ニューラルアーキテクチャサーチの最適停止手法とその評価

坂本　航太郎　　統計的機械学習研究センター　特任助教

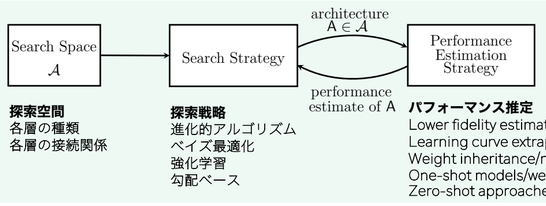
概要

ニューラルアーキテクチャサーチ (Neural architecture search: NAS) はニューラルネットの構造設計を自動化するAutoMLの枠組みである．近年のNASは低コストでの構造探索に成功している一方で，精度と探索コストには依然としてトレードオフがあり，探索の停止のタイミングを決定することは重要な課題である．また， NASの代表の一つである Differentiable Architecture Search (DARTS) においては性能劣化が報告されており， 早期停止による性能劣化を防ぐ手法が提案されている．本研究では構造パラメータに対する汎化誤差の期待値の差分を用いたNASの最適停止手法を提案する．複数のベンチマークデータセット，NAS-Bench-201，NATS-Bench，NAS-Bench-301，NAS-Bench-ASR，を用いて，本提案手法がNASを高精度に保ちつつ，探索コストを削減できることを評価する．

背景

Neural Architecture Search (NAS)

所定のデータセット・タスクに対して優れたパフォーマンスの得られるニューラルネットの構造を比較的短時間で導くことを目指す



NASの概観図
探索空間(構造の集合)$\mathcal{A}$の中から探索戦略が一つの構造 $A \in \mathcal{A}$を選び，そのパフォーマンスを推定する
その推定値に依って，別の構造をAの中から選ぶ…ということを繰り返し，NNの構造を最適化する

探索空間
各層の種類
各層の接続関係

探索戦略
進化的アルゴリズム
ベイズ最適化
強化学習
勾配ベース

パフォーマンス推定
Lower fidelity estimates
Learning curve extrapolation
Weight inheritance/network morphisms
One-shot models/weight sharing
Zero-shot approaches

NASの課題

- 膨大な探索コスト→エコなNAS
- 早期停止：コスト削減と性能改善

Xu, J., et al. KNAS: Green Neural Architecture Search. (2021) https://proceedings.mlr.press/v139/xu21m.html
Baker, B., et al. (2018) https://openreview.net/forum?id=HJqk3N1vG
Li & Talwalkar, (2019) http://proceedings.mlr.press/v115/li20c.html

NASの探索をいつどのように停止するか？

One-shot Neural Architecture Search (One-shot NAS)

ニューラルネットワークの構造とパラメータを一度の学習サイクルにおいて同時に最適化する枠組み

Differentiable Architecture Search (DARTS)

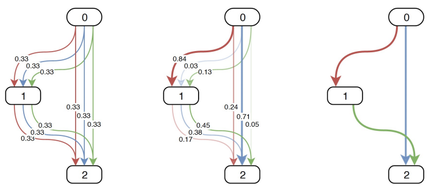
ニューラルネットワーク→有向非巡回グラフ（Directed Acyclic Graph：DAG）

Bi-level最適化問題　$\min_{\alpha}\{f(\alpha) \triangleq \mathcal{L}_{valid}(w^*(\alpha),\alpha)\}$
$s.t.\ w^*(\alpha) = \arg\min_{w} \mathcal{L}_{train}(w,\alpha)$

$w^*(\alpha) \approx w - \xi\nabla_w \mathcal{L}_{train}(w,\alpha)$ と近似し，以下の最適化を交互に行う

$\nabla_w \mathcal{L}_{train}(w,\alpha)$

$\nabla_\alpha \mathcal{L}_{valid}(w - \xi\nabla_w \mathcal{L}_{train}(w,\alpha),\alpha)$



(Taken from ICLR-20-RobustDARTS-slides)

$$x_j = \sum_{\nu=0}^{K-1} \frac{\exp\left(\alpha_\nu^{(i,j)}\right)}{\sum_{\nu'=0}^{K-1}\exp\left(\alpha_{\nu'}^{(i,j)}\right)} \cdot o_\nu(x_i) \qquad o^{(i,j)} \in \arg\max_{\nu\in\mathcal{O}} \alpha_\nu^{(i,j)}$$

