XNAS: Neural Architecture Search with Expert Advice

Asaf Noy

Machine Intelligence Israel Lab (MIIL),

DAMO Academy, Alibaba

Worldwide Cloud Services Partner

DAMO academy

Discovery, Adventure, Momentum and Outlook

MIIL

Israel Machine Intelligence Lab

Alibaba Cloud | Worldwide Cloud Services Partner

https://damo.alibaba.com

Alibaba Cloud | Worldwide Cloud Services Partner

AI is still beyond reach to most companies & people

Even tech-giants are struggling to answer the need in new fields

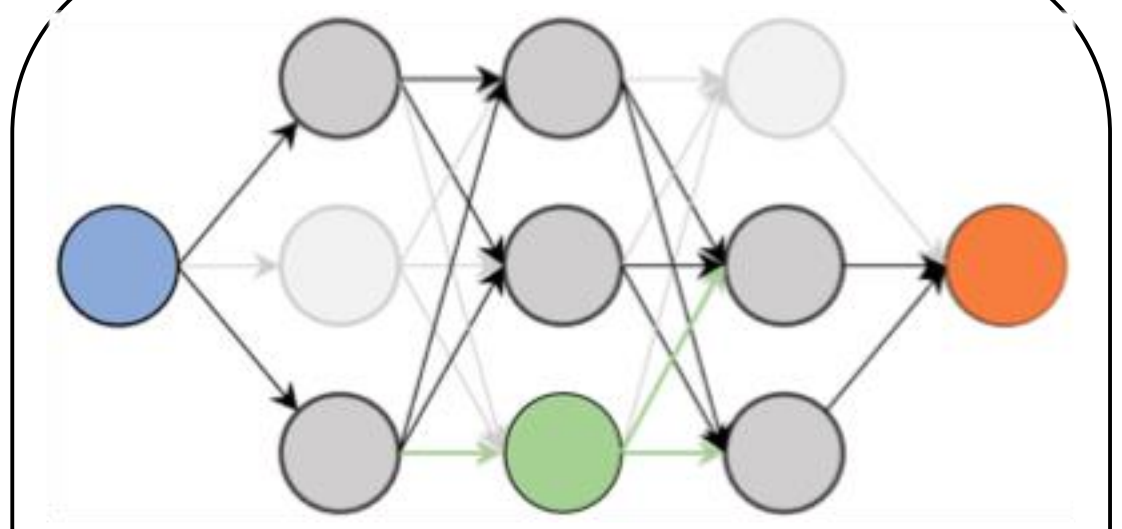
AI Experts are busy tuning parameters

Alibaba Cloud | Worldwide Cloud Services Partner

AutoML::Classification

Upload your data

AutoML



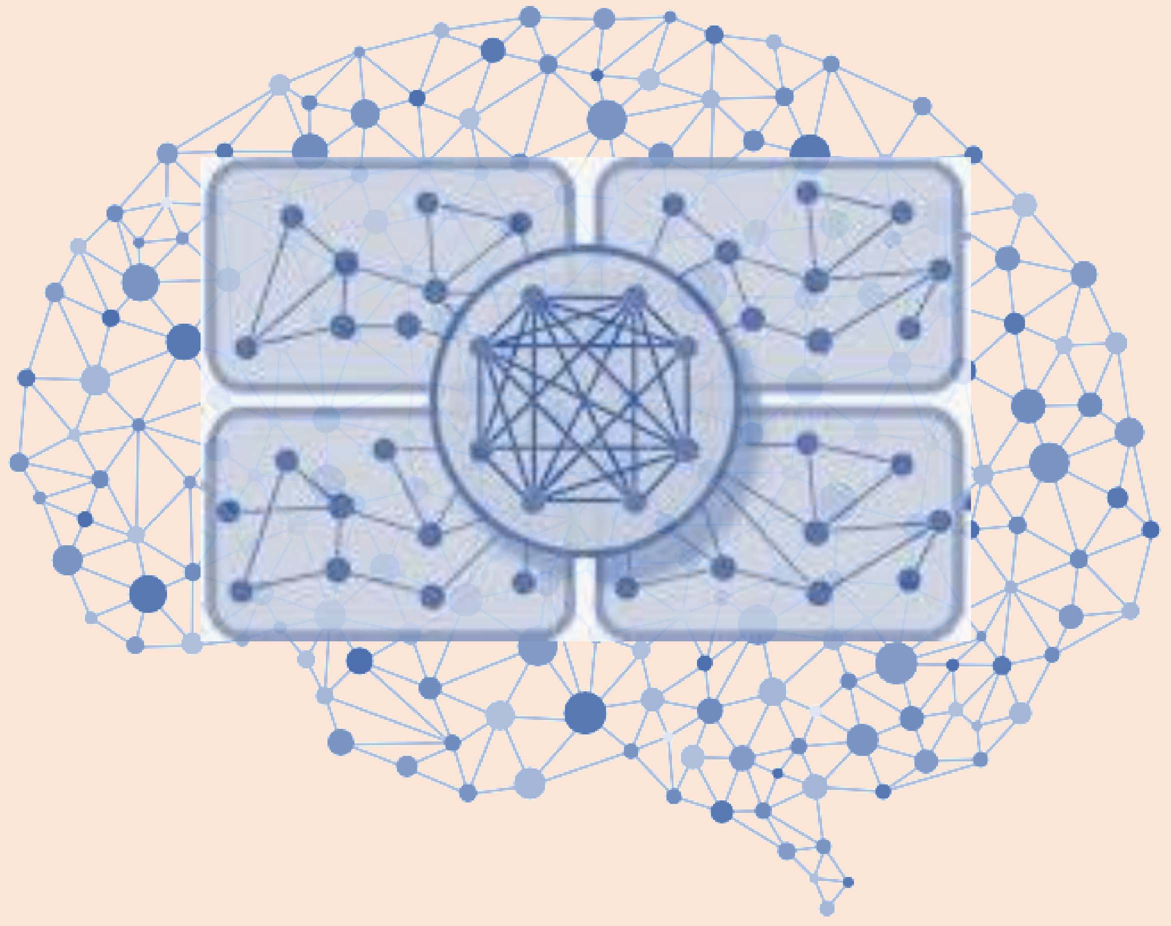
State-of-the-Art Model & Feedback

NAS Overview

- The success of deep learning in perceptual tasks is largely due to its automation of the feature engineering process
- This success led to a rising demand for architecture engineering
- NAS, architecture engineering automation, is a logical next step in the mission of fully automating machine learning

