

論文 / 著書情報  
Article / Book Information

論題(和文)	畳み込み行列の安定ランクを用いた深層学習モデル構造探索
Title(English)	Differentiable Architecture Search Using Stable Rank of Convolutional Layers
著者(和文)	町田 兼悟, 宇都 有昭, 篠田 浩一, 鈴木 大慈
Authors(English)	Kengo Machida, Kuniaki Uto, Koichi Shinoda, Taiji Suzuki
出典(和文)	人工知能学会全国大会 (第35回)論文集, ,
Citation(English)	Proceedings of the Annual Conference of JSAI, , ,
発行日 / Pub. date	2021, 6
Note	<p>ここに掲載した著作物の利用に関する注意 本著作物の著作権は人工知能学会に帰属します。本著作物は著作権者である人工知能学会の許可のもとに掲載するものです。ご利用に当たっては「著作権法」に従うことをお願いいたします。</p> <p>Notice for the use of this material. The copyright of this material is retained by the Japanese Society for Artificial Intelligence (JSAI). This material is published on this web site with the agreement of the author(s) and the JSAI. Please be complied with Copyright Law of Japan if any users wish to reproduce, make derivative work, distribute or make available to the public any part or whole thereof. All Rights Reserved, Copyright (C) The Japanese Society for Artificial Intelligence.</p>

# 畳み込み行列の安定ランクを用いた深層学習モデル構造探索

Differentiable Architecture Search Using Stable Rank of Convolutional Layers

町田 兼悟 \*1

Kengo Machida

宇都 有昭 \*1

Kuniaki Uto

篠田 浩一 \*1

Koichi Shinoda

鈴木 大慈 \*2\*3

Taiji Suzuki

東京工業大学 情報理工学院

School of Computing, Tokyo Institute of Technology

\*2東京大学 大学院情報理工学系研究科

Graduate School of Information Science and Technology, The University of Tokyo, Japan

\*3理研 革新知能統合研究センター

Center for Advanced Intelligence Project, RIKEN, Japan

We propose Neural Architecture Search method MSR-DARTS, which uses the stable rank of a convolutional layer for selecting the optimal network structure. We evaluate MSR-DARTS on CIFAR10 and ImageNet dataset. It achieves an error rate of 2.80% on CIFAR10 and a top-1 error rate of 24.0% of ImageNet.

## 概要

深層学習は、画像処理、音声認識、自然言語処理などの分野で広く用いられており、高い応用可能性から研究が盛んに行われている。深層学習にはモデルの定義が必要であり、モデル構造の最適化を自動で行うタスクである深層学習のモデル構造探索 (Neural Architecture Search; NAS) は、現在注目を集めている研究分野である。本論文では、安定ランクを用いる NAS 手法である MSR-DARTS (Minimum Stable Rank Differentiable Architecture Search) を提案する。安定ランクは行列のランクの代理指標として用いられ、安定ランクが低い層により構成されるモデルは入力にノイズが含まれる場合にノイズを減衰させるため汎化性能が高くなる。提案法は、モデルの畳み込み層を行列とみなした時畳み込み行列の安定ランクにより最適なモデル構造を選択する。CIFAR、ImageNet データセットを用いて評価実験を行い、それぞれ 2.80%、24.0% のテストエラー精度を達成した。

## 1. 序論

モデル構造探索 (Neural Architecture Search; NAS) は深層学習モデルの構造を最適化するタスクであり、既に多くのタスクで精度向上をもたらしている。NAS の登場で、専門知識を有する技術者が構築した従来のモデル構造よりも、より良いモデルが構築可能になった。また、深層学習技術の速い普及も期待される。

NAS 手法の一つである Differentiable Architecture Search (DARTS) [Liu 19] はモデル探索の計算コストが小さく、現在注目されている。DARTS では、探索空間内のモデルをすべて内包する大規模なモデルを探索モデルとして定義し、すべてのモデルを同時に学習していく。探索モデルは、各モデルに含まれる畳み込みなどの重みの他に、構造パラメータと呼ばれる、探索空間のそれぞれのモデルを重み付けするパラメータを含む。探索モデルの重みと構造パラメータを、一方を固定して交互更新を行うことで、両パラメータの同時最

