Quantization Theory & Multi-task NAS

March 26, 2019

Transfer Learning with Neural AutoML (NIPS2018, Google Brain)

Introduction:

- Transfer Neural AutoML that uses knowledge from prior tasks to speed up network design.
- Use a RNN to control the transfer.

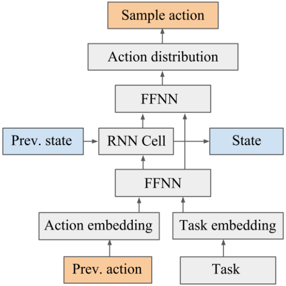


Figure: AutoML Controller

Learned Task Representation

- Tasks are characterized by learning an embedding.
- At each iteration of multitask training, a task is sampled at random. This tasks embedding is fed to the controller, which generates a sequence of actions conditioned on this embedding.
- The child model defined by these actions is trained and evaluated on the task, and the reward is used to update the task-agnostic parameters and the corresponding task embedding.

Action Space

- Deep FFNN: an input embedding module, fully connected layers and a softmax classification layer, regularized with an L2 loss
- Wide-Shallow FFNN: Connects the one-hot token encodings to the softmax classification layer with a linear projection, regularized with a sparse L1 loss.

Parameter	**Search Space**
1) Input embedding modules	Text input: refer to Table 2. Image input: refer to Table 3.
2) Fine-tune input embedding module	{True, False}
3) Number of hidden layers	{1, 2, 3, 5, 7}
4) Hidden layers size	{8, 16, 32, 64, 128, 256}
5) Hidden layers activation	{relu, swish}
6) Hidden layers normalization	{none, batch norm, layer norm}
7) Hidden layers dropout rate	{0.0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6}
8) Deep tower learning rate	{0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0}
9) Deep tower regularization weight	{0.0, 0.00001, 0.0001, 0.001, 0.01, 0.1, disable deep tower}
10) Wide tower learning rate	{0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0}
11) Wide tower regularization weight	{0.0, 0.00001, 0.0001, 0.001, 0.01, 0.1, disable wide tower}
12) Number of training samples	{1000, 3000, 10000, 30000, 100000, 300000, 1000000}

Experiment

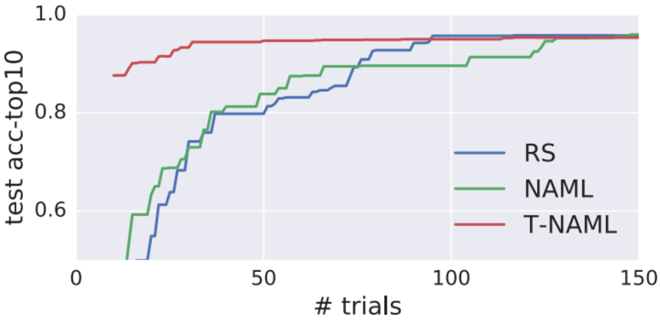


Figure: Comparison on an image classification task, Cifar-10. Mean test accuracy of the top 10 models chosen on the validation set.

Training Quantized Nets: A Deeper Understanding

Given an empirical risk minimization problems of the form:

$$\min_{w \in \mathcal{W}} F(w) := \frac{1}{m} \sum_{i=1}^{m} f_i(w), \tag{1}$$

A Basic Stochastic Quantization (SR):

$$w_b^{t+1} = Q_s(w_b^t - \alpha_t \nabla f(w_b^t)) \tag{2}$$

with quantization as:

$$Q_s(w) = \Delta \cdot \begin{cases} \lfloor \frac{w}{\Delta} \rfloor + 1 & \text{for} \quad p \leq \frac{w}{\Delta} - \lfloor \frac{w}{\Delta} \rfloor \\ \lfloor \frac{w}{\Delta} \rfloor & \text{otherwise} \end{cases} \tag{3}$$

Convergence Analysis

Assumption:

- Loss function F is μ-strongly convex: $f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}||y - x||_2^2$
- Gradient is bounded: $\mathbb{E}||\nabla f(w^t)|| \leq G^2$

Convergence Analysis:

Theorem

Assume that F is μ-strongly convex and the learning rates are given by $\alpha_t = \frac{1}{\mu(t+1)}$. Consider the SR algorithm with updates of the form (2). Then, we have:

$$\mathbb{E}[F(\bar{w}^T) - F(w^*)] \leq \frac{(1 + log(T + 1))G}{2\mu T} + \frac{\sqrt{d}\Delta G}{2} \tag{4}$$

where $\bar{w}^T = \frac{1}{T}\sum_{i=1}^{t} w^t$

Proof of Theorem.1

General Idea:

- Start from quantization error, which is related to quantization resolution, gradient etc.
- Then introduce F by μ-strongly convex.
- Finally telescope sum to reduce intermediate term.

