 25$	$1E+00$	0.7
	$m = 50$	$1E+00$	0.8
	$m = 100$	$1E+00$	0.8
photometric	$m = 25$	$1E-144$	2.5
	$m = 50$	$4E-142$	2.5
	$m = 100$	$1E-142$	2.4

Table 22. ResNet-50 with ℓ_2-normalization: Gradient attribution compared to LRP-β.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\beta = 0$			
Gaussian	$m = 25$	$9E-30$	1.1
	$m = 50$	$9E-39$	1.1
	$m = 100$	$3E-42$	1.2
photometric	$m = 25$	$2E-101$	2.0
	$m = 50$	$9E-98$	2.1
	$m = 100$	$2E-97$	2.1
Grad vs LRP-$\beta = 1$			
Gaussian	$m = 25$	$1E+00$	0.5
	$m = 50$	$1E+00$	0.5
	$m = 100$	$1E+00$	0.5
photometric	$m = 25$	$5E-83$	1.5
	$m = 50$	$4E-79$	1.5
	$m = 100$	$3E-77$	1.5

Table 23. SwinTransformer-V2-Tiny with ℓ_2-normalization: Gradient attribution compared to LRP-β.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\beta = 0$			
Gaussian	$m = 25$	$3E-147$	1.7
	$m = 50$	$5E-143$	1.8
	$m = 100$	$2E-138$	1.8
photometric	$m = 25$	$2E-115$	3.8
	$m = 50$	$1E-110$	4.0
	$m = 100$	$8E-105$	4.2
Grad vs LRP-$\beta = 1$			
Gaussian	$m = 25$	$1E+00$	0.7
	$m = 50$	$1E+00$	0.8
	$m = 100$	$1E+00$	0.8
photometric	$m = 25$	$3E-84$	1.6
	$m = 50$	$2E-84$	1.7
	$m = 100$	$6E-84$	1.8

Table 24. EfficientNet-V2-S with ℓ_2-normalization: Gradient $\times$ Input attribution compared to LRP-β.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\beta = 0$			
Gaussian	$m = 25$	$2E-165$	6.5
	$m = 50$	$2E-165$	7.8
	$m = 100$	$2E-165$	7.7
photometric	$m = 25$	$2E-154$	5.2
	$m = 50$	$6E-151$	5.3
	$m = 100$	$4E-151$	5.2
Grad vs LRP-$\beta = 1$			
Gaussian	$m = 25$	$2E-165$	2.5
	$m = 50$	$2E-165$	3.0
	$m = 100$	$2E-165$	3.0
photometric	$m = 25$	$9E-152$	4.2
	$m = 50$	$8E-149$	4.3
	$m = 100$	$1E-149$	4.1

Table 25. ResNet-50 with ℓ_2-normalization: Gradient $\times$ Input attribution compared to LRP-β.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\beta = 0$			
Gaussian	$m = 25$	$2E-164$	5.5
	$m = 50$	$2E-163$	6.9
	$m = 100$	$4E-160$	8.2
photometric	$m = 25$	$2E-142$	5.1
	$m = 50$	$5E-133$	5.5
	$m = 100$	$2E-121$	5.7
Grad vs LRP-$\beta = 1$			
Gaussian	$m = 25$	$3E-163$	2.4
	$m = 50$	$2E-161$	3.0
	$m = 100$	$1E-156$	3.5
photometric	$m = 25$	$7E-137$	3.9
	$m = 50$	$7E-123$	3.9
	$m = 100$	$6E-117$	4.2

Table 26. SwinTransformer-V2-Tiny with ℓ_2-normalization: Gradient $\times$ Input attribution compared to LRP-β.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\beta = 0$			
Gaussian	$m = 25$	$2E-165$	6.5
	$m = 50$	$7E-162$	7.4
	$m = 100$	$8E-161$	8.0
photometric	$m = 25$	$8E-136$	6.4
	$m = 50$	$1E-123$	6.8
	$m = 100$	$2E-115$	6.8
Grad vs LRP-$\beta = 1$			
Gaussian	$m = 25$	$2E-165$	2.9
	$m = 50$	$3E-158$	3.2
	$m = 100$	$7E-157$	3.5
photometric	$m = 25$	$2E-121$	2.7
	$m = 50$	$2E-112$	2.8
	$m = 100$	$3E-105$	2.9

Table 27. EfficientNet-V2-S with ℓ_2-normalization: Gradient attribution compared to LRP-γ.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	1E+00	0.2
	$m = 50$	1E+00	0.2
	$m = 100$	1E+00	0.2
photometric	$m = 25$	1E+00	0.2
	$m = 50$	1E+00	0.1
	$m = 100$	1E+00	0.1
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	1E+00	0.2
	$m = 50$	1E+00	0.2
	$m = 100$	1E+00	0.2
photometric	$m = 25$	1E+00	0.2
	$m = 50$	1E+00	0.1
	$m = 100$	1E+00	0.1