適化を行う。

本論文では、構造パラメータを学習する DARTS の手法に代わり、構造パラメータを学習せずに固定して重みだけを最適化する。モデルの構造の決定には、探索モデル学習後の畳み込み層のパラメータ行列の安定ランクを用いる。安定ランクは深層学習の汎化性能と関係しており [Arora 18]、モデルを構成する畳み込み層において、安定ランクが低いほど汎化性能が高いと考えられる。精度の高いモデルの探索を目的として、構造を決定する際に安定ランクを用いる Minimum Stable Rank Differentiable ARchiTecture Search (MSR-DARTS) を提案する。

## 2. 従来手法

### 2.1 Neural Architecture Search; NAS

NAS は探索空間と呼ばれる複数のモデルの集合から、最適なモデルを効率よく探索することで実現される。従来、NAS は強化学習や進化的計算を用いて探索を行ってきた [Zoph 17, Zoph 18, Pham, Liu 18, Tang 17, Real 19]。強化学習や進化的計算と比較して高速なモデル探索を行う DARTS は、離散的であるモデルの集合を連続表現にすることで微分可能にして、直接勾配降下法を用いてモデル探索を行う。各モデルの出力に構造パラメータと呼ばれる学習可能パラメータを導入し、各モデルの出力を構造パラメータより算出される重みに基づく重み付き和で表現することにより、全モデルの出力をまとめることで構造パラメータの最適化を行う。その他多くの手法 [Xie 19, Chen 19, Xu 19, Liang 19] が DARTS のアプローチを用いており、効率的な NAS 手法が数多く提案されている。

### 2.2 安定ランク

深層学習モデルは、学習サンプル数よりもはるかに多いパラメータを含む。それにもかかわらず、モデルは過学習せず高い汎化性能を持つ。汎化性能誤差バウンドは深層学習モデルの期待誤差と訓練誤差の差であり、訓練誤差の小さいモデルが小さい汎化性能誤差バウンドを持つ場合に、そのモデルは汎化性能が高いことが保証される。VC 次元や安定ランクなど、汎化性能誤差バウンドを特徴付ける多くの指標が数多く提案されている [Vapnik 98, Harvey 17, Arora 18, Suzuki 20]。

連絡先: machida@ks.c.titech.ac.jp, uto@ks.c.titech.ac.jp, shinoda@c.titech.ac.jp, taiji@mist.i.u-tokyo.ac.jp

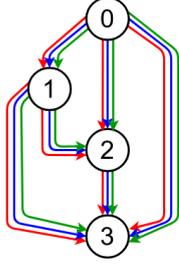


図 1: 4 ノードの有効非巡回グラフ (DAG) で、各エッジが 3 つのオペレーションを持つ場合のグラフ。

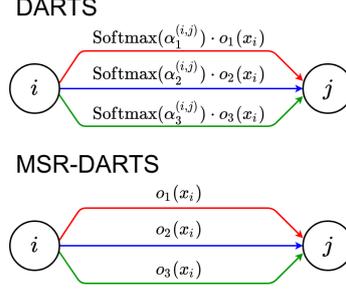


図 2: 従来法 DARTS と提案法 MSR-DARTS の混合エッジの計算。

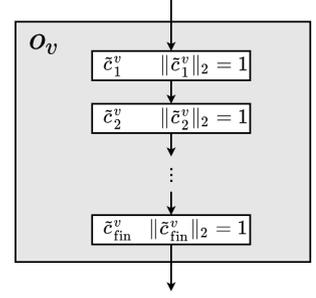


図 3: 各オペレーション  $o_v$  に含まれる畳み込み  $c_p^v$  のスペクトルノルムは定数  $C$  に設定する。本論文では  $C = 1$  とした。

安定ランクは行列の代理指標として用いられ、行列  $A$  の安定ランク  $R(A)$  は  $R(A) = \frac{\|A\|_F^2}{\|A\|_2^2}$  で定義される。ここで、 $\|A\|_F$ 、 $\|A\|_2$  はそれぞれ行列  $A$  のフロベニウスノルム、スペクトルノルムを表す。深層学習モデルの層において、安定ランクが低い層は、入力にランダムノイズが含まれる場合にノイズを減衰させるため、安定ランクが低い層で構成されるモデルは高い汎化性能を持つことが知られている [Arora 18]。