Quantization Error:

$$r^t = Q_s(w_b^t - \alpha_t \nabla f(w_b^t)) - (w_b^t - \alpha_t \nabla f(w_b^t))$$

is bounded by:

$$\mathbb{E}||r^t||^2 \leq \sqrt{d}\Delta\alpha_t G \tag{5}$$

Weights update:

$$w^{t+1} = w^t - \alpha_t \nabla f(w^t) + r^t \rightarrow w^{t+1} - w^* = (w^t - w^*) - (\alpha_t \nabla f(w^t) - r^t)$$

Proof of Theorem.1 (Cont.)

$\tilde{f}$: arbitrary f from F.

$$\begin{aligned}
&\mathbb{E}\|w^{t+1} - w^*\|^2 \\
&= \|w^t - w^*\|^2 - 2\underbrace{\mathbb{E}\langle w^t - w^*, \alpha_t \nabla \tilde{f}(w^t) - r^t\rangle}_{\mathbb{E}[r^t]=0} + \underbrace{\mathbb{E}\|\alpha_t \nabla \tilde{f}(w^t) - r^t\|^2}_{\mathbb{E}[r^t]=0} \\
&= \|w^t - w^*\|^2 - 2\alpha_t\langle w^t - w^*, \nabla F(w^t)\rangle + \alpha_t^2 \mathbb{E}\|\nabla \tilde{f}(w^t)\|^2 + \mathbb{E}\|r^t\|^2 \\
&\leq \|w^t - w^*\|^2 - 2\alpha_t\langle w^t - w^*, \nabla F(w^t)\rangle + \alpha_t^2 G^2 + \underbrace{\sqrt{d}\Delta\alpha_t G}_{\mathbb{E}\left\|r^t\right\|_2^2 \leq \sqrt{d}\Delta\alpha_t G.},
\end{aligned}$$

By μ-strongly convex:

$F(w^*) - F(w^t) \geq \left\langle w^* - w^t, \nabla F(w^t)\right\rangle + \frac{\mu}{2}||w^* - w^t||^2 \rightarrow$

$$\begin{aligned}
\mathbb{E}\|w^{t+1} - w^*\|^2 \leq (1 - \alpha_t\mu)\|w^t - w^*\|^2 - 2\alpha_t(F(w^t) - F(w^*)) \\
+\alpha_t^2 G^2 + \sqrt{d}\Delta\alpha_t G.
\end{aligned}$$

Proof of Theorem.1 (Cont.)

Re-arranging the terms, taking expectation, and asssume that the stepsize decreases with the rate $\alpha_t = 1/\mu(t+1)$. Then we have:

$$\mathbb{E}(F(w^t) - F(w^*)) \leq \frac{\mu t}{2}\mathbb{E}\|w^t - w^*\|^2 - \frac{\mu(t+1)}{2}\mathbb{E}\|w^{t+1} - w^*\|^2 + \frac{1}{2\mu(t+1)}G^2 + \frac{\sqrt{d}\Delta G}{2}.$$

Averaging over $t = 0$ to T, we get a telescoping sum on the right hand side:

$$\{t = t\} : \underbrace{\frac{\mu t}{2}\mathbb{E}\|w^t - w^*\|^2}_{\text{eliminate}} - \frac{\mu(t+1)}{2}\mathbb{E}\|w^{t+1} - w^*\|^2$$

$$\{t = t-1\} : \underbrace{\frac{\mu(t-1)}{2}\mathbb{E}\|w^{t-1} - w^*\|^2}_{\text{eliminate when } t=1} - \underbrace{\frac{\mu t}{2}\mathbb{E}\|w^t - w^*\|^2}_{\text{eliminate}}$$

Proof of Theorem.1 (Cont.)

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}(F(w^t)-F(w^*)) &\leq \frac{G^2}{2\mu T}\sum_{t=0}^{T}\frac{1}{t+1}+\frac{\sqrt{d}\Delta G}{2} \\ &- \frac{\mu(T+1)}{2}\mathbb{E}\|w^{T+1}-w^*\|^2 (\text{Get rid of}) \\ &\leq \frac{(1+\log(T+1))G^2}{2\mu T}+\frac{\sqrt{d}\Delta G}{2}.\end{aligned}$$

