Unleashing the Power of Gradient Signal-to-Noise Ratio for Zero-Shot NAS – Supplementary Material –

Zihao Sun[1, 3 †], Yu Sun[2, 3 †], Longxing Yang[1, 3], Shun Lu[1, 3], Jilin Mei[1], Wenxiao Zhao[2, 3 *], Yu Hu[1, 3 *]
[1]Research Center for Intelligent Computing Systems,
Institute of Computing Technology, Chinese Academy of Sciences
[2]Key Laboratory of Systems and Control,
Academy of Mathematics and Systems Science, Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences
{sunzihao18z, yanglongxing20b, lushun19s, meijilin, huyu}@ict.ac.cn, {sunyu211}@mails.ucas.ac.cn, {wxzhao}@amss.ac.cn

A. Theoretical Proof

For the sake of proof, we assume that $f(\boldsymbol{x}, \theta) : \mathbb{R}^P \to \mathbb{R}$ is the neural network function with one-dimensional output. For any $\mathcal{D}$ and $f_{\mathcal{D}}(\theta)$, dataset and neural network function satisfy $||x_i|| \leq 1$, $|y_i| \leq M$, $|f_{\mathcal{D}}| \leq M$. As for the loss function, we use the mean squared error loss function $\mathcal{L}_{\mathcal{D}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(f(x_i, \theta) - y_i)^2$.

A.1. The Proof of Theorem 1

Theorem 1 $\forall 1 > \epsilon > 0$, $j = 1, ..., P$, $\exists M_1$ *such that with probability at least* $(1-\epsilon)$ *over randomly initialized parameters* θ^0*,*

$$\frac{M_1}{(16M^2 - M_1) + \frac{16M^2}{GSNR(\theta_j^0)}} \leq gsnr(\theta_j^0) \tag{1}$$

where M is the upper bound of the output.

Proof. For randomly initialized parameters θ_0, the performance of the neural network is probably not good. So there exist M_1 such that

$$P\left\{\mathbb{E}_{(x,y)\sim\mathcal{Z}}\left((f(x,\theta^0) - y)^2(\mathrm{g}')^2(x,\theta_j^0)\right) > M_1\left(\mathbb{E}_{(x,y)\sim\mathcal{Z}}(\mathrm{g}')^2(x,\theta_j)\right)\right\} > 1-\epsilon \tag{2}$$

over randomly initialized parameters θ^0. Because $\mathrm{g}(x,y,\theta_j) = \frac{\partial\mathcal{L}}{\partial f}(x,y,\theta)(\mathrm{g}')^2(x,\theta_j)$, it can be inferred that

$$P\left\{\mathbb{E}_{(x,y)\sim\mathcal{Z}}\mathrm{g}^2(x,y,\theta_j) > M_1(\mathbb{E}_{(x,y)\sim\mathcal{Z}}(\mathrm{g}')^2(x,\theta_j))\right\} > 1-\epsilon \tag{3}$$

and $|\frac{\partial\mathcal{L}}{\partial f}| = 2|f(x,\theta) - y| \leq 2|f(x,y,\theta)| + 2|y| \leq 4M$, then

$$GSNR(\theta_j^0) \tag{4}$$

$$= \frac{(\mathbb{E}_{(x,y)\sim\mathcal{Z}}\mathrm{g}(x,y,\theta_j^0))^2}{(\mathbb{E}_{(x,y)\sim\mathcal{Z}}\mathrm{g}^2(x,y,\theta_j^0)) - (\mathbb{E}_{(x,y)\sim\mathcal{Z}}\mathrm{g}(x,y,\theta_j^0))^2} \tag{5}$$

$$= \frac{(\mathbb{E}_{(x,y)\sim\mathcal{Z}}(\frac{\partial\mathcal{L}}{\partial f})\mathrm{g}'(x,\theta_j^0))^2}{(\mathbb{E}_{(x,y)\sim\mathcal{Z}}\mathrm{g}^2(x,y,\theta_j^0)) - (\mathbb{E}_{(x,y)\sim\mathcal{Z}}(\frac{\partial\mathcal{L}}{\partial f})\mathrm{g}'(x,\theta_j^0))^2} \tag{6}$$

$$\tag{7}$$

[†]Equal contribution. [*]Corresponding authors.

$$\leq \frac{16M^2(\mathbb{E}_{(x,y)\sim\mathcal{Z}}\mathrm{g}'(x,\theta_j^0))^2}{M_1(\mathbb{E}_{(x,y)\sim\mathcal{Z}}(\mathrm{g}')^2(x,\theta_j^0)) - M_2(\mathbb{E}_{(x,y)\sim\mathcal{Z}}\mathrm{g}'(x,\theta_j^0))^2} \tag{8}$$

$$= \frac{16M^2}{M_1\frac{(\mathbb{E}_{(x,y)\sim\mathcal{Z}}(\mathrm{g}')^2(x,\theta_j^0))}{(\mathbb{E}_{(x,y)\sim\mathcal{Z}}\mathrm{g}'(x,\theta_j^0))^2} - M_2} \tag{9}$$

$$= \frac{16M^2}{-(M_2 - M_1) + \frac{M_1}{gsnr(\theta_j^0)}} \tag{10}$$

where M_2 is a contstant for fixed θ_0 satisfying $16M^2 \geq M_2 \geq M_1$ and $(\mathbb{E}_{(x,y)\sim\mathcal{Z}}\mathrm{g}^2(x,y,\theta_j^0)) - (\mathbb{E}_{(x,y)\sim\mathcal{Z}}(\frac{\partial\mathcal{L}}{\partial f})\mathrm{g}'(x,\theta_j^0))^2 > M_1(\mathbb{E}_{(x,y)\sim\mathcal{Z}}(\mathrm{g}')^2(x,\theta_j^0)) - M_2(\mathbb{E}_{(x,y)\sim\mathcal{Z}}\mathrm{g}'(x,\theta_j^0))^2 > 0$. Such constant exists because