### 2.3 DARTS

DARTS は NAS 手法の 1 つで、探索空間内のモデルの全重みパラメータを同時に最適化する NAS 手法である。DARTS の探索空間は  $N$  ノードの Directed Acyclic Graph (DAG) で定義されたセル (cell) により表現される。DAG 内で、ノード  $x_i$  ( $i \in 0, \dots, N-1$ ) は畳み込みの特徴マップなどの潜在表現であり、ノードを結ぶエッジ  $(i, j)$  ( $i < j$ ) はノード  $i$  からノード  $j$  への特徴マップの流れ及び変形を表す。  $K$  個のオペレーションによる集合  $\mathcal{O}$  を  $\mathcal{O} = \{o_0, o_1, \dots, o_{K-1}\}$  で表し、各オペレーション  $o_v$  ( $v \in \{1, \dots, K-1\}$ ) は畳み込みの重みなどの、学習可能な重みパラメータ  $w_v^{(i,j)}$  を含む。図 1 に 4 ノードの DAG の例を示す。ノード間の各エッジは  $\mathcal{O}$  で定義されたすべてのオペレーションを含む混合エッジであり、図 1 は  $|\mathcal{O}| = 3$  の場合の例である。構造パラメータ  $\alpha_v^{(i,j)}$  は、ノード  $i$  とノード  $j$  をつなぐ混合エッジにおいて、 $v$  番目のオペレーションの適正度を表すパラメータであり、オペレーション  $o_v^{(i,j)}$  を適用した結果への重みとして導入する。 $i < j$  で、ノード  $i$  からノード  $j$  への計算は、図 2 の上図で表され、

$$f_{i,j}(x_i) = \sum_{v=0}^{K-1} \frac{\exp(\alpha_v^{(i,j)})}{\sum_{v'=0}^{K-1} \exp(\alpha_{v'}^{(i,j)})} \cdot o_v(x_i) \quad (1)$$

で表現される。

ノード  $j$  の出力は、ノード  $i$  ( $i < j$ ) からのすべてのエッジにおける  $f_{i,j}(x_i)$  の和、すなわち  $x_j = \sum_{i < j} f_{i,j}(x_i)$  で計算される。セルにおいて、はじめの 2 つのノード  $x_0, x_1$  は入力ノードである。ノード全体の出力は最終ノード  $x_{N-1}$  であり、 $x_{N-1}$  は入力ノードを除くすべての中間ノードのチャネル方向への結合で定義される。DARTS では 2 種類のセルが存在する。それぞれ normal cell, reduction cell であり、前者は入力解像度を保持して出力し、後者は入力解像度を削減して出力する。normal cell は構造パラメータを共有し、reduction cell も同様であるため、最適化対象は  $(\alpha_{\text{normal}})^v^{(i,j)}$ 、 $(\alpha_{\text{reduce}})^v^{(i,j)}$  となる。

DARTS には 2 つのステージが存在し、それぞれ探索段階と評価段階と呼ぶ。探索段階ではセルのエッジを混合エッジとして、 $L_{\text{search}}$  個のセルを積み重ねて学習を行う。学習は重み  $w_v^{(i,j)}$  と構造パラメータ  $\alpha_v^{(i,j)}$  を、一方を固定して交互に更新する。訓練後、構造パラメータに従って簡略化された離散モデル (normal cell と reduction cell) を構築する。離散モデルはノード  $i, j$  間のエッジ強度  $S_{i,j}$  を  $S_{i,j} = \max_{o \in \mathcal{O}} \frac{\exp(\alpha_o^{(i,j)})}{\sum_{o' \in \mathcal{O}} \exp(\alpha_{o'}^{(i,j)})}$  で定義し、各ノードはエッジ強度の強い 2 つの先祖ノードからの接続を採用する。採用されたエッジは  $\alpha_v^{(i,j)}$  が最大となるオペレーションに置き換える。評価段階では、探索段階で最適化したセルを  $L_{\text{eval}}$  個積み重ねてフルスクラッチ学習を行う。評価段階で構築するモデルも、 $L_{\text{eval}}$  個のセルの深さ  $\frac{1}{3}, \frac{2}{3}$  の位置に reduction cell を挿入する。