NAS

Automatic **N**eural **A**rchitecture **S**earch



NAS Overview

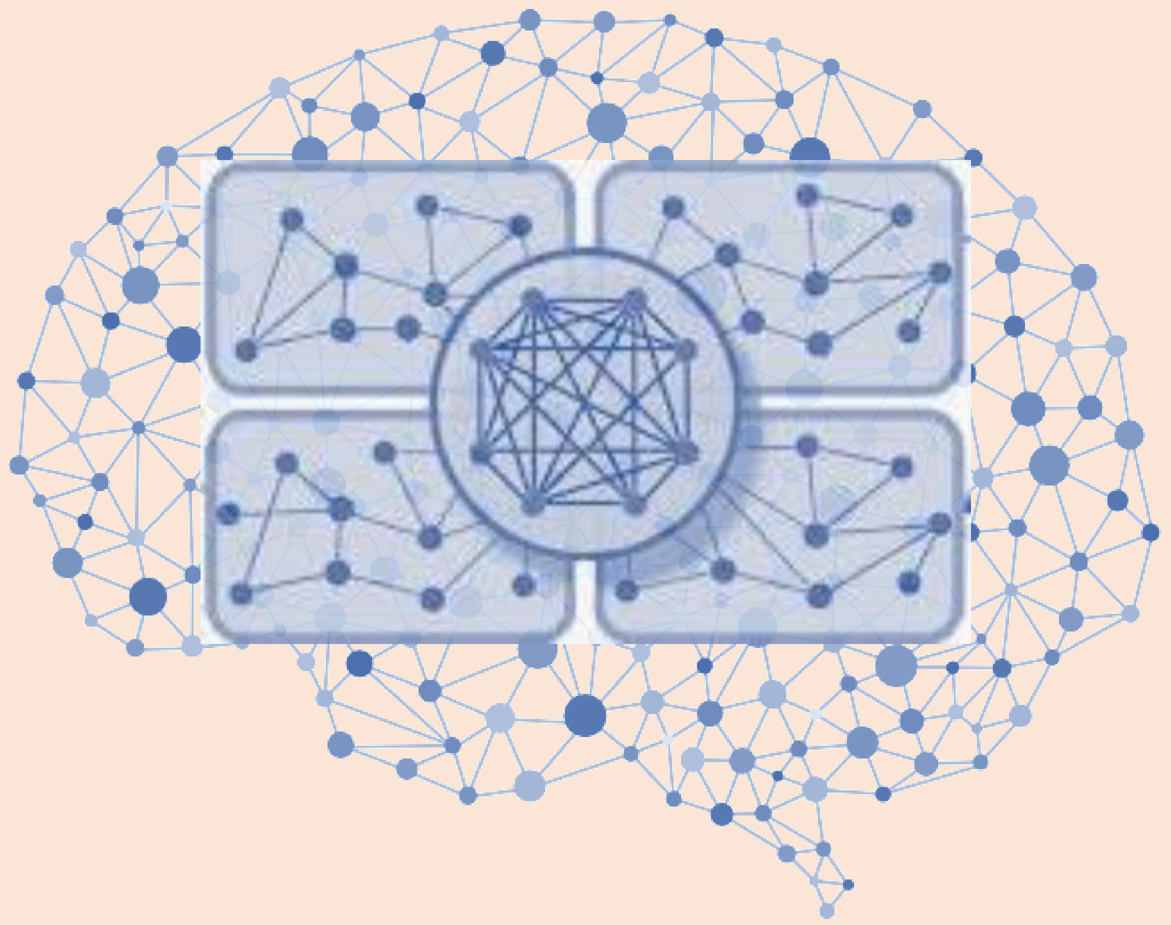
- The goal: Dataset → Architecture
- A difficult optimization problem
  - Expensive evaluations
  - A huge categorial space
- Existing solutions are biased towards current human understanding of Neural Networks structure
- Once, a game for tech-giants only

NAS [Zoph]	AmoebaNet [Liu]	DARTS [Liu]	XNAS (ours)
22,000 days	3,150 days	1 day	4-6 hours
2016	2017	2018	2019

Search duration over CIFAR-10 with a single GPU.