$$(\mathbb{E}_{(x,y)\sim\mathcal{Z}}\mathrm{g}^2(x,y,\theta_j^0)) - \left(\mathbb{E}_{(x,y)\sim\mathcal{Z}}(\frac{\partial\mathcal{L}}{\partial f})\mathrm{g}'(x,\theta_j^0)\right)^2 > 0 \tag{11}$$

$$(\mathbb{E}_{(x,y)\sim\mathcal{Z}}\mathrm{g}^2(x,y,\theta_j^0)) - \left(\mathbb{E}_{(x,y)\sim\mathcal{Z}}(\frac{\partial\mathcal{L}}{\partial f})\mathrm{g}'(x,\theta_j^0)\right)^2 > M_1(\mathbb{E}_{(x,y)\sim\mathcal{Z}}(\mathrm{g}')^2(x,\theta_j^0)) - 16M^2(\mathbb{E}_{(x,y)\sim\mathcal{Z}}\mathrm{g}'(x,\theta_j^0))^2 \tag{12}$$

$$\mathbb{E}_{(x,y)\sim\mathcal{Z}}\left((\frac{\partial\mathcal{L}}{\partial f})^2(\mathrm{g}')^2(x,\theta_j^0)\right) > M_1\left(\mathbb{E}_{(x,y)\sim\mathcal{Z}}(\mathrm{g}')^2(x,\theta_j)\right) \tag{13}$$

it implies

$$gsnr(\theta_j^0) \tag{14}$$

$$\geq \frac{M_1}{(M_2 - M_1) + \frac{16M^2}{GSNR(\theta_j^0)}} \tag{15}$$

$$\geq \frac{M_1}{(16M^2 - M_1) + \frac{16M^2}{GSNR(\theta_j^0)}} \tag{16}$$

□

A.2. The Proof of Theorem 2

Theorem 2 *Under Assumption 1, for fixed initialization parameters θ^0, if $\nabla_\theta^2\mathcal{L}_\mathcal{D}(\theta^t)$ is semi-positive definite matrix, $\mathbb{E}_{(x,y)\sim\mathcal{Z}}|(f(x,\theta^t) - y)|$ is small enough, $\forall t = 1, 2...$, there exist $0 < \alpha_t < 1$ and $\frac{1}{\sqrt{n\alpha_t gsnr(\theta_j^0)}} < r < 1, j = 1, 2...P$, such that ,*

$$\lambda_{max}(\nabla_\theta^2\mathcal{L}_{\mathcal{D}'}(\theta^t)) \leq \frac{n(1+r)^2}{(1-r)^2}\lambda_{max}(\nabla_\theta^2\mathcal{L}_\mathcal{D}(\theta^t)) \tag{17}$$

with probability at least

$$1 - \sum_{j=1}^{P}\frac{2n}{r^2\alpha_t gsnr(\theta_j^0)} \tag{18}$$

over randomly chosen possible distributions for all training sets $\mathcal{D}$ and validation sets $\mathcal{D}'$ which have the same number of data.

Proof. For the Hessian matrix of the loss function,

$$\nabla_\theta^2\mathcal{L}_\mathcal{D}(\theta^t) \tag{19}$$

$$= \nabla_\theta^2\frac{1}{n}\sum_{i=1}^{n}(f(x_i,\theta^t) - y_i)^2 \tag{20}$$

$$= \frac{2}{n}\sum_{i=1}^{n}(\nabla_\theta f(x_i,\theta^t))(\nabla_\theta f(x_i,\theta^t))^T + \frac{2}{n}\sum_{i=1}^{n}(f(x_i,\theta^t) - y_i)(\nabla_\theta^2 f(x_i,\theta^t)) \tag{21}$$

Within the bounded area $||\nabla^2_\theta f(x_i, \theta^t)||_F < M_3$, when $\frac{2}{n}\sum_{i=1}^{n} |(f(x_i, \theta^t) - y_i)|$ is small enough, we have

$$||\frac{2}{n}\sum_{i=1}^{n}(f(x_i, \theta^t) - y_i)(\nabla^2_\theta f(x_i, \theta^t))||_F \tag{22}$$

$$\leq \frac{2}{n}\sum_{i=1}^{n} |(f(x_i, \theta^t) - y_i)|||(\nabla^2_\theta f(x_i, \theta^t))||_F \tag{23}$$

$$\leq M_3 \frac{2}{n}\sum_{i=1}^{n} |(f(x_i, \theta^t) - y_i)| \to 0 \tag{24}$$

So

$$\nabla^2_\theta \mathcal{L}_\mathcal{D}(\theta^t) \approx \frac{2}{n}\sum_{i=1}^{n}(\nabla_\theta f(x_i, \theta^t))(\nabla_\theta f(x_i, \theta^t))^T \tag{25}$$