## 3. 提案手法

### 3.1 MSR-DARTS

DARTS では、畳み込みなどの重みパラメータ  $w_v^{(i,j)}$  と構造パラメータ  $\alpha_v^{(i,j)}$  を交互に更新することでモデル探索を実現した。しかしこのアルゴリズムは、特に学習初期段階で、収束しきっていない重みに従って  $\alpha_v^{(i,j)}$  を更新するため、学習の速いモデルに重みが集中し、精度の面で最適なモデルを探索することができない。そこで本論文では、 $\alpha_v^{(i,j)}$  を 1 に固定して、混合エッジを図 2 の下図のように変更し、重み  $w_v^{(i,j)}$  のみを学習させる。探索モデル学習後は、汎化性能誤差解析に用いられる安定ランク (Stable Rank) を用いて精度の面で最適なモデルを構築する MSR-DARTS を提案する。

本論文では、オペレーション  $o_v \in \mathcal{O}$  は複数の畳み込みで構成されると仮定し、 $c_p^{(i,j)v}$ 、 $c_{\text{fin}}^{(i,j)v}$  をそれぞれ、ノード  $i, j$  間の混合エッジの  $o_v$  を構成する  $p$  番目の畳み込み、 $o_v$  を構成する畳み込みの最終層と定義する。探索モデルの重み  $w_v^{(i,j)}$  の学習後、各混合エッジで最適なオペレーション  $o_v$  を決定する際、汎化性能と安定ランクの関係を利用し、畳み込みを行列計算による表現とみなして安定ランクの最も低いオペレーションを採用する。安定ランクの低いオペレーションは、層への入力にノイズが含まれる場合に、ノイズを減衰させつつ入力情報を伝搬する効果が期待され [Arora 18]、これにより安定ランクの低いオペレーションを選択することで汎化性能の高いモデルの構築が期待できる。

提案法では、学習を安定させるため、図 3 に示すように、各オペレーション  $o_v$  に含まれるすべての畳み込み  $c_p^{(i,j)v}$  のスペ

表 1: CIFAR10 データセットにおける DARTS と MSR-DARTS の比較 .

Architecture	Test Err. (%)	Test Loss	Params (M)
MSR-DARTS	<b>2.84</b>	<b>0.124</b>	<b>1.63</b>
DARTS	3.13	0.134	1.65

クトルノルムを 1 に調整しながら探索モデルの学習を行う . 探索モデル学習後 , 各ノード  $i, j$  間の混合エッジで , 学習後のオペレーション  $o_v$  の畳み込みの最終層  $c_{\text{fin}}^{(i,j)v}$  の安定ランクのセルにわたる平均が最も低いオペレーション , すなわち

$$\forall i, j (i < j) \arg \min_{o_v \in \mathcal{O}} \bar{R}_{\mathcal{T}}(c_{\text{fin}}^{(i,j)v}) \quad (2)$$

を採用する . ここで  $\mathcal{T}$  はセルの種類をあらわし ,  $\mathcal{T} \in \{\text{normal cell, reduction cell}\}$  であり ,  $\bar{R}_{\mathcal{T}}$  はセルタイプ  $\mathcal{T}$  のセルにわたる安定ランクの平均を表す . また , normal cell , reduction cell それぞれで , セルの構造パラメータを共有するため , 安定ランクを各セルタイプに渡り平均したことに注意されたい . また , ノード  $i, j$  間のエッジ強度  $S'_{i,j}$  も同様に  $c_{\text{fin}}^{(i,j)v}$  の安定ランクで定義し ,

$$S'_{i,j} = \max_{o_v \in \mathcal{O}} -\bar{R}_{\mathcal{T}}(c_{\text{fin}}^{(i,j)v}) \quad (3)$$

と定義する . DARTS と同様 , normal cell と reduction cell の各ノードはエッジ強度  $S'_{i,j}$  の強い 2 つの先祖ノードからの接続を採用する .