NAS

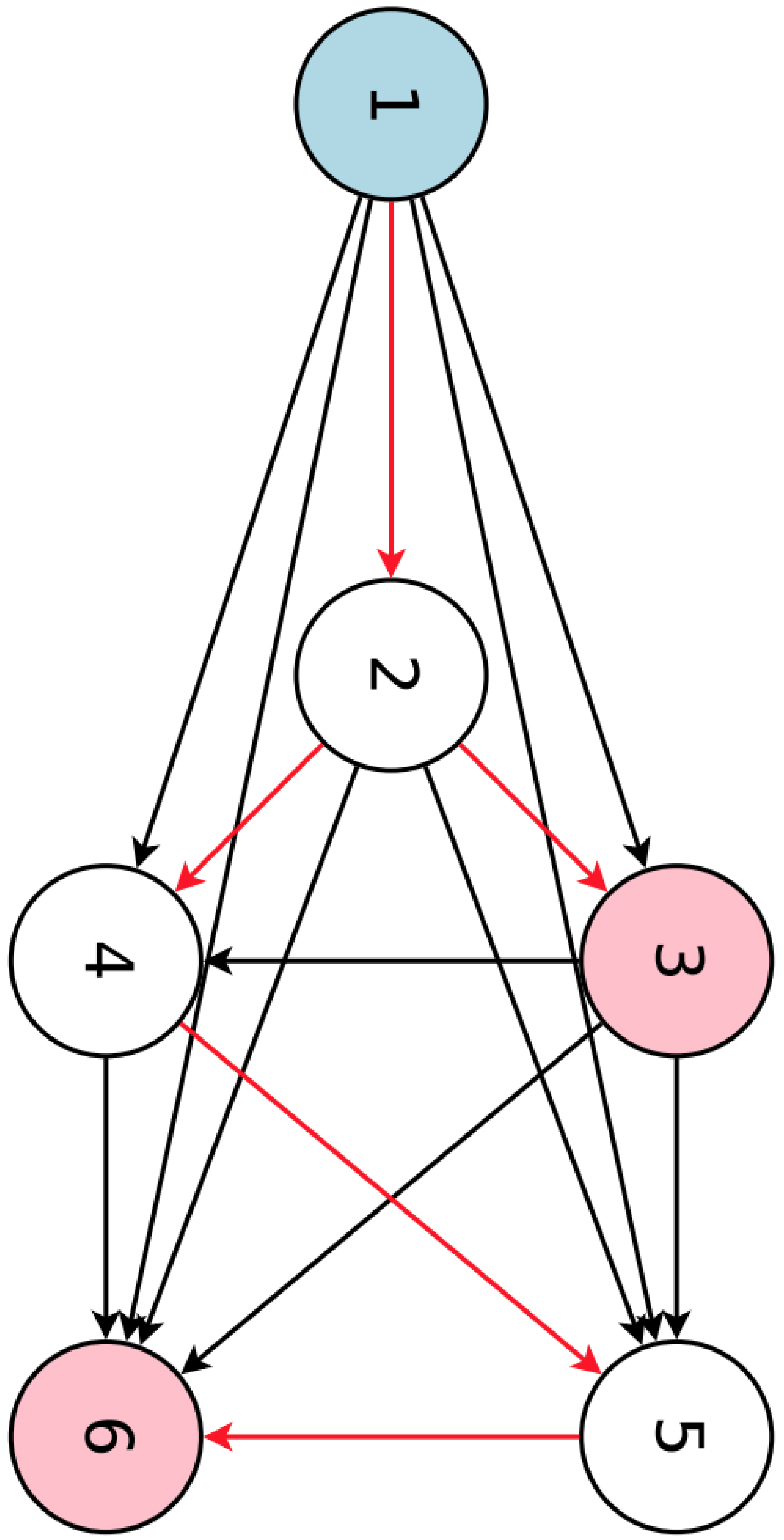
Automatic **N**eural **A**rchitecture **S**earch



Architecture Space and Optimization

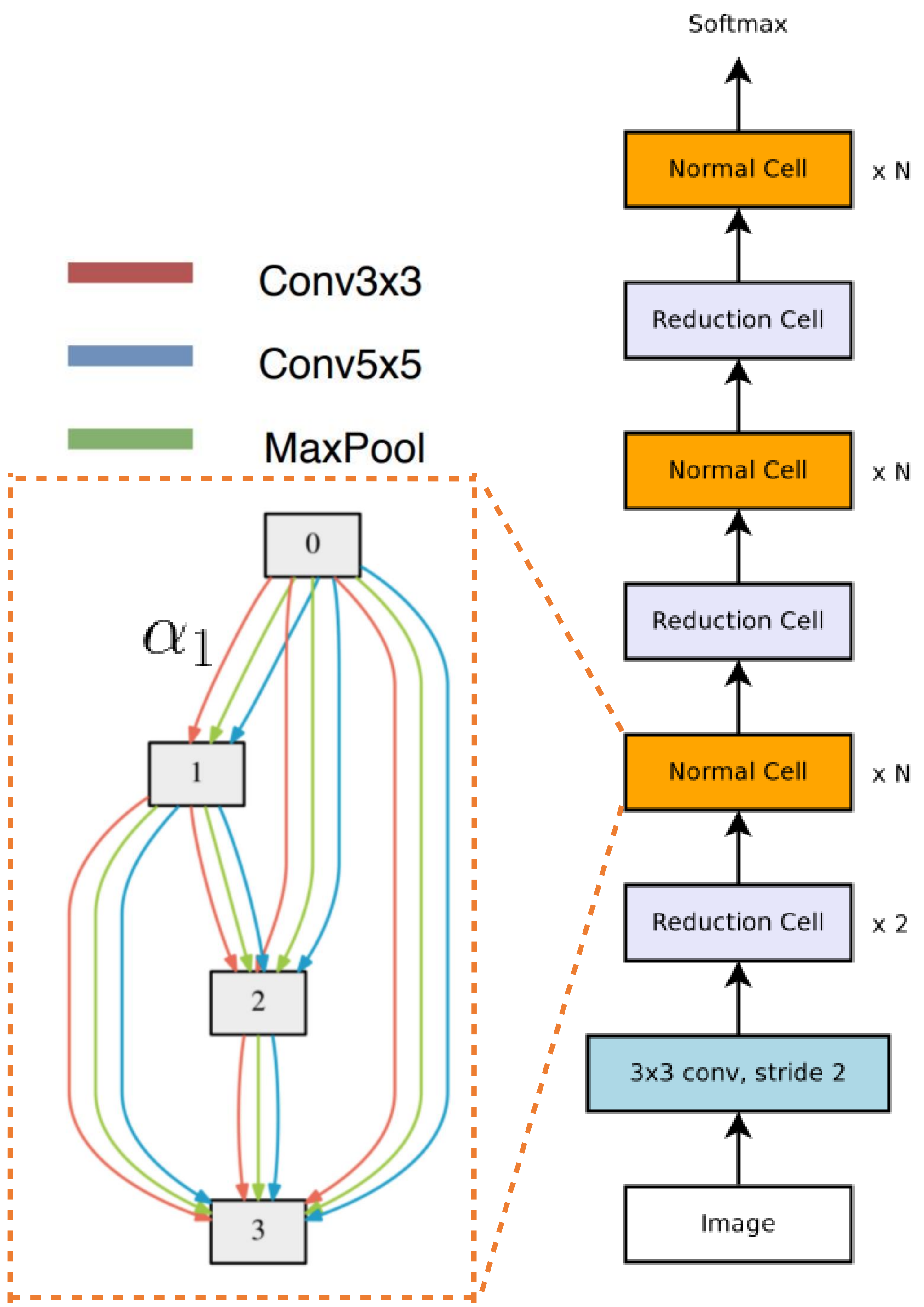
- The architectures space is represented by a super-graph
  - Nodes are features-maps (tensors)
  - Edges are operations over tensors (layers)
  - Paths are architectures
  - Space size is the number of different paths (10^{30})