The maximum eigenvalue is

$$\lambda_{max}(\nabla^2_\theta \mathcal{L}_\mathcal{D}(\theta^t)) \tag{26}$$

$$\approx \lambda_{max}\left(\frac{2}{n}\sum_{i=1}^{n}(\nabla_\theta f(x_i, \theta^t))(\nabla_\theta f(x_i, \theta^t))^T\right) \tag{27}$$

$$\geq \bar{\lambda}\left(\frac{2}{n}\sum_{i=1}^{n}(\nabla_\theta f(x_i, \theta^t))(\nabla_\theta f(x_i, \theta^t))^T\right) \tag{28}$$

$$= \frac{2}{n^2} tr\left(\sum_{i=1}^{n}(\nabla_\theta f(x_i, \theta^t))(\nabla_\theta f(x_i, \theta^t))^T\right) \tag{29}$$

$$= \frac{2}{n^2}\sum_{i=1}^{n} ||\nabla_\theta f(x_i, \theta^t)||^2 \tag{30}$$

$$= \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{P}(g')^2(x_i, \theta^t_j) \tag{31}$$

where $tr(\cdot)$ means the trace of matrix, $\bar{\lambda}(\cdot)$ means average eigenvalue, $\nabla^2_\theta \mathcal{L}_\mathcal{D}(\theta^t)$ need to be a semi-positive definite matrix. Also,

$$\lambda_{max}(\nabla^2_\theta \mathcal{L}_\mathcal{D}(\theta^t)) \tag{32}$$

$$\approx \lambda_{max}(\frac{2}{n}\sum_{i=1}^{n}(\nabla_\theta f(x_i, \theta^t))(\nabla_\theta f(x_i, \theta^t))^T) \tag{33}$$

$$\leq \frac{2}{n}||\sum_{i=1}^{n}(\nabla_\theta f(x_i, \theta^t))(\nabla_\theta f(x_i, \theta))^T||_F \tag{34}$$

$$\leq \frac{2}{n}\sum_{i=1}^{n} ||(\nabla_\theta f(x_i, \theta^t))(\nabla_\theta f(x_i, \theta^t))^T||_F \tag{35}$$

$$= \frac{2}{n}\sum_{i=1}^{n} ||\nabla_\theta f(x_i, \theta^t)||^2 \tag{36}$$

$$= \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{P}(g')^2(x_i, \theta^t_j) \tag{37}$$

For any two independent datasets $\mathcal{D} = \{(x_i, y_i)_{i=1}^n\} \sim \mathcal{Z}^n$, $\mathcal{D}' = \{(x'_i, y'_i)_{i=1}^n\} \sim \mathcal{Z}^n$. For any x, by chebyshev's theorem,

$$P\{|g'(x, \theta^t_j) - \mathbb{E}_{(x,y)\sim\mathcal{Z}} g'(x, \theta^t_j)| \leq \delta\} \geq 1 - \frac{Var_{(x,y)\sim\mathcal{Z}} g'(x, \theta^t_j)}{\delta^2} \tag{38}$$

By letting $\delta = r|\mathbb{E}_{(x,y)\sim\mathcal{Z}} g'(x, \theta_j^t)|$

$$P\{|g'(x, \theta_j^t) - \mathbb{E}_{(x,y)\sim\mathcal{Z}} g'(x, \theta_j^t)| \leq r|\mathbb{E}_{(x,y)\sim\mathcal{Z}} g'(x, \theta_j^t)|\} \geq 1 - \frac{1}{r^2 gsnr(\theta_j^t)} \tag{39}$$

Through independence, with probability at least $(\prod_{j=1}^{P}(1 - \frac{1}{r^2 gsnr(\theta_j^t)}))^{2n}$, for $i = 1, ..., n$, $j = 1, ..., P$,

$$\begin{aligned}
(1-r)^2(\mathbb{E}_{(x,y)\sim\mathcal{Z}} g'(x, \theta_j^t))^2 &\leq (g')^2(x_i', \theta_j^t) && (40)\\
&\leq (1+r)^2(\mathbb{E}_{(x,y)\sim\mathcal{Z}} g'(x, \theta_j^t))^2 && (41)\\
(1-r)^2(\mathbb{E}_{(x,y)\sim\mathcal{Z}} g'(x, \theta_j^t))^2 &\leq (g')^2(x_i', \theta_j^t) && (42)\\
&\leq (1+r)^2(\mathbb{E}_{(x,y)\sim\mathcal{Z}} g'(x, \theta_j^t))^2 && (43)\\
\Rightarrow (g')^2(x_i, \theta_j^t) &\leq \frac{(1+r)^2}{(1-r)^2}(g')^2(x_i', \theta_j^t) && (44)
\end{aligned}$$

Then

$$\begin{aligned}
&\lambda_{max}(\nabla_\theta^2 \mathcal{L}_{\mathcal{D}'}(\theta^t)) && (45)\\
&\leq \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{P}(g')^2(x_i', \theta_j^t) && (46)\\
&\leq \frac{2}{n}\frac{(1+r)^2}{(1-r)^2}\sum_{i=1}^{n}\sum_{j=1}^{P}(g')^2(x_i, \theta_j^t) && (47)\\
&\leq \frac{n(1+r)^2}{(1-r)^2}\lambda_{max}(\nabla_\theta^2 \mathcal{L}_{\mathcal{D}}(\theta^t)) && (48)
\end{aligned}$$

$\forall 1 > \epsilon_1 > 0$, $\mathbb{E}_{(x,y)\sim\mathcal{Z}}|(f(x, \theta^t) - y)|$ is small enough means with probability at least $(1 - \epsilon_1)$, $\frac{2}{n}\sum_{i=1}^{n}|(f(x_i, \theta) - y_i)|$ and $\frac{2}{n}\sum_{i=1}^{n}|(f(x_i', \theta) - y_i')|$ is small enough. By the continuity of $gsnr(\theta_j), j = 1, ..., P$, there exists a neighborhood $B(\theta^0)$ of θ^0 such that $\theta^1, ..., \theta^t \in B(\theta^0)$, α_t be the constant satisfing

$$gsnr(\theta_j) \geq \alpha_t gsnr(\theta_j^0),\ j = 1, ..., P \tag{49}$$

$\forall \theta \in B(\theta^0)$. When $\epsilon_1 \to 0$, $(1 - \epsilon_1)(\prod_{j=1}^{P}(1 - \frac{1}{r^2 gsnr(\theta_j)}))^{2n}$ will be larger than $1 - \sum_{j=1}^{P}\frac{2n}{r^2\alpha_t gsnr(\theta_j^0)}$. So with probability at least $1 - \sum_{j=1}^{P}\frac{2n}{r^2\alpha_t gsnr(\theta_j^0)}$ over randomly chosen possible distributions for all training sets $\mathcal{D}$ and validation sets $\mathcal{D}'$,

$$\lambda_{max}(\nabla_\theta^2 \mathcal{L}_{\mathcal{D}'}(\theta^t)) \leq \frac{n(1+r)^2}{(1-r)^2}\lambda_{max}(\nabla_\theta^2 \mathcal{L}_{\mathcal{D}}(\theta^t)) \tag{50}$$