Using Jensen's inequality, we have:

$$\begin{aligned}\mathbb{E}(F(\bar{w}^T)-F(w^*)) &\leq \frac{1}{T}\sum_{t=0}^{T}\mathbb{E}(F(w^t)-F(w^*)) \\ &\leq \frac{(1+\log(T+1))G^2}{2\mu T}+\frac{\sqrt{d}\Delta G}{2}\end{aligned}$$

ProxQuant: Quantized Neural Networks via Proximal Operators

Weights update:

$$\theta_{t+1} = \mathsf{prox}_{\eta_t \lambda_t R}\left(\theta_t - \eta_t \tilde{\nabla} L(\theta_t)\right). \tag{6}$$

Compare with Stochastic Quantization:

$$w_b^{t+1} = Q_s(w_b^t - \alpha_t \nabla f(w_b^t))$$

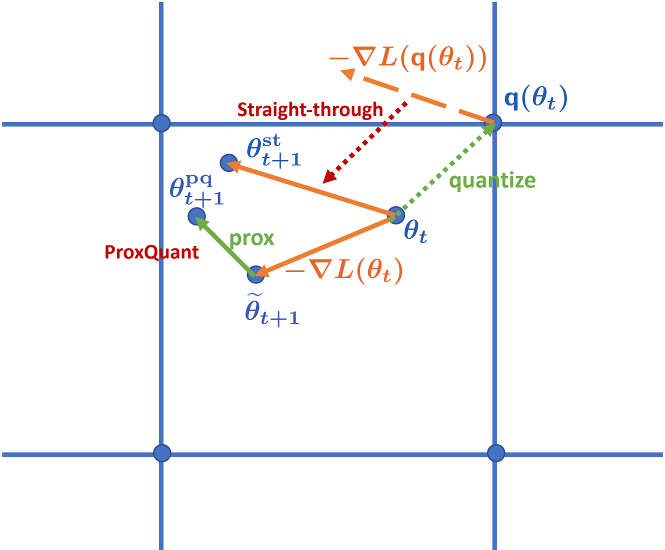
Proximal Operation:

$$\mathsf{prox}_{\lambda R}(\theta) := \arg\min_{\tilde{\theta} \in \mathbb{R}^d} \left\{ \frac{1}{2} ||\tilde{\theta} - \theta||_2^2 + \lambda R(\tilde{\theta}) \right\}.$$

In the case $\mathcal{Q} = \{\pm 1\}^d$ for example, one could take

$$R(\theta) = R_{\mathrm{bin}}(\theta) = \sum_{j=1}^{d} \min\{|\theta_j - 1|, |\theta_j + 1|\}. \tag{7}$$

ProxQuant



$-\nabla L(\mathbf{q}(\theta_t))$
$\mathbf{q}(\theta_t)$
Straight-through
$\theta_{t+1}^{\mathrm{st}}$
quantize
$\theta_{t+1}^{\mathrm{pq}}$
θ_t
prox
ProxQuant
$-\nabla L(\theta_t)$
$\widetilde{\theta}_{t+1}$

Convergence Analysis of ProxQuant

Theorem (Convergence of ProxQuant)

Assume that the loss L is β-smooth (i.e. has β-Lipschitz gradients) and the regularizer R is differentiable. Let $F_\lambda(\theta) = L(\theta) + \lambda R(\theta)$ be the composite objective and assume that it is bounded below by $F_\star$. Running ProxQuant with batch gradient ∇L, constant stepsize $\eta_t \equiv \eta = \frac{1}{2\beta}$ and $\lambda_t \equiv \lambda$ for T steps, we have the convergence guarantee

$$||\nabla F_\lambda(\theta_{T_{\text{best}}})||_2^2 \leq \frac{C\beta(F_\lambda(\theta_0) - F_\star)}{T} \tag{8}$$

where

$$T_{\text{best}} = \arg\min_{1 \leq t \leq T} ||\theta_t - \theta_{t-1}||_2, \tag{9}$$

where $C > 0$ is a universal constant.

General Idea

f is β-smooth if:

$$||\nabla f(y) - \nabla f(x)|| \leq \beta||x - y|| \tag{10}$$

this can implies:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2}||y - x||^2 \tag{11}$$

- Construction comparison between weight change $(\theta_{t+1} - \theta_t)$ and regularized loss change $(F_\lambda(\theta_{t+1}) - F_\lambda(\theta_t))$.
- Telescoping to attain minimal convergence step $(\min_{0 \leq t \leq T-1} ||\theta_{t+1} - \theta_t||_2^2)$ is bounded by discrepancy between regularized loss and optimal loss w.r.t iteration T.
- Use smoothness to convert $\min_{0 \leq t \leq T-1} ||\theta_{t+1} - \theta_t||_2^2$ to $||\nabla F_\lambda(\theta_{T_{\text{best}}})||_2^2$.