Table 28. ResNet-50 with ℓ_2-normalization: Gradient attribution compared to LRP-γ.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	4E-02	1.1
	$m = 50$	7E-02	1.1
	$m = 100$	1E+00	1.1
photometric	$m = 25$	1E+00	0.3
	$m = 50$	1E+00	0.2
	$m = 100$	1E+00	0.2
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	1E+00	0.9
	$m = 50$	1E+00	0.8
	$m = 100$	1E+00	0.7
photometric	$m = 25$	1E+00	0.2
	$m = 50$	1E+00	0.2
	$m = 100$	1E+00	0.1

Table 29. SwinTransformer-V2-Tiny with ℓ_2-normalization: Gradient attribution compared to LRP-γ.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	1E+00	0.2
	$m = 50$	1E+00	0.2
	$m = 100$	1E+00	0.1
photometric	$m = 25$	1E+00	0.2
	$m = 50$	1E+00	0.2
	$m = 100$	1E+00	0.2
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	1E+00	0.2
	$m = 50$	1E+00	0.1
	$m = 100$	1E+00	0.1
photometric	$m = 25$	1E+00	0.2
	$m = 50$	1E+00	0.2
	$m = 100$	1E+00	0.1

Table 30. EfficientNet-V2-S with ℓ_2-normalization: Gradient attribution compared to LRP-γ-lifted.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	3E-24	1.6
	$m = 50$	2E-04	1.5
	$m = 100$	1E+00	1.0
photometric	$m = 25$	9E-20	1.7
	$m = 50$	1E-01	1.4
	$m = 100$	1E+00	0.9
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	3E-01	1.4
	$m = 50$	1E+00	1.0
	$m = 100$	1E+00	0.6
photometric	$m = 25$	9E-01	1.4
	$m = 50$	1E+00	0.8
	$m = 100$	1E+00	0.5

Table 31. ResNet-50 with ℓ_2-normalization: Gradient attribution compared to LRP-γ-lifted.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	9E-30	1.1
	$m = 50$	9E-39	1.1
	$m = 100$	3E-42	1.2
photometric	$m = 25$	6E-99	2.0
	$m = 50$	4E-95	2.0
	$m = 100$	7E-92	2.0
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	9E-30	1.1
	$m = 50$	9E-39	1.1
	$m = 100$	3E-42	1.2
photometric	$m = 25$	7E-99	2.0
	$m = 50$	5E-94	2.0
	$m = 100$	8E-91	2.0

Table 32. SwinTransformer-V2-Tiny with ℓ_2-normalization: Gradient attribution compared to LRP-γ-lifted.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	3E-148	2.1
	$m = 50$	5E-143	2.2
	$m = 100$	1E-139	2.3
photometric	$m = 25$	3E-118	3.5
	$m = 50$	4E-110	3.6
	$m = 100$	7E-105	3.9
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	3E-148	2.1
	$m = 50$	5E-143	2.2
	$m = 100$	1E-139	2.3
photometric	$m = 25$	1E-115	2.8
	$m = 50$	1E-106	2.7
	$m = 100$	1E-98	2.5

Table 33. EfficientNet-V2-S with ℓ_2-normalization: Gradient $\times$ Input attribution compared to LRP-γ.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	1E+00	0.8
	$m = 50$	1E+00	0.7
	$m = 100$	1E+00	0.7
photometric	$m = 25$	1E+00	0.3
	$m = 50$	1E+00	0.2
	$m = 100$	1E+00	0.2
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	1E+00	0.8
	$m = 50$	1E+00	0.7
	$m = 100$	1E+00	0.7
photometric	$m = 25$	1E+00	0.3
	$m = 50$	1E+00	0.2
	$m = 100$	1E+00	0.2

Table 34. ResNet-50 with ℓ_2-normalization: Gradient $\times$ Input attribution compared to LRP-γ.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	5E-163	5.3
	$m = 50$	7E-161	6.5
	$m = 100$	1E-152	7.4
photometric	$m = 25$	1E+00	0.8
	$m = 50$	1E+00	0.6
	$m = 100$	1E+00	0.4
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	5E-159	4.5
	$m = 50$	2E-149	4.9
	$m = 100$	8E-117	4.6
photometric	$m = 25$	1E+00	0.6
	$m = 50$	1E+00	0.4
	$m = 100$	1E+00	0.3