□

A.3. The Proof of Theorem 3

Theorem 3 *Under Assumption 1, for fixed initialization parameters θ^0, if the learning rate η is small enough, $\forall t = 1, 2...$, there exist $0 < \alpha_t < 1$ and $\frac{1}{\sqrt{n\alpha_t gsnr(\theta_j^0)}} < r < 1, j = 1, 2...P$, such that,*

$$\mathcal{L}_{\mathcal{D}}(\theta^{t+1}) - \mathcal{L}_{\mathcal{D}}(\theta^t) < -\eta\alpha_t(1-r)^2(\frac{\partial \mathcal{L}_{\mathcal{D}}}{\partial f_{\mathcal{D}}}(\theta^t))^2 \mathbb{E}_{\mathcal{D}\sim\mathcal{Z}^n}(\sum_{j=1}^{P}(\mathrm{g}_{\mathcal{D}}'(\theta_j^0))^2)$$

with probability at least

$$1 - \sum_{j=1}^{P}\frac{1}{nr^2\alpha_t gsnr(\theta_j^0)} \tag{51}$$

over randomly chosen possible distributions for all training sets $\mathcal{D}$.

Proof. Because the learning rate λ is small enough,

$$\mathcal{L}_{\mathcal{D}}(\theta^{t+1}) - \mathcal{L}_{\mathcal{D}}(\theta^{t}) \tag{52}$$

$$= -\eta(\nabla_{\theta}\mathcal{L}_{\mathcal{D}}(\theta^{t}))^{T}(\nabla_{\theta}\mathcal{L}_{\mathcal{D}}(\theta^{t})) + O(\lambda^2) \tag{53}$$

$$= -\lambda(\frac{\partial \mathcal{L}_{\mathcal{D}}}{\partial f_{\mathcal{D}}}(\theta^{t}))^2(\nabla_{\theta} f_{\mathcal{D}}(\theta^{t}))^{T}(\nabla_{\theta} f_{\mathcal{D}}(\theta^{t})) + O(\lambda^2) \tag{54}$$

$$= -\lambda(\frac{\partial \mathcal{L}_{\mathcal{D}}}{\partial f_{\mathcal{D}}}(\theta^{t}))^2||\nabla_{\theta} f_{\mathcal{D}}(\theta^{t})||^2 + O(\lambda^2) \tag{55}$$

$$= -\lambda(\frac{\partial \mathcal{L}_{\mathcal{D}}}{\partial f_{\mathcal{D}}}(\theta^{t}))^2(\sum_{j=1}^{P}(\mathrm{g}'_{\mathcal{D}}(\theta^{t}_{j}))^2) + O(\lambda^2) \tag{56}$$

by chebyshev's theorem,

$$P\{|\mathrm{g}'_{\mathcal{D}}(\theta^{t}_{j}) - \mathbb{E}_{\mathcal{D}\sim\mathcal{Z}^n}\mathrm{g}'_{\mathcal{D}}(\theta^{t}_{j})| \le \delta\} \ge 1 - \frac{Var_{\mathcal{D}\sim\mathcal{Z}^n}\mathrm{g}'_{\mathcal{D}}(\theta^{t}_{j})}{\delta^2} \tag{57}$$

Let $\delta = |\mathbb{E}_{\mathcal{D}\sim\mathcal{Z}^n}\mathrm{g}'_{\mathcal{D}}(\theta^{t}_{j})|r$, then

$$P\{|\mathrm{g}'_{\mathcal{D}}(\theta^{t}_{j})| \ge |\mathbb{E}_{\mathcal{D}\sim\mathcal{Z}^n}\mathrm{g}'_{\mathcal{D}}(\theta^{t}_{j})|(1-r)\} \ge 1 - \frac{1}{r^2\frac{(\mathbb{E}_{\mathcal{D}\sim\mathcal{Z}^n}\mathrm{g}'_{\mathcal{D}}(\theta^{t}_{j}))^2}{Var_{\mathcal{D}\sim\mathcal{Z}^n}\mathrm{g}'_{\mathcal{D}}(\theta^{t}_{j})}} \tag{58}$$

Because

$$\frac{(\mathbb{E}_{\mathcal{D}\sim\mathcal{Z}^n}\mathrm{g}'_{\mathcal{D}}(\theta_j))^2}{Var_{\mathcal{D}\sim\mathcal{Z}^n}\mathrm{g}'_{\mathcal{D}}(\theta_j)} \tag{59}$$

$$= \frac{(\mathbb{E}_{(x,y)\sim\mathcal{Z}}\mathrm{g}'_{\mathcal{D}}(x,\theta_j))^2}{\frac{1}{n}Var_{(x,y)\sim\mathcal{Z}}\mathrm{g}'(x,\theta_j)} \tag{60}$$

$$= n\ gsnr(\theta_j) \tag{61}$$

we can deduce that

$$P\{|\mathrm{g}'_{\mathcal{D}}(\theta^{t}_{j})|^2 \ge |\mathbb{E}_{\mathcal{D}\sim\mathcal{Z}^n}\mathrm{g}'_{\mathcal{D}}(\theta^{t}_{j})|^2(1-r)^2\} \ge 1 - \frac{1}{nr^2 gsnr(\theta^{t}_{j})} \tag{62}$$