- The *NAS objective*:
  - Select the path which maximizes the validation accuracy
    - Paths are sampled and scored via different techniques



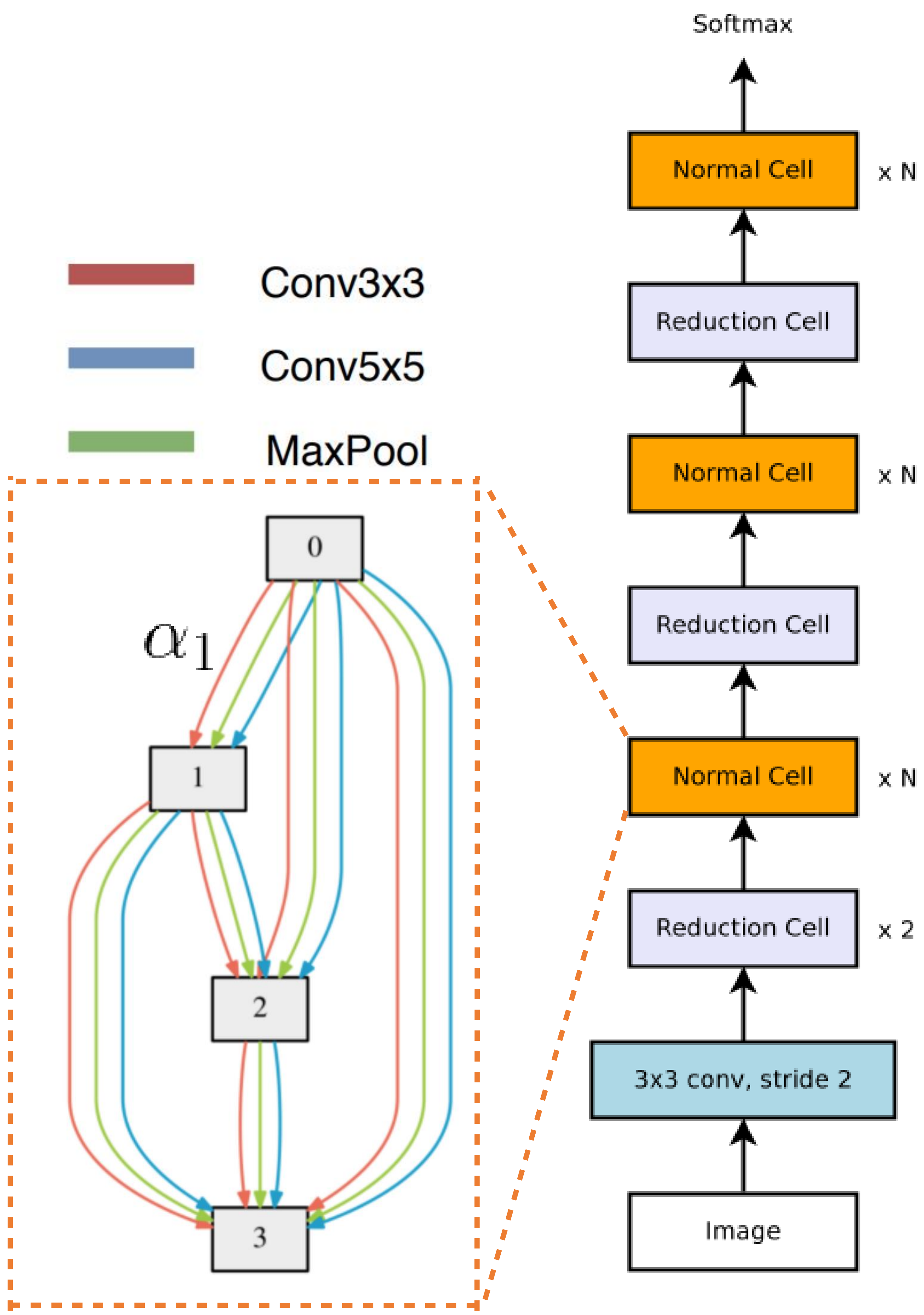
Differential Architecture Space

- The search space can be reduced to a sub-space of repetitive *cells* [NASNet, Zoph 2017]
- DARTS replaced path-sampling with super-graph training [Liu, 2018]
  - Introduced architecture-weights **α**
  - The sampling scheme is relaxed to *joint maximization*
  - Efficient search via gradient-descent