Table 35. Swin-V2-Tiny with ℓ_2-normalization: Gradient $\times$ Input attribution compared to LRP-γ.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	1E+00	0.8
	$m = 50$	1E+00	0.7
	$m = 100$	1E+00	0.6
photometric	$m = 25$	1E+00	0.4
	$m = 50$	1E+00	0.3
	$m = 100$	1E+00	0.2
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	1E+00	0.8
	$m = 50$	1E+00	0.6
	$m = 100$	1E+00	0.5
photometric	$m = 25$	1E+00	0.4
	$m = 50$	1E+00	0.3
	$m = 100$	1E+00	0.2

Table 36. EfficientNet-V2-S with ℓ_2-normalization: Gradient $\times$ Input attribution compared to LRP-γ-lifted.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	9E-162	5.7
	$m = 50$	2E-156	5.8
	$m = 100$	2E-160	4.0
photometric	$m = 25$	1E-62	3.0
	$m = 50$	7E-28	2.3
	$m = 100$	1E-07	1.6
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	1E-154	4.9
	$m = 50$	2E-135	3.8
	$m = 100$	2E-140	2.4
photometric	$m = 25$	2E-19	2.3
	$m = 50$	8E-01	1.3
	$m = 100$	1E+00	0.8

Table 37. ResNet-50 with ℓ_2-normalization: Gradient $\times$ Input attribution compared to LRP-γ-lifted.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	2E-164	5.5
	$m = 50$	2E-163	6.9
	$m = 100$	4E-160	8.2
photometric	$m = 25$	9E-140	5.2
	$m = 50$	4E-131	5.4
	$m = 100$	8E-116	5.6
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	2E-164	5.5
	$m = 50$	2E-163	6.9
	$m = 100$	4E-160	8.2
photometric	$m = 25$	1E-139	5.2
	$m = 50$	5E-131	5.3
	$m = 100$	9E-116	5.5

Table 38. Swin-V2-Tiny with ℓ_2-normalization: Gradient $\times$ Input attribution compared to LRP-γ-lifted.

Augmentation	Sample size	p-value	ratio
Grad vs LRP-$\gamma = 10^3$			
Gaussian	$m = 25$	2E-165	8.1
	$m = 50$	3E-162	9.3
	$m = 100$	2E-161	10.1
photometric	$m = 25$	3E-136	5.8
	$m = 50$	1E-124	6.1
	$m = 100$	3E-121	6.3
Grad vs LRP-$\gamma = 10^2$			
Gaussian	$m = 25$	2E-165	8.1
	$m = 50$	3E-162	9.2
	$m = 100$	2E-161	9.9
photometric	$m = 25$	5E-135	4.8
	$m = 50$	1E-123	4.6
	$m = 100$	7E-117	4.1

22. Code changes for Lifted LRP-γ

```
class gammaconv_lifted(torch.nn.Module):

    def __init__(self, conv, ignorebias, gamma):
        super().__init__()
        self.gamma = gamma
        # computes convolution with only pos contributions
        # w_{+}x_{+} + w_{-}x_{-}
        self.pnconv=posnegconv(conv,ignorebias)
        self.conv=clone_module(conv)

        if ignorebias==True:
            self.conv.bias=None
        else:
            if conv.bias is not None:
                self.conv.bias=
                nn.Parameter( conv.bias.data.clone() )

    def forward(self,x):
        vp= self.pnconv(x)
        u= self.conv(x)

        # the new changes to compute gamma_{lift}
        epsthresh = 1e-10
        vpstab  = torch.where(vp>epsthresh, vp, epsthresh )
        mingamma = (1.0- u/vpstab)**2
        return vp + u/torch.maximum(self.gamma, mingamma)
        # return vp + u/self.gamma
```

23. Additional Faithfulness measurements

In order to complement the results in Fig. 1, we also measured faithfulness according to the definition laid out in the research of [12]. We performed it for the case of EfficientNet and SwinV2-Tiny in order to compare LRP-γ versus LRP using the autotuned lifted-γ. See Tab. 39 for results. The faithfulness is always higher for the lifted-γ variant. All t-test p-values are below $p = 3 \cdot 10^{-5}$. It should be noted that this particular measure of faithfulness puts more weight on regions with low attribution scores, simply because it samples regions uniformly, while the set of regions with top-p-percentiles of attribution scores is typically small for $p > 50\%$ for images from ImageNet, in which discriminative cues are sparse.

$\gamma =$	100	5	lifted-5	100	5	lifted-5
faithfulness	6e-3	3e-3	**1e-2**	3e-2	< 0	**4e-2**

Table 39. Faithfulness according to [12] (where higher is better differing from MoRF measures) for $\gamma = 10^2$, 5, lifted-5 for EffNet-V2-S (left) and SwinV2-Tiny (right). All t-test p-values are below $3 \cdot 10^{-5}$.