So

$$P\left\{\left(\sum_{j=1}^{P}(\mathrm{g}'_{\mathcal{D}}(\theta^{t}_{j}))^2\right) \ge \mathbb{E}_{\mathcal{D}\sim\mathcal{Z}^n}\left(\sum_{j=1}^{P}(\mathrm{g}'_{\mathcal{D}}(\theta^{t}_{j}))^2\right)(1-r)^2\right\} \ge 1 - \sum_{j=1}^{P}\frac{1}{nr^2 gsnr(\theta^{t}_{j})} \tag{63}$$

By the continuity of $gsnr(\theta_j), j = 1, ..., P$ and $\mathbb{E}_{\mathcal{D}\sim\mathcal{Z}^n}\nabla_{\theta} f_{\mathcal{D}}(\theta)$, there exists a neighborhood $B(\theta^0)$ of θ^0 such that $\theta^1, ..., \theta^t \in B(\theta^0)$, α_t be the constant satisfing

$$gsnr(\theta_j) \ge \alpha_t gsnr(\theta^{0}_{j}),\ j = 1, ..., P \tag{64}$$

$$\mathbb{E}_{\mathcal{D}\sim\mathcal{Z}^n}\left(\sum_{j=1}^{P}(\mathrm{g}'_{\mathcal{D}}(\theta_j))^2\right) \ge \alpha_t\mathbb{E}_{\mathcal{D}\sim\mathcal{Z}^n}\left(\sum_{j=1}^{P}(\mathrm{g}'_{\mathcal{D}}(\theta^{0}_{j}))^2\right) \tag{65}$$

$\forall\theta \in B(\theta^0)$. Combine (63),

$$P\left\{\left(\sum_{j=1}^{P}(\mathrm{g}'_{\mathcal{D}}(\theta^{t}_{j}))^2\right) \ge \alpha_t(1-r)^2\mathbb{E}_{\mathcal{D}\sim\mathcal{Z}^n}\left(\sum_{j=1}^{P}(\mathrm{g}'_{\mathcal{D}}(\theta^{0}_{j}))^2\right)\right\} \tag{66}$$

$$\ge \prod_{j=1}^{P}\left(1 - \frac{1}{nr^2\alpha_t gsnr(\theta^{0}_{j})}\right) \tag{67}$$

$$\ge 1 - \sum_{j=1}^{P}\frac{1}{nr^2\alpha_t gsnr(\theta^{0}_{j})} \tag{68}$$

So, with probability at least $1 - \sum_{j=1}^{P} \frac{1}{nr^2\alpha_t gsnr(\theta_j^0)}$ over randomly choosed possible distributions for all training sets $\mathcal{D}$,

$$\mathcal{L}_{\mathcal{D}}(\theta^{t+1}) - \mathcal{L}_{\mathcal{D}}(\theta^t) < -\lambda\alpha_t(1-r)^2(\frac{\partial \mathcal{L}_{\mathcal{D}}}{\partial f_{\mathcal{D}}}(\theta^t))^2\mathbb{E}_{\mathcal{D}\sim\mathcal{Z}^n}(\sum_{j=1}^{P}(\mathrm{g}'_{\mathcal{D}}(\theta_j^0))^2) \tag{69}$$

□

B. Search Algorithm

We adopt different search algorithms to verify the effectiveness of ξ-*GSNR*. Specifically, the pruning-based Zero-Shot NAS algorithm (Algorithm 1) for NAS-Bench-201 [4] and DARTS [6] search space, and the evolutionary Zero-Shot NAS algorithm (Algorithm 2) for ProxylessNAS [2] and BurgerFormer [8] search space.

Algorithm 1: Pruning-based Zero-Shot NAS via ξ-*GSNR*

```
Input: Search Space $\mathcal{S}$; Supernet $\mathcal{N}_0$ stacked by cells, each cell has $E$ edges, each edge has $|O|$ operators, step $t = 0$;
       Batch Size $\mathcal{B}$, Batch Numbers $N$, random $\xi$.
Output: Optimal Architecture.
Initialize parameters $\theta$ of Supernet $\mathcal{N}_0$ stacked by cells;
while $\mathcal{N}_t$ is not a single-path network do
    for each operation $o_m$ in $\mathcal{N}_t$ do
        Compute $\xi$-GSNR_1 of $\mathcal{N}_t$ by Eq.(10);
        Compute $\xi$-GSNR_2 of $\mathcal{N}_t \setminus_{o_m}$ by Eq.(10);
        Get score $s(o_m) = \xi$-GSNR_1 $-$ $\xi$-GSNR_2
    $\mathcal{N}_{t+1} = \mathcal{N}_t$
    for each edge $e_n, n = 1, 2...E$ do
        $m^* = \mathrm{argmin}\{s(o_m) : o_m \in e_n\}$;
        $\mathcal{N}_{t+1} = \mathcal{N}_{t+1} \setminus_{m^*}$
    $t = t + 1$
Return Optimal Architecture
```

Algorithm 2: Evolutionary Zero-Shot NAS via ξ-*GSNR*

```
Input: Search Space $\mathcal{S}$, Inference budget constraints $C$, search iterations $T$, population size $P$; Batch Size $\mathcal{B}$, Batch
       Numbers $N$, random $\xi$.
Output: Optimal Architecture.
Initialize population $P = \{A_0, A_1, ..., A_P\}$ that meets the inference budget constraints $C$;
Compute $\xi$-GSNR score $S = \{S_1, S_2, ..., S_P\}$ in $P$ by Eq.(10);
for $t = 1, 2, ..., T$ do
    Randomly sample $A_t \in P$;
    $\tilde{A}_t$ = randomly mutate architecture $A_t$ that meets the inference budget constraints $C$;
    Compute $\xi$-GSNR score $S_{\tilde{A}_t}$ of $\tilde{A}_t$ by Eq.(10);
    if $S_{\tilde{A}_t} > \min\{S_1, S_2, ..., S_P\}$ then
        Remove the architecture with the lowest $\xi$-GSNR score in $P$;
        Add $\tilde{A}_t$ to population $P$;
    else
        Skip $S_{\tilde{A}_t}$
$A^*$ = the architecture with the highest $\xi$-GSNR score in $P$
Return Optimal Architecture $A^*$
```

