Neural Architecture Search in A Proxy Validation Loss Landscape (Supplementary Material)

Yanxi Li [1] **Minjing Dong** [1] **Yunhe Wang** [2] **Chang Xu** [1]

1. Proof of The Algorithm Consistency

This section provides the proof of the consistency to learn a validation loss estimator ψ_t^* by sampling architectures to be evaluated with a sampler whose distribution centre is updated by:

$$\boldsymbol{A}' \leftarrow \boldsymbol{A} - \eta \cdot \nabla_{\boldsymbol{A}} \psi_t^*(\tilde{\boldsymbol{H}}), \tag{1}$$

where $\boldsymbol{A}$ is the previous sampler centre, $\boldsymbol{A}'$ is the new sampler centre and $\tilde{\boldsymbol{H}}$ is the previous sampled architecture.

We firstly give the definition of the *empirical loss* weighted by the probability that an architecture is sampled and the *expected loss* on the entire data space.

Definition 1 (Empirical Loss)**.** The weighted *empirical loss* of a given estimator ψ at time step T is

$$L_T(\psi) = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{p_t} \ell(\psi(H_t), \mathcal{L}_t) \tag{2}$$

where p_t is the probability that H_t is sampled, $\mathcal{L}_t$ is the label of H_t.

Definition 2 (Expected Loss)**.** Let $\mathbb{H}$ be an architecture search space, $D = \{(H, \mathcal{L}) | H \sim \mathbb{H}\}$ be a data space, and ℓ be a loss function, the *expected loss* of a estimator ψ on the data space D is

$$L(\psi) = \mathbb{E}_{(H,\mathcal{L})\sim D} \left[\ell(\psi(H), \mathcal{L})\right] \tag{3}$$

We demand that the gap, $|L_T(\psi) - L(\psi)|$, between the empirical loss and the expected loss of the validation loss estimator ψ is bounded within an arbitrarily small value ϵ, with probability arbitrarily to 1:

$$\mathbf{P}[|L_T(\psi) - L(\psi)| < \epsilon] > 1 - \delta, \tag{4}$$

where δ is a constant close to 0. To find this bound, a concentration inequality of random process called Azuma's Inequality (Azuma, 1967) can be used.

Theorem 1 (Azuma's Inequality)**.** *Suppose* $\{X_k, k = 0, 1, 2, \dots\}$ *is a martingale and* $|X_k - X_{k-1}| < c_k$ *almost surely. Then for all positive integers* N *and all positive reals* ϵ*,*

$$\mathbf{P}\left[|X_N - X_0| \geq \epsilon\right] \leq 2 \exp\left(\frac{-\epsilon^2}{2\sum_{k=1}^{N} c_k^2}\right) \tag{5}$$

Now, we give our conclusion of the bound in Theorem 2 and demonstrate the proof of it.

Theorem 2. *Let* Ψ *be a hypothesis class containing all the possible hypothesises of estimator* ψ*. For any* $\delta > 0$*, with probability at lest* $1 - \delta$*,* $\forall \psi \in \Psi$*:*

$$|L_T(\psi) - L(\psi)| < \sqrt{\frac{2\left(d + \ln\frac{2}{\delta}\right)}{T}} \tag{6}$$

where d *is the Pollard's pseudo-dimension of* Ψ*.*

Proof. To use Azuma's inequality, we need our target sequence to be a martingale. We consider a sequence of random variables $U_1, \dots, U_T$, with $U_t = \frac{1}{p_t}\ell(\psi(H_t), \mathcal{L}_t) - L(\psi)$. Since probability $p_t \in [0, 1]$ and loss $\ell(\cdot) \in [0, 1]$, we have $|U_t| \leq 1$. Letting $Z_t = \sum_{i=1}^{t} U_i$ and $Z_0 = 0$, Z_t is a martingale. For any $1 \leq t \leq T$:

$$\begin{aligned}
&\mathbf{E}[Z_t | Z_{t-1}, \dots, Z_0] \\
=&\mathbf{E}[U_t + Z_{t-1} | Z_{t-1}, \dots, Z_0] \\
=&\mathbf{E}[l(\psi(H_t), \mathcal{L}_t) - L(H) + Z_{t-1} | Z_{t-1}, \dots, Z_0] \\
=&Z_{t-1}
\end{aligned} \tag{7}$$

We can apply Azuma's inequality on Z_t. Given $|Z_t - Z_{t-1}| = |U_t| \leq 1$ and $|Z_T - Z_0| = |Z_T| = |T(L_T(\psi) - L(\psi))|$, for any real value $\lambda > 0$, we have:

$$\mathbf{P}\left[|L_T(\psi) - L(\psi)| \geq \frac{\lambda}{\sqrt{T}}\right] \leq 2e^{-\lambda^2/2} \tag{8}$$

Setting $\lambda = \sqrt{2(d + \ln\frac{2}{\delta})}$, we can have the desired result in Theorem 2. □

[1] School of Computer Science, University of Sydney [2] Noah's Ark Lab, Huawei Technoligies. Correspondence to: Yanxi Li <yali0722@uni.sydney.edu.au>, Chang Xu <c.xu@sydney.edu.au>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