DARTSの課題

skip-connection (恒等写像)が多く含まれる構造が獲得されやすいなどの性能劣化（ランダムサーチより悪化）

Yang, A., et al., Nas evaluation is frustratingly hard. (2020) https://openreview.net/forum?id=HygrdpVKvr

既存手法：早期停止による解決法

- DARTS+
  単一のセル内でskip-connectionが2つ以上になったら探索を停める/構造パラメータが安定したら停める
- RobustDARTS
  損失地形のシャープさを表しているとされる検証損失のヘシアンの支配的な固有値をモニターし，過去5エポック分に対する最大固有値の平均値の比が0.75以下になった（固有値が大きく増加した）時に停止する

導入

確率緩和に基づくOne-shot NAS：ASNG-NAS

- 構造をスーパーネットからサンプリングし，重み　　は構造間で共有
- カテゴリ変数空間で定義される確率分布族
- 確率分布のもとでの目的関数の期待値の最

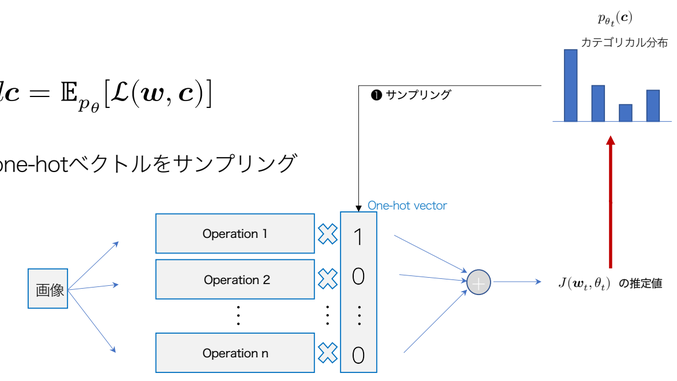
$$J(w,\theta) = \int_{\boldsymbol{c}\in\mathcal{C}} \mathcal{L}(\boldsymbol{w},\boldsymbol{c})p_\theta(\boldsymbol{c})d\boldsymbol{c} = \mathbb{E}_{p_\theta}[\mathcal{L}(\boldsymbol{w},\boldsymbol{c})]$$

❶ カテゴリカル分布から構造を選択するone-hotベクトルをサンプリング

❷ サンプルした構造を順伝播＆重み更新
$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \eta_{\boldsymbol{w}}\nabla_{\boldsymbol{w}} J(\boldsymbol{w}_t,\theta_t)$

❸ 検証データを使って順伝播＆分布更新
$\theta_{t+1} = \theta_t + \eta_\theta^t \widetilde{\nabla}_\theta J(\boldsymbol{w}_{t+1},\theta_t)$

❹ 学習率を更新



[1] Y. Akimoto, S. Shirakawa, N. Yoshinari, K. Uchida, S. Saito, and K. Nishida, "Adaptive stochastic natural gradient method for one-shot neural architecture search," ICML 2019

提案法

停止基準の構成

- NASの目的は汎化誤差を最小化するニューラルネットの構造を獲得することであるので汎化誤差の差分に注目して停止基準を構成することは自然である
- 探索ステップ前後での汎化誤差のハイパーパラメータの分布に対する期待値の差分は汎化誤差が探索時には得られないので直接計算できない
- そこでその上界を評価して閾値と比較することで停止タイミングを決定する
- 差分の上界を評価するにあたり，以下のPinskerの不等式を考える

定理 ([2], Fact 9). $p_{\theta_t}(\boldsymbol{c})$と$p_{\theta_{t-1}}(\boldsymbol{c})$ を任意の確率分布とする. このとき任意の可測関数 $\mathcal{L}(\boldsymbol{c}) \in [a,b]$ について，以下の不等式が成り立つ:

$$\mathbb{E}_{p_{\theta_t}(\boldsymbol{c})}[\mathcal{L}(\boldsymbol{w},\boldsymbol{c})] - \mathbb{E}_{p_{\theta_{t-1}}(\boldsymbol{c})}[\mathcal{L}(\boldsymbol{w},\boldsymbol{c})] \le (b-a)\sqrt{\frac{1}{2}\mathrm{D}_{KL}\left[p_{\theta_t}(\boldsymbol{c}) \parallel p_{\theta_{t-1}}(\boldsymbol{c})\right]}$$

[2] D. Russo and B. V. Roy, "An information-theoretic analysis of Thompson sampling," Journal of Machine Learning Research, vol. 17, no. 68, pp. 1–30, 2016.