C. Experimental Setup

C.1. Details of Ranking Consistency Experiments

We conduct the ranking consistency experiments in NAS-Bench-201, NAS-Bench-101, and NDS search space strictly following NASWOT [7] experimental settings. To compute ξ-*GSNR* proxy score, we set the batch size to 64, the number of batches to 8, with ξ=1e-8 in NAS-Bench-201 and NDS search space, and the batch size to 1, the number of batches to 8, with ξ=1e-8 in NAS-Bench-101. All the other baseline methods are based on public code [1] with a batch size of 64 in NAS-Bench-101 and NAS-Bench-201, and a batch size of 128 in NDS search space.

C.2. Details of Searching Experiments

NAS-Bench-201. To directly search for the optimal architecture on different datasets, including CIFAR-10, CIFAR-100, and ImageNet-16-120, we use the Algorithm 1 following the experimental settings same as TE-NAS [3]. We set the batch size to 64, and the number of batches to 8, with ξ=1e-8, which are the same as the ranking consistency experimental settings. The validation accuracy and test accuracy of the searched architecture can be directly indexed on the benchmark.

DARTS Space. Based on Algorithm 1, we search for the optimal normal cell and reduction cell directly on CIFAR-10 and ImageNet. We set the batch size to 4, and the number of batches to 2, with ξ=1e-8 for balancing the efficiency and effectiveness.

To evaluate the performance of the searched architecture, we follow DARTS [6] settings to retrain the target model. On CIFAR-10, we train the target network consisting of 20 cells with an initial channel size of 36 on the whole training dataset from scratch. We use an SGD optimizer with an initial learning rate starting from 0.025 that follows the cosine annealing strategy to a minimum of 0, and with a weight decay of 3×10^{-4} and a momentum of 0.9. The network is trained for 600 epochs with a batch size of 96. On ImageNet, the target network consists of 14 cells with 48 initial channels. We use the SGD optimizer with an initial learning rate of 0.5, weight decay of 3×10^{-5}, and momentum of 0.9. The network is trained for 250 epochs with a batch size of 1024.

ProxylessNAS Space. The search space contains about 6^{19} different networks. We utilize Algorithm 2 to search for the optimal architecture directly on ImageNet. The inference budget is limited to around 400M FLOPs. The search iterations T are 2000 with a population size P of 128. To compute the ξ-*GSNR* score, the batch size is set to 64, and the number of batches is set to 2, with ξ=1e-8.

To evaluate the performance of the searched architecture, we follow DNA [5] experimental settings to retrain the model. The target model has Squeeze-Excitation (SE) attention and Swish activation. We retrain the model with a batch size of 1024 for 500 epochs. We use the RMSprop optimizer with a momentum of 0.9, an initial learning rate of 0.064 that is decayed by 0.963 every 3 epochs.

BurgerFormer Space. The search space includes a large number of ViT-Like structures by a micro-meso-macro design. We utilize Algorithm 2 to search for the optimal architecture directly on ImageNet. The search iteration T is 2000 with a population of 128. We limit the resource budget to obtain architectures with similar FLOPs or parameters to baselines.

To evaluate the performance of the searched architecture, we follow the training pipeline of BurgerFormer. The searched network is retrained using AdamW with a learning rate of 1×10^{-3}, weight decay of 0.05, and batch size of 1024. Data augmentations include MixUp, CutMix, CutOut and RandAugment. The training epochs are 300 with 10 epochs warmup.

D. The variants of GSNR

In addition to ImageNet-16-120, we also compare the different variants of *GSNR* on CIFAR-10 and CIFAR-100 datasets in NAS-Bench-201. We randomly sample 200 architectures and then compute the proxy score by leveraging the variants of *GSNR* listed in Tab.12. We set the batch size to 64, the number of batches to 8 across all variant proxies, as well as ξ=1e-8. We can see that (4) *GSNR* achieves better ranking consistency than either (1) the gradient's squared mean and (2) the inverse of gradient's variance on different datasets. Moreover, a small ξ added to the variance of (3) can improve the performance over the simple gradient's variance. Therefore, we develop an efficient Zero-Shot proxy by adding the ξ to the standard *GSNR* as shown in (5). As a result, our ξ-*GSNR* proxy achieves the best ranking.

E. Visualization

Ranking Results. We visualize the architecture's test accuracy and its corresponding ξ-*GSNR* proxy score in NAS-Bench-201 on three different datasets as shown in Fig.4. The ranking consistency on CIFAR-10 is 0.845 of Spearman's ρ and 0.661

of Kendall's τ, respectively. The ranking consistency on CIFAR-100 is 0.840 of Spearman's ρ and 0.658 of Kendall's τ, respectively. The ranking consistency on ImageNet-16-120 is 0.793 of Spearman's ρ and 0.608 of Kendall's τ, respectively.