## 4. 実験

提案法 MSR-DARTS を画像分類タスクに適用し , CIFAR10 [Krizhevsky 09] と ImageNet [Deng 09] の 2 つのデータセットで実験を行った . CIFAR10 は縦横 32 ピクセルの小さな画像を 10 クラスに分類するタスクで , 各クラス 5000 枚の訓練画像を含む . ImageNet は 1000 クラス用意され , 訓練データは , 縦横様々なサイズの画像を 1.28M 枚含む . 本論文では縦横 224 にリサイズして入力を行った .

本実験でのオペレーションセット  $\mathcal{O}$  は , DARTS で定義されたオペレーションセットに含まれるもののうち , 畳み込みに限定したもの , すなわち  $\{\text{“sep conv 3x3”}, \text{“sep conv 5x5”}, \text{“dil conv 3x3”}, \text{“dil conv 5x5”}\}$  の 4 つのオペレーションを含むものとした . 実験は GPU (TITAN RTX) を 1 枚用いて行った . また , トレーニングセットを 4:1 に分割し , 前者を探索モデルの訓練 , 後者を探索モデルのバリデーションに利用した . 探索モデルのバリデーションエラーが最も小さかったエポック終了時点での重みを利用し , 3.1 節に従って離散モデルの抽出を行った .

### 4.1 CIFAR10

DARTS と実験設定を揃え , 探索時と評価時のセル数  $L_{\text{search}}$  と  $L_{\text{eval}}$  をそれぞれ 8 とした結果を表 1 に示す . 提案法はパラメータ数 1.6M のモデルを 0.3 日で探索し , テストエラー精度 2.84% を達成した . DARTS と比較して 0.29 ポイントのテストエラー精度を改善し , より汎化性能の高いモデルを探索した .

表 2: CIFAR10 データセットにおける DARTS 由来の state of the art 手法と MSR-DARTS の比較 .

Architecture	Test Err. (%)	Params (M)	GPU-days
DARTS	2.76	3.3	1.0
P-DARTS	2.50	3.4	0.3
PC-DARTS	2.57	3.6	0.1
MSR-DARTS	2.84	<b>1.65</b>	0.3

また , 探索モデルにおいてセルごとに最適化した場合の比較を表 2 に示す . 提案法 MSR-DARTS はパラメータ数 1.65M のモデルを探索し , テストエラー精度 2.80% を達成した . DARTS と比較するとパラメータ数が半分のモデルで同等の精度を達成し , 探索コストを 3 分の 1 に削減した .

### 4.2 ImageNet

探索時のセル数  $L_{\text{search}}$  を 14 に設定して実験を行った . また , 評価時のセル数と  $L_{\text{eval}}$  を 14 とし , cosine スケジューリングで学習率減衰を施しながら 300 エポックの学習を行った . その他の実験設定は DARTS に従った . 提案法では , モデル探索に , CIFAR を用いる実験と ImageNet データセットのサブセットを用いる実験の 2 つを行った . 結果の比較を表 3 に示す . 提案法は , CIFAR データセットを探索に用いた場合 , モデル探索を 0.5 日で終え , Top-1 テストエラー精度 24.0% , Top-5 テストエラー精度 7.2% を達成した . また , ImageNet データセットを探索に用いた場合 , モデル探索を 2.6 日で終え , Top-1 テストエラー精度 24.1% , Top-5 テストエラー精度 7.7% を達成した .

## 5. 結論

現在注目されている NAS 手法 DARTS において , 構造パラメータを学習せず固定して畳み込みなどの重みだけを学習し , モデル構造最適化に安定ランクを用いる MSR-DARTS を提案した . 提案法を画像分類タスクを解く深層学習モデルに対して適用した . CIFAR10 , ImageNet データセットを用いて評価実験を行い , それぞれ 2.80% , 24.0% のテストエラー精度を達成した .

## 謝辞

This work was supported by JST CREST JPMJCR1687 and NEDO JPNP18002. TS was partially supported by MEXT Kakenhi (15H05707, 18K19793 and 18H03201), and Japan Digital Design.

## 参考文献

[Arora 18] Arora, S., Ge, R., Neyshabur, B., and Zhang, Y.: Stronger Generalization Bounds for Deep Nets via a Compression Approach, in Dy, J. G. and Krause, A. eds., *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10-15, 2018*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 254-263, PMLR (2018)

表 3: ImageNet データセットにおける DARTS 由来の state of the art 手法と MSR-DARTS の比較 . . † は探索に ImageNet データセットを直接利用したことを表す.