- NASにおいて重みパラメータは更新されるので定理は単純適用ができない
- そこでテイラー展開によりその影響を近似する方針をとる

$J(\boldsymbol{w},\theta)$ を$\boldsymbol{w}$ について $\boldsymbol{w}_t$まわりで n 次のテイラー展開したときの値の値を$T_n(\boldsymbol{w},\theta)$,
剰余項$R_n(\boldsymbol{w},\theta) := J(\boldsymbol{w},\theta) - T_n(\boldsymbol{w},\theta)$を$T_n(\boldsymbol{w},\theta)$から０次の項を取り除いたものを$T_n'(\boldsymbol{w},\theta)$とする.

$$J(\boldsymbol{w}_t,\theta_t) - J(\boldsymbol{w}_{t-1},\theta_{t-1}) = T_n(\boldsymbol{w}_t,\theta_t) + R_n(\boldsymbol{w}_t,\theta_t) - J(\boldsymbol{w}_{t-1},\theta_{t-1})$$
$$= J(\boldsymbol{w}_{t-1},\theta_t) - J(\boldsymbol{w}_{t-1},\theta_{t-1}) + T_n'(\boldsymbol{w}_t,\theta_t) + R_n(\boldsymbol{w}_t,\theta_t)$$

上式の第１項，第２項に定理を適用し以下の不等式を得る：

$$J(\boldsymbol{w}_t,\theta_t) - J(\boldsymbol{w}_{t-1},\theta_{t-1}) \le (b-a)\sqrt{\frac{1}{2}D_{KL}[p_{\theta_t}(\boldsymbol{c}) \parallel p_{\theta_{t-1}}(\boldsymbol{c})]} + T_n'(\boldsymbol{w}_t,\theta_t) + R_n(\boldsymbol{w}_t,\theta_t)$$

Jを$\boldsymbol{w}$について C^n級の関数とし,$\boldsymbol{w} \in [\boldsymbol{w}_{t-1}, \boldsymbol{w}_{t-1}+d]$ かつ$\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\| < d$ の全てにおいて$J(\boldsymbol{w},\theta) \le M$と仮定したとき,
剰余項 $R_n(\boldsymbol{w},\theta)$に対してテイラーの不等式を適用することで以下の不等式が成り立つ：

$$\mathbb{E}_{p_{\theta_t}(\boldsymbol{c})}[\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{c})] - \mathbb{E}_{p_{\theta_{t-1}}(\boldsymbol{c})}[\mathcal{L}(\boldsymbol{w}_{t-1},\boldsymbol{c})]$$
$$\le (b-a)\sqrt{\frac{1}{2}\mathrm{D}_{KL}\left[p_{\theta_t}(\boldsymbol{c}) \parallel p_{\theta_{t-1}}(\boldsymbol{c})\right]} + \underbrace{T_n'(\boldsymbol{w}_t,\theta_t)}_{} + \frac{M}{n!}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^n$$

$J(w,\theta)$をw について w_tまわりでn 次のテイラー展開したときの値から０次の項を取り除いたもの

$M \ge J(w,\theta)$

上記の不等式を停止基準として採用するためには$\mathcal{L}$の範囲 [a, b] とMを予め決めないといけないが$M = (b-a)$と仮定でき, 0次の項でテイラー展開したときは$(\mathrm{b}-\mathrm{a})$を消去できる. ただし, 高次展開では, $(\mathrm{b}-\mathrm{a})$は残るので, 目的関数の上界を決めておく必要がある.