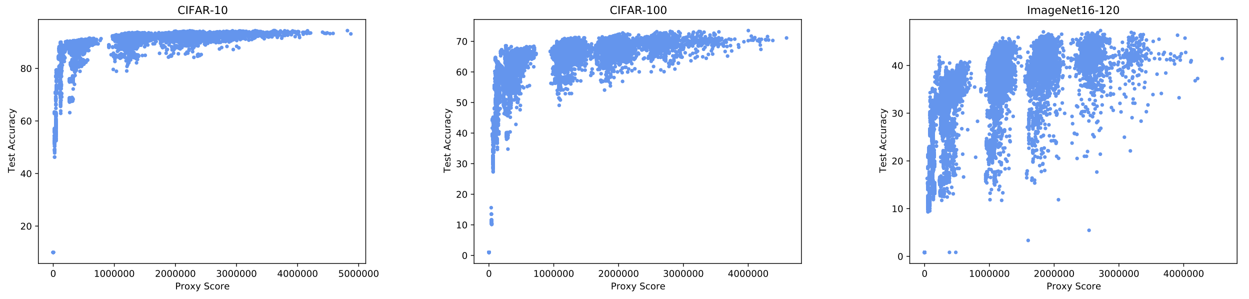
Architectures in DARTS Space. We visualize the searched normal cells and reduction cells directly on CIFAR-10 from three independent experiments with different random seeds. As shown in Fig.5, the test error and parameters of the network based on (a)(b) are 2.47% and 3.66M respectively; the test error and parameters of the network based on (c)(d) are 2.54% and 3.06M respectively; the test error and parameters of the network based on (e)(f) are 2.56% and 3.64M respectively. In addition, we visualize the searched normal cell and reduction cell directly on ImageNet in Fig.6. The architecture achieves Top-1 accuracy of 75.5% and Top-5 accuracy of 92.5%.

Architectures in ProxylessNAS Space. We visualize the chain-style architecture searched directly on ImageNet in Fig.7. Each layer contains an MBConv block with a different kernel size and expansion ratio. The parameters and FLOPs are 5.4M and 409M, respectively. We achieve the state-of-the-art performance of 78.2% Top-1 accuracy.

Architectures in BurgerFormer Space. We visualize the searched ViT-Like architecture in Fig.8. The architecture with 29M Params and 4.5G FLOPs obtains 83.1% Top-1 accuracy on ImageNet dataset.

	CIFAR-10		**CIFAR-100**	
	Spearman's ρ	**Kendall's τ**	**Spearman's ρ**	**Kendall's τ**
(1) $\sum_{j=1}^{P}(\mathbb{E}(\text{g}(x,y,\theta_j)))^2$	0.423	0.316	0.533	0.387
(2) $\sum_{j=1}^{P}\frac{1}{Var(\text{g}(x,y,\theta_j))}$	0.250	0.277	0.212	0.244
(3) $\sum_{j=1}^{P}\frac{1}{Var(\text{g}(x,y,\theta_j))+\xi}$	0.256	0.289	0.218	0.256
(4) $\sum_{j=1}^{P}\frac{(\mathbb{E}(\text{g}(x,y,\theta_j)))^2}{Var(\text{g}(x,y,\theta_j))}$	0.763	0.595	0.748	0.574
(5) $\sum_{j=1}^{P}\frac{(\mathbb{E}(\text{g}(x,y,\theta_j)))^2}{Var(\text{g}(x,y,\theta_j))+\xi}$	**0.871**	**0.696**	**0.854**	**0.683**

Table 12. Comparison Ranking Consistency of GSNR proxy variants on CIFAR-10 and CIFAR-100 in NAS-Bench-201 space.

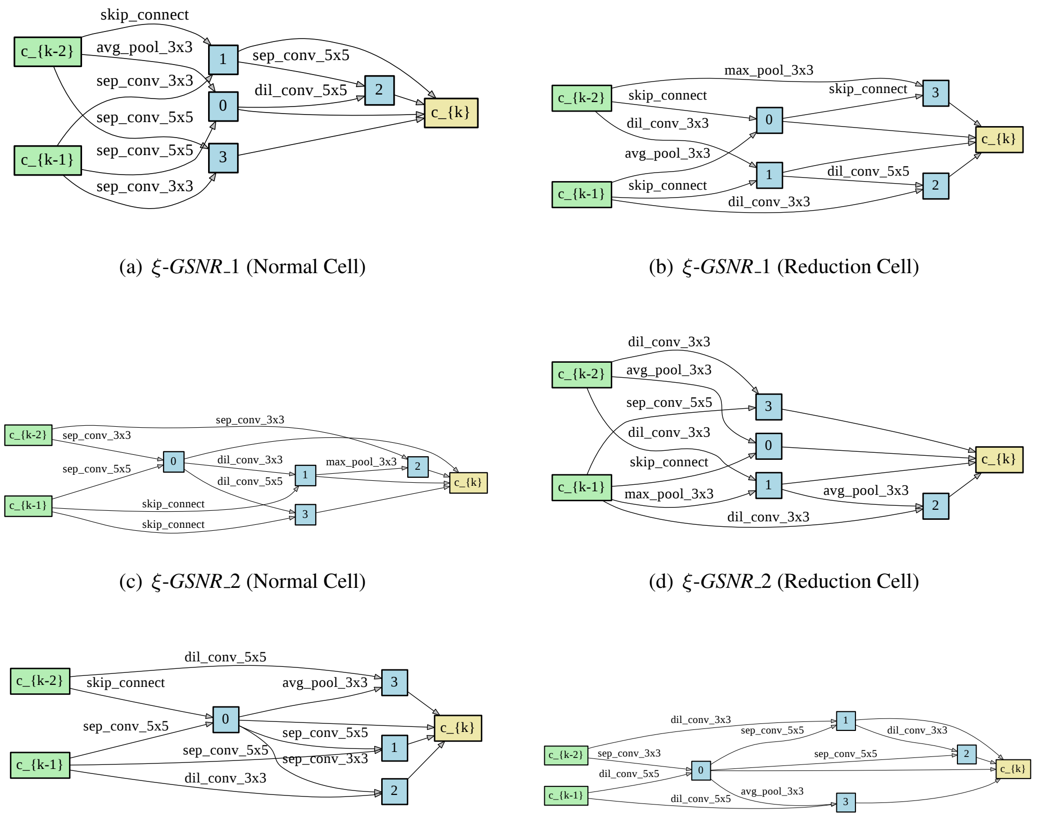


(a) CIFAR-10

(b) CIFAR-100

(c) ImageNet-16-120

Figure 4. Visualization of the test accuracy vs. ξ-*GSNR* proxy score in NAS-Bench-201 on different datasets, including (a) CIFAR-10, (b) CIFAR-100, and (c) ImageNet-16-120.