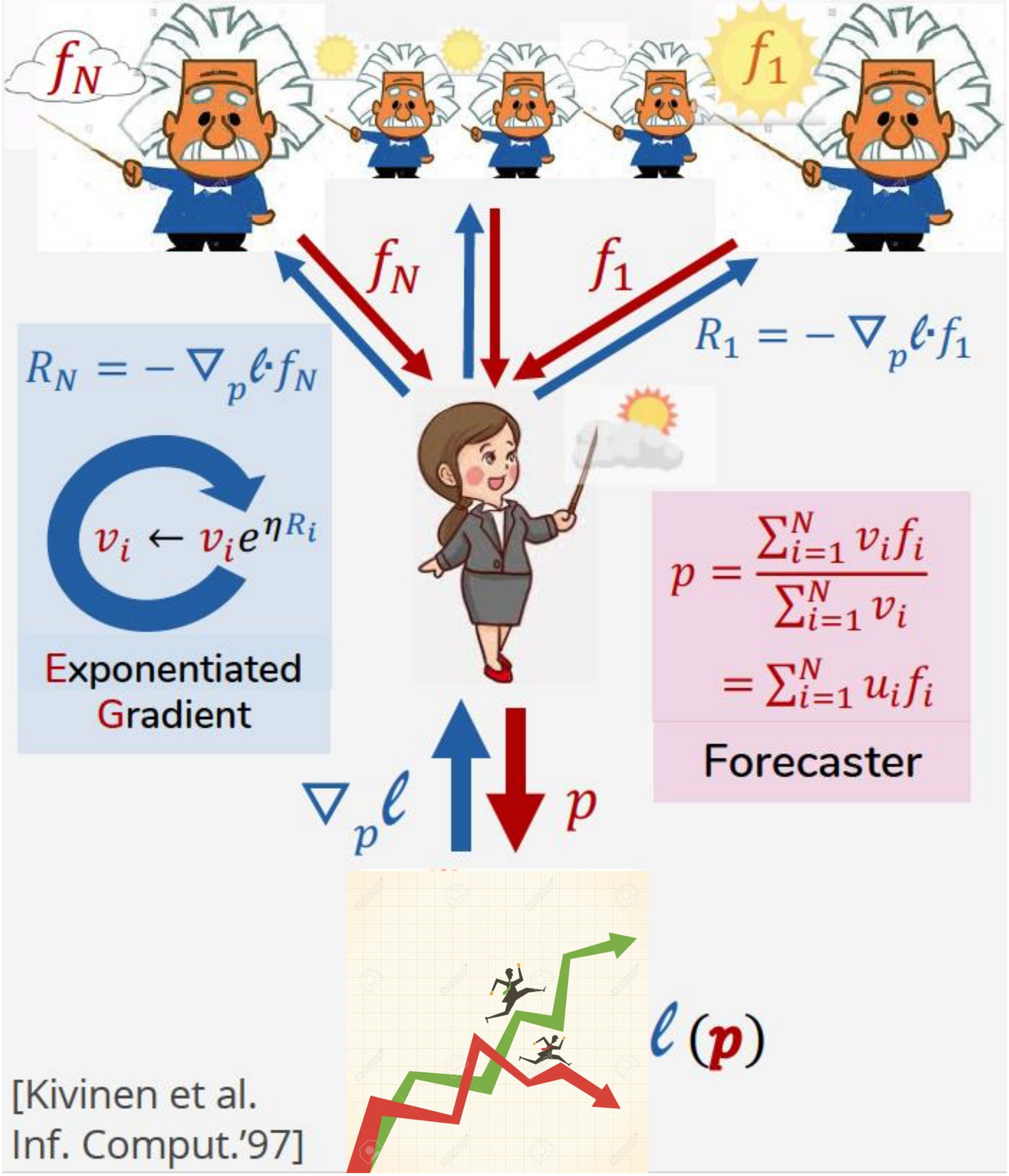


Differential Architecture Space

- DARTS Search algorithm:
  - For steps 1..T do:
    - Gradient-descent step over network weights **w**
    - Gradient-descent step over Architecture weights **α**
  - Prune all operations except the best ones (largest **α**)
- Output architectures are practically *random [Li, 2019]*
- We argue that this optimization process is *inefficient*
  1. Start: Diverse operations → Parameterization bias
  2. End: Harsh final pruning → Relaxation bias
- To address that, we ask for ***expert advice***