Proof of Convergence Analysis of ProxQuant

Combine:

$$\theta = \mathsf{prox}_{\eta_t \lambda_t R}\left(\theta_t - \eta_t \nabla L(\theta_t)\right),$$

and

$$\mathsf{prox}_{\lambda R}(\theta) := \arg\min_{\tilde{\theta} \in \mathbb{R}^d} \left\{ \frac{1}{2} ||\tilde{\theta} - \theta||_2^2 + \lambda R(\tilde{\theta}) \right\}.$$

we have:

$$\begin{aligned} ||\tilde{\theta} - \theta||_2^2 &= ||\theta - \theta_t + \eta_t \nabla L(\theta_t)||_2^2 \\ = \quad & ||\theta - \theta_t||_2^2 + 2\left\langle \theta - \theta_t, \nabla L(\theta_t) \right\rangle + \underbrace{\eta^2 (\nabla L(\theta_t))^2}_{fixed} \end{aligned}$$

Incorporate $L(\theta_t)$ and $\lambda R(\tilde{\theta})$:

$$\theta_{t+1} = \arg\min_{\theta \in \mathbb{R}^d} \left\{ L(\theta_t) + \langle \theta - \theta_t, \nabla L(\theta_t) \rangle + \frac{1}{2\eta} ||\theta - \theta_t||_2^2 + \lambda R(\theta) \right\}.$$

Proof (Cont.)

θ_{t+1} minimizes the above objective:

$$
\begin{aligned}
&F_\lambda(\theta_t) = L(\theta_t) + \lambda R(\theta_t) \\
&\geq \underbrace{L(\theta_t) + \langle \theta_{t+1} - \theta_t, \nabla L(\theta_t) \rangle}_{\beta-\text{smooth}} + \frac{1}{2\eta} ||\theta_{t+1} - \theta_t||_2^2 + \lambda R(\theta_{t+1}) \\
&\geq L(\theta_{t+1}) + \underbrace{\left(\frac{1}{2\eta} - \frac{\beta}{2} \right)}_{\eta = \frac{1}{2\beta}} ||\theta_{t+1} - \theta_t||_2^2 + \lambda R(\theta_{t+1}) \\
&= F_\lambda(\theta_{t+1}) + \frac{\beta}{2} ||\theta_{t+1} - \theta_t||_2^2.
\end{aligned}
$$

Proof (Cont.)

Telescoping the above bound for $t = 0, \ldots, T-1$, we get that

$$\sum_{t=0}^{T-1} ||\theta_{t+1} - \theta_t||_2^2 \leq \frac{2(F_\lambda(\theta_0) - F_\lambda(\theta_T))}{\beta} \leq \frac{2(F_\lambda(\theta_0) - F_\star)}{\beta}.$$

Proximity guarantee

$$\min_{0 \leq t \leq T-1} ||\theta_{t+1} - \theta_t||_2^2 \leq \frac{2(F_\lambda(\theta_0) - F_\star)}{\beta T}. \tag{12}$$

Using $\frac{1}{2\eta}$-smooth and first-order optimality condition for θ_{t+1} gives

$$\nabla L(\theta_t) + \frac{1}{\eta}(\theta_{t+1} - \theta_t) + \lambda \nabla R(\theta_{t+1}) = 0.$$

Proof (Cont.)

Combining the above equality and the smoothness of L, we get

$$||\nabla F_\lambda(\theta_{t+1})||_2 = ||\nabla L(\theta_{t+1}) + \lambda \nabla R(\theta_{t+1}) + \nabla L(\theta_t) - \nabla L(\theta_t)||_2$$
$$= ||\frac{1}{\eta}(\theta_t - \theta_{t+1}) + \underbrace{\nabla L(\theta_{t+1}) - \nabla L(\theta_t)}_{\text{smoothness}}||_2$$
$$\leq \left(\frac{1}{\eta} + \beta\right)||\theta_{t+1} - \theta_t||_2 = 3\beta||\theta_{t+1} - \theta_t||_2.$$

Choosing $t = T_{\text{best}} - 1$ and applying the proximity guarantee (12), we get

$$||\nabla F_\lambda(\theta_{T_{\text{best}}})||_2^2 \leq 9\beta^2||\theta_{T_{\text{best}}} - \theta_{T_{\text{best}}-1}||_2^2$$
$$= 9\beta^2 \min_{0\leq t\leq T-1} ||\theta_{t+1} - \theta_t||_2^2 \leq \frac{18\beta(F_\lambda(\theta_0) - F_\star)}{T}.$$

This is the desired bound.