$$r_t^1 = (b-a)\sqrt{\frac{1}{2}\mathrm{D}_{KL}\left[p_{\theta_t}(\boldsymbol{c}) \parallel p_{\theta_{t-1}}(\boldsymbol{c})\right]} + \nabla_{\boldsymbol{w}} J(\boldsymbol{w}_{t-1},\theta_t)\cdot(\boldsymbol{w}_t - \boldsymbol{w}_{t-1}) + (b-a)\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|$$

$$r_t^2 = (b-a)\sqrt{\frac{1}{2}\mathrm{D}_{KL}\left[p_{\theta_t}(\boldsymbol{c}) \parallel p_{\theta_{t-1}}(\boldsymbol{c})\right]} + \nabla_{\boldsymbol{w}} J(\boldsymbol{w}_{t-1},\theta_t)\cdot(\boldsymbol{w}_t - \boldsymbol{w}_{t-1}) + \frac{1}{2}(\boldsymbol{w}_t - \boldsymbol{w}_{t-1})^{\mathrm{T}}\nabla_{\boldsymbol{w}}^2 J(\boldsymbol{w}_{t-1},\theta_t)(\boldsymbol{w}_t - \boldsymbol{w}_{t-1}) + \frac{(b-a)}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2$$

- 誤差関数の有界性が求められる

→bi-tempered logistic lossを採用する
→より一般的な交差エントロピー等も仮定することで使用可能

停止基準の自動決定

- 上界が閾値を下回ったら構造探索を停止する
- 検証データが利用可能であるので検証データを用いて停止基準を自動決定することを考えたい
- Makarova, et al. (2021)は予測モデルのハイパーパラメータのベイズ最適化における過剰適合の解決策として，simple regretの上界が汎化誤差の標準偏差を下回ったときに訓練を早期停止することを提案している
- 汎化誤差の標準偏差はk-fold交差検証, 訓練及び検証データから以下の式により推定可能である

$$\lambda_t = \sqrt{\frac{1}{k} + \frac{|S_{\mathrm{val}}|}{|S_{\mathrm{train}}|}}\,\hat{s}_t$$

- 提案法の停止基準は汎化誤差の期待値の差分をもとにしており，汎化誤差の標準偏差の推定値と比較することは自然である
- そこで汎化誤差の標準偏差の推定値を閾値として採用する

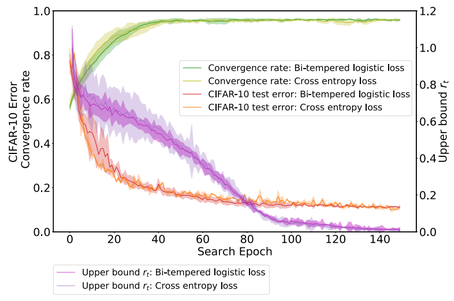
DARTSの停止基準への応用

- 提案法はASNG-NASの設定を停止基準の定式化に用いており，よりポピュラーなone-shot NASの手法であるDARTSへ直接適用はできない
- DARTSでは(1)最適化対象が汎化誤差の期待値ではない (2)構造パラメータは確率緩和をベースにしたものではない (連続緩和)
- そこでオペレーション候補の選択を担うsoftmaxを確率分布と見做して探索エポック前後の変化を評価することで停止基準として用いる方針を取る
- 停止基準に用いるはsoftmaxにより計算される事後分布である
- 閾値の自動決定についてはone-hot化の離散化を行った後に5-fold交差検証によって計算する

評価実験

ASNG-NASの最適停止

パラメータθの収束率との比較

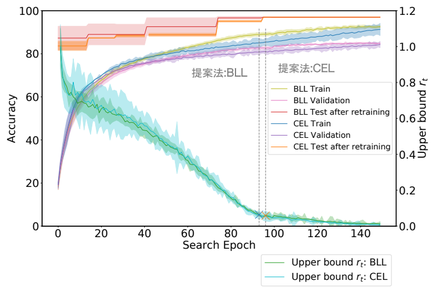


パラメータθの収束率

$$\frac{1}{d}\sum_{i=1}^{d}\max_{\mathrm{j}}[\theta]_{i,j}$$

i: cell index
j: operation index

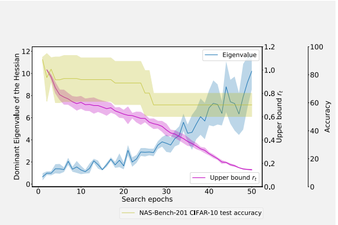
上界の評価値と停止タイミング



停止 epoch	Test Accuracy	Test Accuracy (Re-trained)
0	19.2%	96.91%
14	69.09%	96.89%
27	79.4%	97.02%
41	83.4%	97.04
74	87.19%	97.12%
92	87.6%	97.14%
95	88.02%	96.86%
100	90.34%	95.57%