Prediction with expert advice



Prediction with expert advice

Alibaba Cloud | Worldwide Cloud Services Partner

- a forecaster relies on the advice of N experts,

  forming an attention-vector,

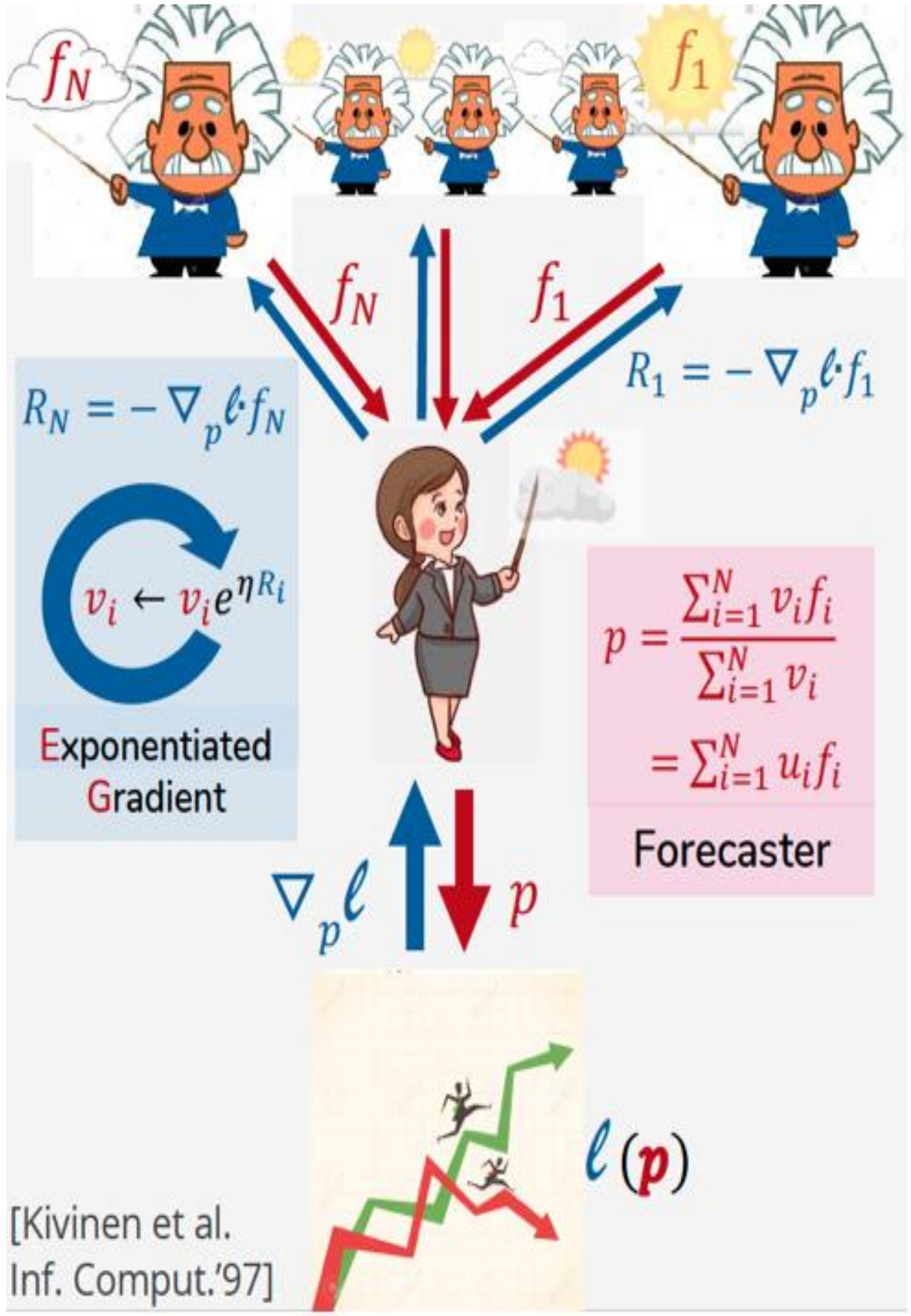
$$\Delta_t = \{ \boldsymbol{w} \in \mathbb{R}^N : \sum_{i=1}^{N} \boldsymbol{w}_i = 1,\ \boldsymbol{w} \geq 0 \}\ ,\ \hat{h_t} = \sum_{i=1}^{N} \boldsymbol{w}_{i,t} \hat{f_{i,t}}(\boldsymbol{x}_t)$$

- Define the *accumulated regret*,

$$R_{T,N} = \sum_{t=1}^{T} \ell(\hat{h}_t, y_t) - \min_{i=1,..,n} \sum_{t=1}^{T} \ell_{i,t}$$

  - A measure for the forecaster regret for not following best expert's advice, in hindsight
  - We optimize the mixture, and select the best operation at the end

➢ Our experts represent **operations (layers)** and our forecaster is their **mixture tensor**