2. Proof of The Label Complexity

In this section, we discuss the bound of label complexity of our method on an architecture search space $\mathbb{H}$ with size N. According to our sampling strategy in Eq. 1, the next sample to be requested is determined by the current sample and current estimator. Thus after each time step, the architecture to be sampled $\boldsymbol{H}_{t+1}$ at time step $t+1$ satisfies that the estimated validation loss of it is smaller than the estimated validation loss of architecture previously sampled at time step t, i.e. $\psi_t(\boldsymbol{H}_{t+1}) \le \psi_t(\boldsymbol{H}_t)$, as long as the step size η is within a reasonable range (e.g. small enough). According to the standard label complexity, the validation loss estimator ψ_t has an maximum allowed slack $\Delta_t = \sqrt{(8/t)(d + \ln(2/\delta))}$, which means the estimated validation loss at most larger than the ground truth than Δ_t, i.e. $\psi_t(\boldsymbol{H}_t) \le \mathcal{L}_t + \Delta_t$. This is equivalent to shrink the current search space into a smaller subset $\mathbb{H}_{t+1}$:

$$\mathbb{H}_{t+1} = \{\boldsymbol{H} \in \mathbb{H}_t : \\ \psi_t(\boldsymbol{H}) \le \psi_t(\boldsymbol{H}_t) \le \mathcal{L}(\boldsymbol{H}_t) + \Delta_t\}. \quad (9)$$

The new sample $\boldsymbol{H}_{t+1}$ should locate in the subset $\mathbb{H}_{t+1}$. The initial search space is set to be the entire search space: $\mathbb{H}_0 = \mathbb{H}$.

We define a *pseudo-metric* for architectures in Definition 3. With this pseudo-metric, the probability that an architecture is sampled can be defined.

Definition 3 (Pseudo-metric)**.** At time step t, the *pseudo-metric* of architecture $H \in \mathbb{H}_t$ is

$$\rho_t(H) = \max_{H' \in \mathbb{H}_t} \psi_{t-1}(H') - \psi_{t-1}(H) \quad (10)$$

Equation 10 measures the distance from a given architecture to the worst one in the current sub search space. According to our purpose, the further the distance is, the more likely it is to sample the corresponding architecture. With the pseudo-metric, the probability that an architecture $H_n \in \mathbb{H}_t$ is sampled at time step t is $p_n = \rho_t(H_n)$.

Based on the above settings, we give the upper bound of label complexity of our algorithm in Theorem 3.

Theorem 3. *With probability at least* $1 - \delta$*, to learn an estimator* ψ *with error bound* $\epsilon \le \sqrt{(8/N)(d + \ln(2/\delta))}$*, the number of labels requested by the algorithm is at most the order of*

$$\mathcal{O}\left(\sqrt{N(d + \ln(2/\delta))}\right) \quad (11)$$

We need the following lemma for the proof of Theorem 3.

Lemma 4. *For all sub search spaces* $\mathbb{H}_t$*, for all concepts* $\mathcal{L}$ *to be learned, for all* $\delta > 0$*, with probability at least* $1 - \delta$*, for all samples* $H_1, H_2 \in \mathbb{H}_t$ *and all time steps* t

$$|(\psi_t(H_1) - \psi_t(H_2)) - (\mathcal{L}(H_1) - \mathcal{L}(H_2))| \le \Delta_t \quad (12)$$

Proof. Given the definition of allowed slack Δ_t, we can have $|(\psi_t(H) - \mathcal{L}(H)| \le \Delta_t$ at time step t for $H \in \mathbb{H}_t$. Thus $|(\psi_t(H_1) - \psi_t(H_2)) - (\mathcal{L}(H_1) - \mathcal{L}(H_2))| = |\psi_t(H_1) - \psi_t(H_2) - \mathcal{L}(H_1) + \mathcal{L}(H_2)| = |(\psi_t(H_1) - \mathcal{L}(H_1)) - (\psi_t(H_2) - \mathcal{L}(H_2))| \le \Delta_t$. □

Proof of Theorem 3. By applying Lemma 4, for two samples $H_1, H_2 \in \mathbb{H}_t$, we have $\mathcal{L}(H_1) - \mathcal{L}(H_2) \le \psi_{t-1}(H_1) - \psi_{t-1}(H_2) + \Delta_{t-1} \le \psi_{t-1}(H^*) + \Delta_{t-1} - \psi_{t-1}(H^*) + \Delta_{t-1} = 2\Delta_{t-1}$.

Thus the expected value of sample probability $\mathbf{E}[p_N] = \mathbf{E}_{H \sim \mathbb{H}_N}[\max_{H' \sim \mathbb{H}_N} \psi_N(H') - \psi_N(H_N)] \le \mathcal{L}(H_1) - \mathcal{L}(H_2) = 2\Delta_{N-1}$. By $N \cdot \mathbf{E}[p_N]$, the expected number of architectures sampled is at most

$$\mathcal{O}\left(\sqrt{N(d + \ln(2/\delta))}\right) \quad (13)$$

□

References

Azuma, K. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.