Architecture	Test Err.(%)		Params (M)	×+ (M)	Search Cost GPU-days	Search Method
	top-1	top-5				
DARTS (2nd order) [Liu 19]	26.7	8.7	4.7	574	4.0	gradient-based
P-DARTS(CIFAR-10) [Chen 19]	24.4	7.4	4.9	557	0.3	gradient-based
P-DARTS(CIFAR-100) [Chen 19]	24.7	7.5	5.1	577	0.3	gradient-based
PC-DARTS(CIFAR-10) [Xu 19]	25.1	7.8	5.3	586	0.1	gradient-based
PC-DARTS(ImageNet) † [Xu 19]	24.2	7.3	5.3	597	3.8	gradient-based
MSR-DARTS(CIFAR-10) (ours)	24.0	7.2	6.7	766	0.5	Stable Rank
MSR-DARTS(ImageNet) (ours) †	24.1	7.7	6.6	748	2.6	Stable Rank

- [Chen 19] Chen, X., Xie, L., Wu, J., and Tian, Q.: Progressive differentiable architecture search: Bridging the depth gap between search and evaluation, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1294–1303 (2019)
- [Deng 09] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database, in *CVPR09* (2009)
- [Harvey 17] Harvey, N., Liaw, C., and Mehrabian, A.: Nearly-tight VC-dimension bounds for piecewise linear neural networks, in Kale, S. and Shamir, O. eds., *Proceedings of the 2017 Conference on Learning Theory*, Vol. 65 of *Proceedings of Machine Learning Research*, pp. 1064–1068, Amsterdam, Netherlands (2017), PMLR
- [Krizhevsky 09] Krizhevsky, A.: Learning multiple layers of features from tiny images, Technical report (2009)
- [Liang 19] Liang, H., Zhang, S., Sun, J., He, X., Huang, W., Zhuang, K., and Li, Z.: Darts+: Improved differentiable architecture search with early stopping, *arXiv preprint arXiv:1909.06035* (2019)
- [Liu 18] Liu, H., Simonyan, K., Vinyals, O., Fernando, C., and Kavukcuoglu, K.: Hierarchical Representations for Efficient Architecture Search, in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net (2018)
- [Liu 19] Liu, H., Simonyan, K., and Yang, Y.: DARTS: Differentiable Architecture Search, in *International Conference on Learning Representations* (2019)
- [Pham] Pham, H., Guan, M. Y., Zoph, B., Le, Q. V., and Dean, J.: Efficient Neural Architecture Search via Parameter Sharing
- [Real 19] Real, E., Aggarwal, A., Huang, Y., and Le, Q. V.: Regularized Evolution for Image Classifier Architecture Search, in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 4780–4789, AAAI Press (2019)
- [Suzuki 20] Suzuki, T., Abe, H., and Nishimura, T.: Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network, in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net (2020)
- [Tang 17] Tang, J., Golbabaee, M., and Davies, M. E.: Gradient Projection Iterative Sketch for Large-Scale Constrained Least-Squares, in Precup, D. and Teh, Y. W. eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Vol. 70 of *Proceedings of Machine Learning Research*, pp. 3377–3386, PMLR (2017)
- [Vapnik 98] Vapnik, V. N.: *Statistical Learning Theory*, Wiley-Interscience (1998)
- [Xie 19] Xie, S., Zheng, H., Liu, C., and Lin, L.: SNAS: stochastic neural architecture search, in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net (2019)
- [Xu 19] Xu, Y., Xie, L., Zhang, X., Chen, X., Qi, G.-J., Tian, Q., and Xiong, H.: Pc-darts: Partial channel connections for memory-efficient differentiable architecture search, *arXiv preprint arXiv:1907.05737* (2019)
- [Zoph 17] Zoph, B. and Le, Q. V.: Neural Architecture Search with Reinforcement Learning, in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, OpenReview.net (2017)
- [Zoph 18] Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V.: Learning transferable architectures for scalable image recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8697–8710 (2018)