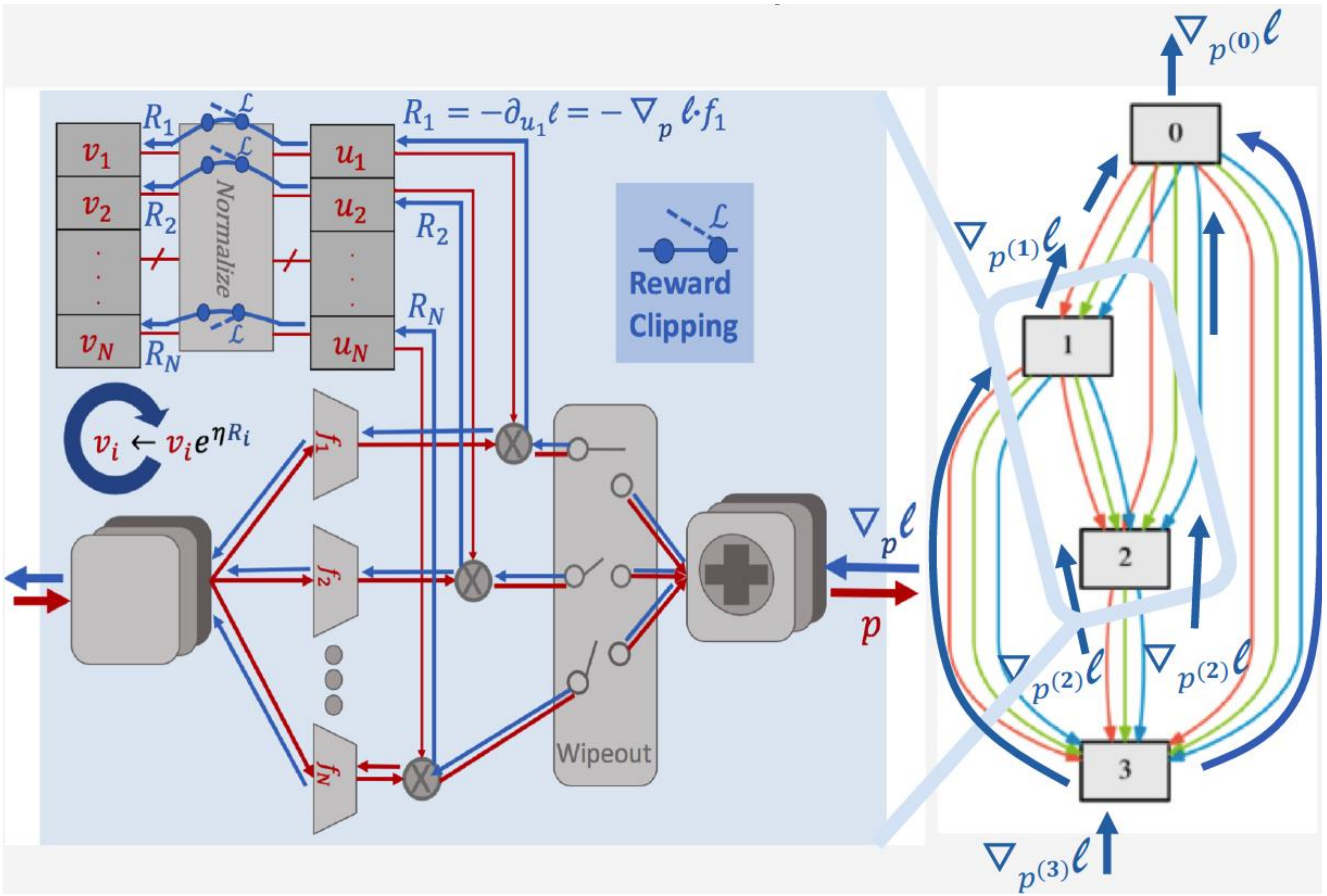


[Kivinen et al. Inf. Comput.'97]

XNAS: NAS with eXpert Advice

Algorithm 1 XNAS for a single forecaster

Input: step size η,
loss-gradient bound $\mathcal{L}$,
Experts predictions $\{f_{t,i}\}_{i=1}^{N}$ $\forall t = 1, \ldots, T$
Init: $I_0 = \{1, \ldots, N\}$, $v_{0,i} \leftarrow 1$, $\forall i \in I_0$
for rounds $t = 1, \ldots, T$ **do**
    Update $\boldsymbol{\omega}$ by descending $\nabla_{\boldsymbol{\omega}} \ell_{\text{train}}(\boldsymbol{\omega}, \boldsymbol{v})$
    $p_t \leftarrow \frac{\sum_{i \in I_{t-1}} v_{t-1,i} \cdot f_{t-1,i}}{\sum_{i \in I_{t-1}} v_{t-1,i}}$ #Predict
    {loss gradient revealed: $\nabla_{p_t} \ell_{\text{val}}(p_t)$}
    for $i \in I_{t-1}$ **do**
        $R_{t,i} = -\nabla_{p_t} \ell_{\text{val}}(p_t) \cdot f_{t,i}$ #Rewards
        $v_{t,i} \leftarrow v_{t-1,i} \cdot \exp\{\eta R_{t,i}\}$ #EG step
    end for
    $\theta_t \leftarrow \max_{i \in I_{t-1}} \{v_{t,i}\} \cdot \exp\{-2\eta \mathcal{L}(T - t)\}$
    $I_t \leftarrow I_{t-1} \setminus \{i \mid v_{t,i} < \theta_t\}$ #Wipeout
end for



XNAS: Wipeout

- Wipeout is a *safe* procedure,

Lemma 1. *In XNAS, the optimal expert in hindsight cannot be wiped-out.*

- *Advantages* of dynamic wipeout of inferior operations,
  1. Speeds up the search
  2. Decreases the network's complexity
  3. Mitigates the relaxation bias

Algorithm 1 XNAS for a single forecaster

Input: step size η,
loss-gradient bound $\mathcal{L}$,
Experts predictions $\{f_{t,i}\}_{i=1}^{N}$ $\forall t = 1, \ldots, T$
Init: $I_0 = \{1, \ldots, N\}$, $v_{0,i} \leftarrow 1$, $\forall i \in I_0$
for rounds $t = 1, \ldots, T$ **do**
  Update $\boldsymbol{\omega}$ by descending $\nabla_{\boldsymbol{\omega}} \ell_{\text{train}}(\boldsymbol{\omega}, \boldsymbol{v})$
  $p_t \leftarrow \frac{\sum_{i \in I_{t-1}} v_{t-1,i} \cdot f_{t-1,i}}{\sum_{i \in I_{t-1}} v_{t-1,i}}$ #Predict
  $\{$loss gradient revealed: $\nabla_{p_t} \ell_{\text{val}}(p_t)\}$
  for $i \in I_{t-1}$ **do**
    $R_{t,i} = -\nabla_{p_t} \ell_{\text{val}}(p_t) \cdot f_{t,i}$ #Rewards
    $v_{t,i} \leftarrow v_{t-1,i} \cdot \exp\{\eta R_{t,i}\}$ #EG step
  end for
  $\theta_t \leftarrow \max_{i \in I_{t-1}} \{v_{t,i}\} \cdot \exp\{-2\eta \mathcal{L}(T-t)\}$
  $I_t \leftarrow I_{t-1} \setminus \{i \mid v_{t,i} < \theta_t\}$ #Wipeout
end for