(a) ξ-*GSNR*_1 (Normal Cell)

(b) ξ-*GSNR*_1 (Reduction Cell)

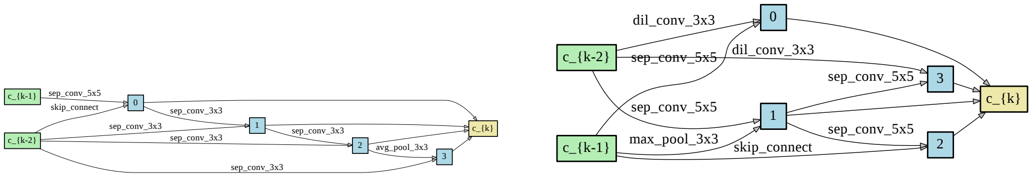
(c) ξ-*GSNR*_2 (Normal Cell)

(d) ξ-*GSNR*_2 (Reduction Cell)

(e) ξ-*GSNR*_3 (Normal Cell)

(f) ξ-*GSNR*_3 (Reduction Cell)

Figure 5. Visualization of the searched normal cells and reduction cells directly on CIFAR-10 in DARTS search space.



(a) ξ-*GSNR* (Normal Cell)

(b) ξ-*GSNR* (Reduction Cell)

Figure 6. Visualization of the searched normal cell and reduction cell directly on ImageNet in DARTS search space.

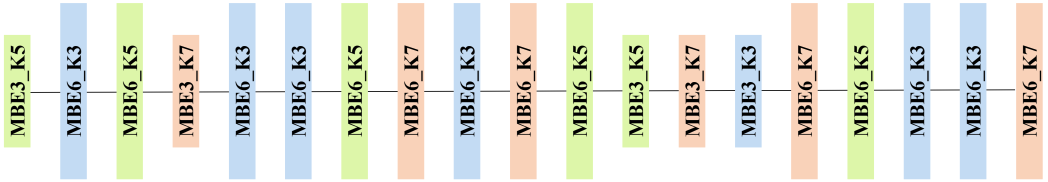


Figure 7. Visualization of the searched architecture directly on ImageNet in Proxyless search space. “MB” denotes MobileNetV2 block; “E” denotes expansion ratio; “K” denotes kernel size.

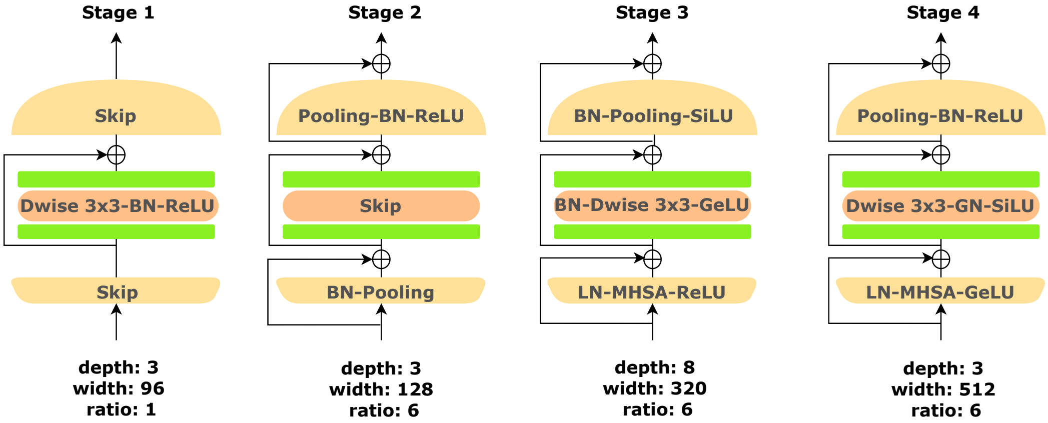


Figure 8. Visualization of the searched architecture directly on ImageNet in BurgerFormer search space.

References

[1] Mohamed S Abdelfattah, Abhinav Mehrotra, Łukasz Dudziak, and Nicholas Donald Lane. Zero-cost proxies for lightweight nas. In *ICLR*, 2020. 7

[2] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. In *ICLR*, 2018. 6

[3] Wuyang Chen, Xinyu Gong, and Zhangyang Wang. Neural architecture search on imagenet in four gpu hours: A theoretically inspired perspective. In *ICLR*, 2021. 7

[4] Xuanyi Dong and Yi Yang. Nas-bench-201: Extending the scope of reproducible neural architecture search. In *ICLR*, 2020. 6

[5] Changlin Li, Jiefeng Peng, Liuchun Yuan, Guangrun Wang, Xiaodan Liang, Liang Lin, and Xiaojun Chang. Block-wisely supervised neural architecture search with knowledge distillation. In *CVPR*, 2020. 7

[6] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. In *ICLR*, 2018. 6, 7

[7] Joe Mellor, Jack Turner, Amos Storkey, and Elliot J Crowley. Neural architecture search without training. In *ICML*, 2021. 7

[8] Longxing Yang, Yu Hu, Shun Lu, Zihao Sun, Jilin Mei, Yinhe Han, and Xiaowei Li. Searching for burgerformer with micro-meso-macro space design. In *ICML*, 2022. 6