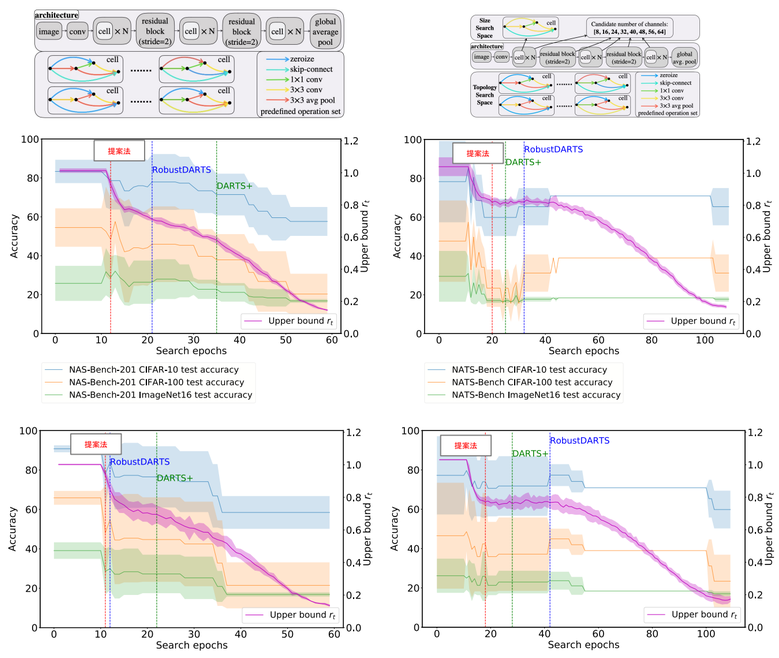
- CIFAR-10 のtrain/validation/test 精度と提案手法の停止基準による停止タイミングの一つを示している
- テスト精度は停止タイミングの後も増加しているが構造探索による性能向上はこれ以上望めないと考えられる

DARTSの最適停止

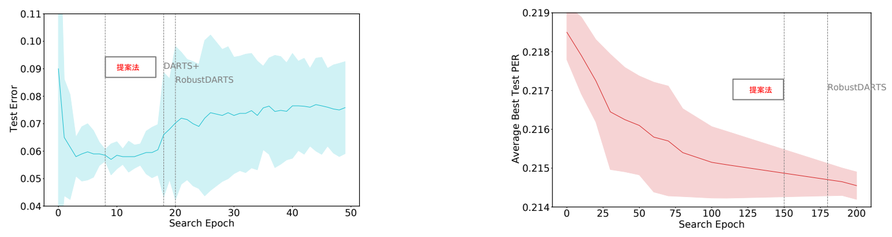


- RobustDARTS で注目するヘシアンの最大固有値は性能劣化が起きてから大きく増加
- 一方で我々の手法で注目する上界は減少を続けており適切な値を与えれば早期に停止することが可能

NAS Benchmark-201　　NATS Benchmark



NAS Benchmark-301　　NAS-Bench-ASR



*Bi-tempered logistic loss/CIFAR-10で訓練　　*Bi-tempered logistic lossで訓練/Phoneme error rate (PER)

- 既存の早期停止手法であるRobustDARTS及びDARTS+に比べ提案法は非常に早期に停止できている
- DARTSにおいては冗長な探索が行われていたことを示唆している

謝辞

九州工業大学石橋英朗氏，横浜国立大学白川真一氏，筑波大学秋本洋平氏，佐藤怜氏，統計数理研究所日野英逸氏との共同研究です．
本研究は国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託業務（JPNP18002）の支援を受けました．