XNAS: Theoretical guarantees

- The *aggregated wipeout factor measures the extent of the wipeout,*

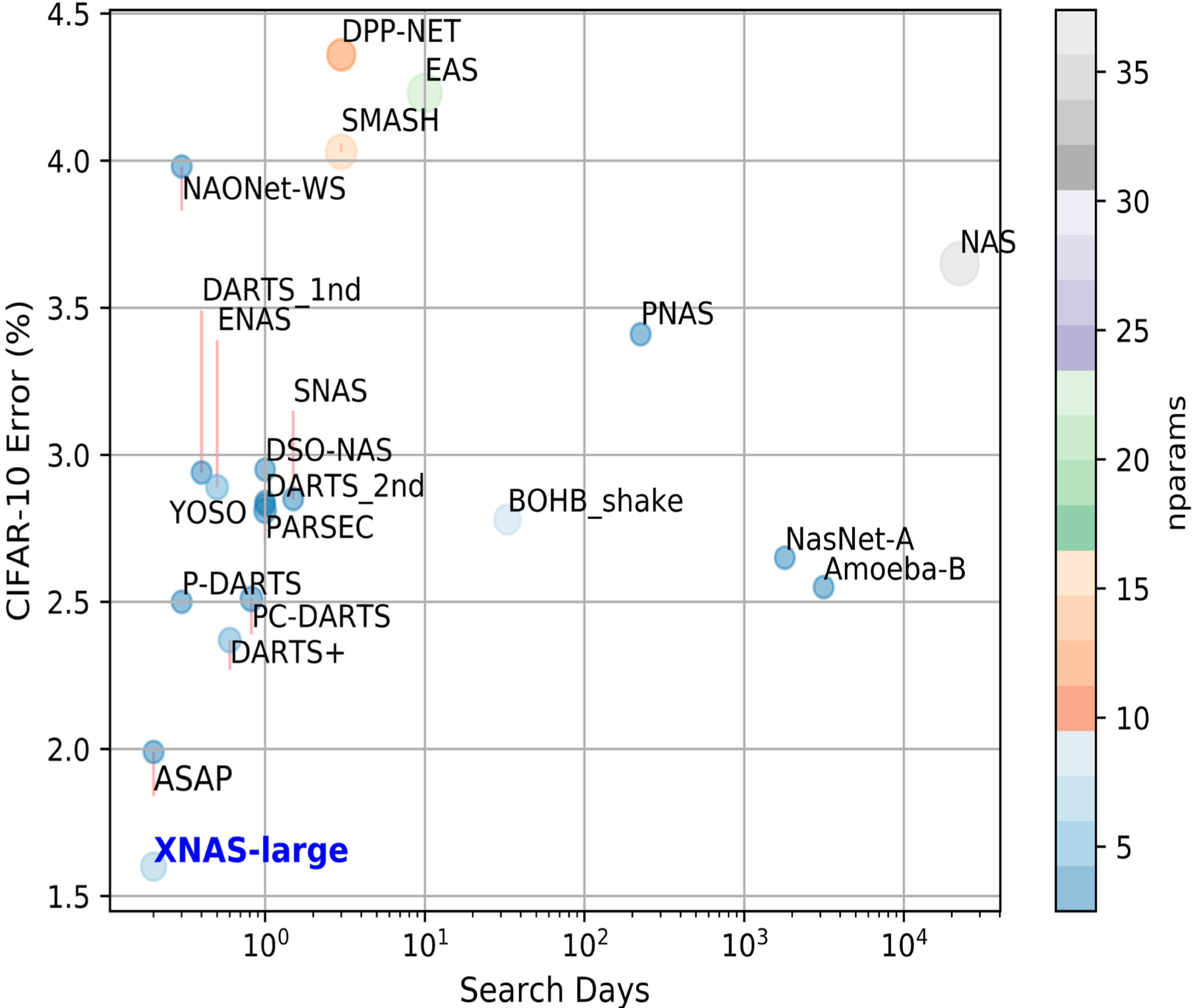
$$\gamma_t := \prod_{t=1}^{T} \frac{\sum_{i \in I_{t-1}} v_{t,i}}{\sum_{i \in I_t} v_{t,i}}$$

- Tight worst-case regret upper-bound is achieved,

Theorem 1 (XNAS Regret Bound). *The regret of the XNAS algorithm 1, with N experts and learning rate η, incurring a sequence of T non-negative convex losses of $\mathcal{L}$-bounded rewards, satisfies,*

$$\eta^* = \sqrt{\frac{2 \ln N}{T \mathcal{L}^2}} \quad ; \quad \text{Regret}_T \leq \mathcal{L}\sqrt{2T \ln N}\left(1 - \frac{1}{2}\frac{\ln \gamma_T}{\ln N}\right) \tag{3}$$

Results: CIFAR-10



Results

- Public datasets:

Datasets Architecture	CIFAR100 Error	FMNIST Error	SVHN Error	Freiburg Error	CINIC10 Error	ImageNet Error	Search cost
Known SotA	**10.7** (3)	3.65 (24)	**1.02** (3)	10.7 (14)	6.83 (14)	**15.6 (6)**	-
SNAS (22)	16.5	3.72	1.98	14.7	7.13	27.3	1.5
PNAS (10)	15.9	3.72	1.83	12.3	7.03	25.8	150
Amoeba-A (16)	15.9	3.8	1.93	11.8	7.18	25.5	3150
NASNet (25)	15.8	3.71	1.96	13.4	6.93	26.0	1800
DARTS (11)	15.7	3.68	1.95	10.8	6.88	26.7	1
ASAP (14)	15.6	3.71	1.81	10.7	6.83	26.7	0.2
XNAS	13.6	**3.64**	1.72	**6.3**	**6.0**	23.9	0.3

- Internal Alibaba datasets:

➢ Competitive results with tailor-made models in several tasks

➢ State-of-the-art results with ‘*AliExpress*’ : ***1,000,000 classes*, 86% accuracy!**

Thanks!

Alibaba Cloud | Worldwide Cloud Services Partner

Have a computer vision task? Give us data and get predictions (for free)

Want to know more about AutoML? Stay tuned for future events by **MIIL**

Interested at what we do? Let’s get in touch: **Asaf.noy@alibaba-inc.com**