Memory-Efficient Models for Scene Text Recognition via Neural Architecture Search

SeulGi Hong DongHyun Kim Min-Kook Choi

hutom

NASFW 2020

WACV 2020

March 1-5, Aspen, CO

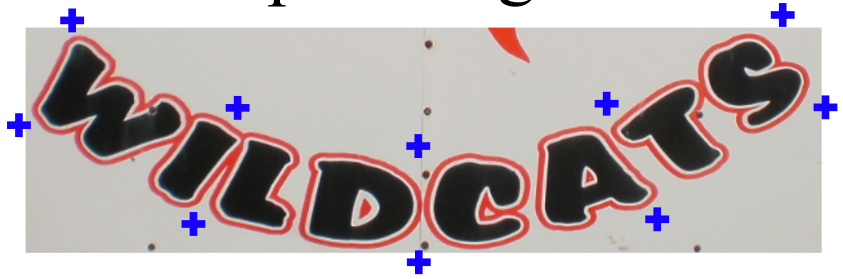
Motivation

- **What is STR:**
  - Scene Text Recognition
  - on diverse appearances or in imperfect conditions

- **Why STR + NAS:**
  - **Automate the process** of designing STR model
  - **Expand the field** of NAS

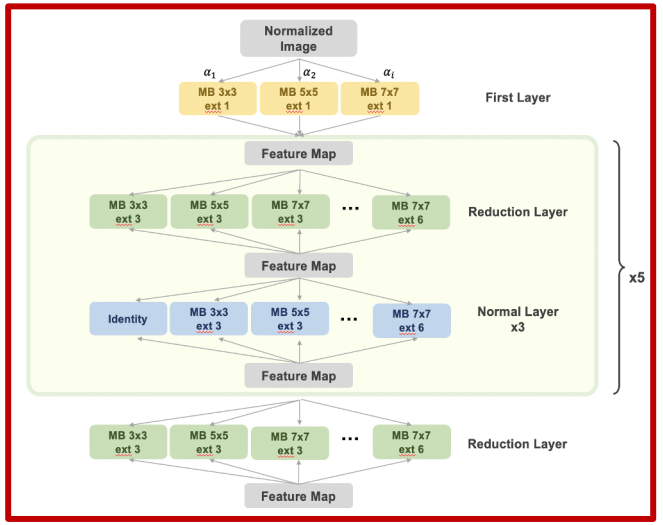
Pipeline of Scene Text Recognition:

Input Image X



Transformation Stage

Feature Extraction Stage



Sequence Modeling Stage

Prediction Stage

Prediction Y

WILDCATS

Our Contributions:

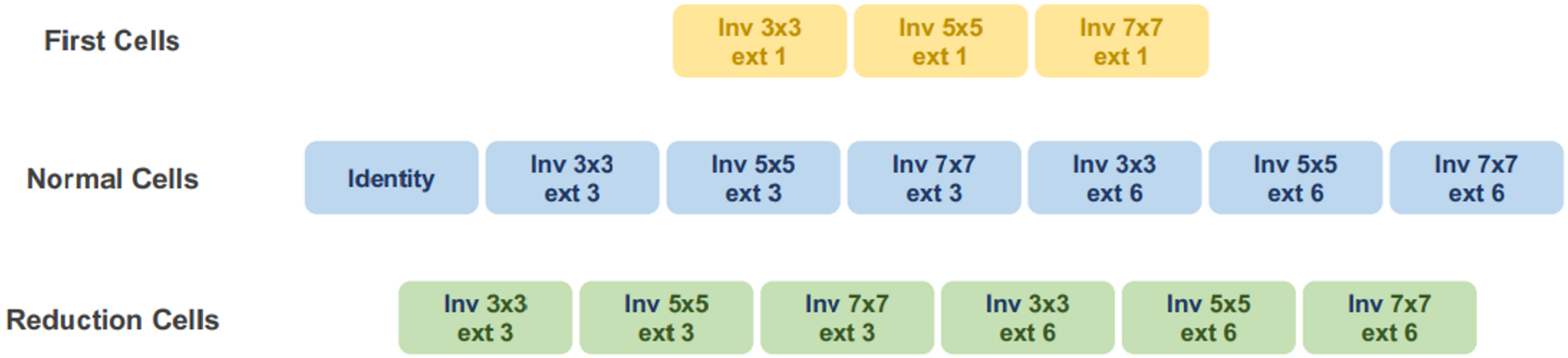
1. **Expand the application field** of NAS
2. NAS for CNN based modules (**transformation, feature extraction**) of STR
3. Proxyless approach;
   let meta-learner **directly find well-adapted modules** on STR scenario

Method:

- Architecture Space
- Rectifier Search
- Feature Extractor Search
- How to train

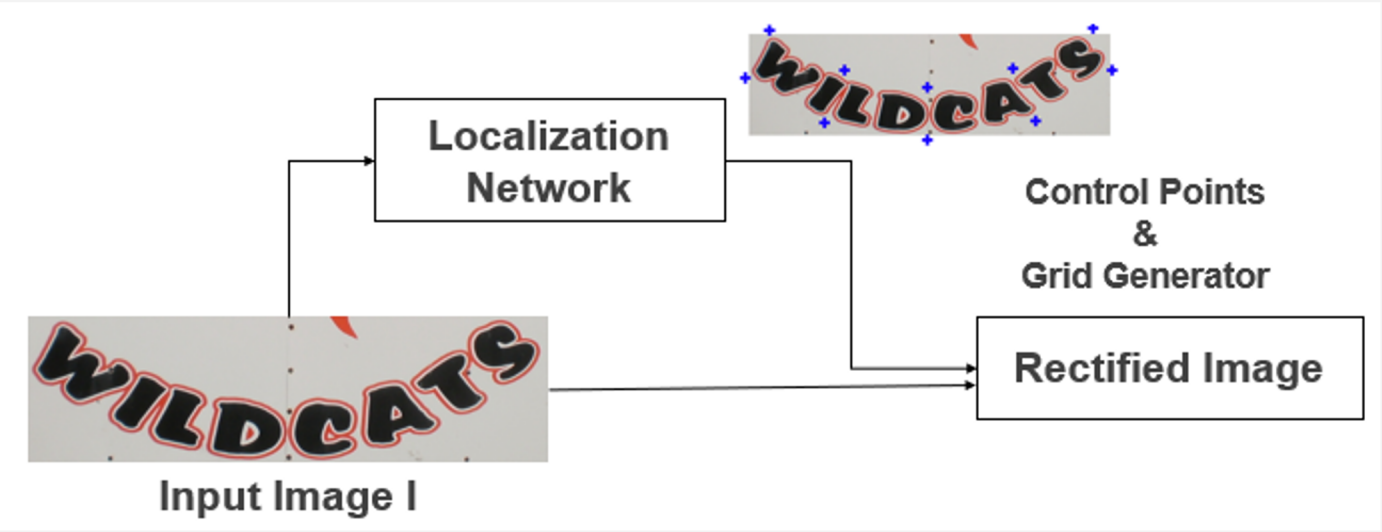
Architecture Space:

- **tree-structured architecture space**
- first, normal, reduction cell
  - MobileNet v2 based: for parameter-efficient
  - narrow down the search space; kernel size & expansion ratio

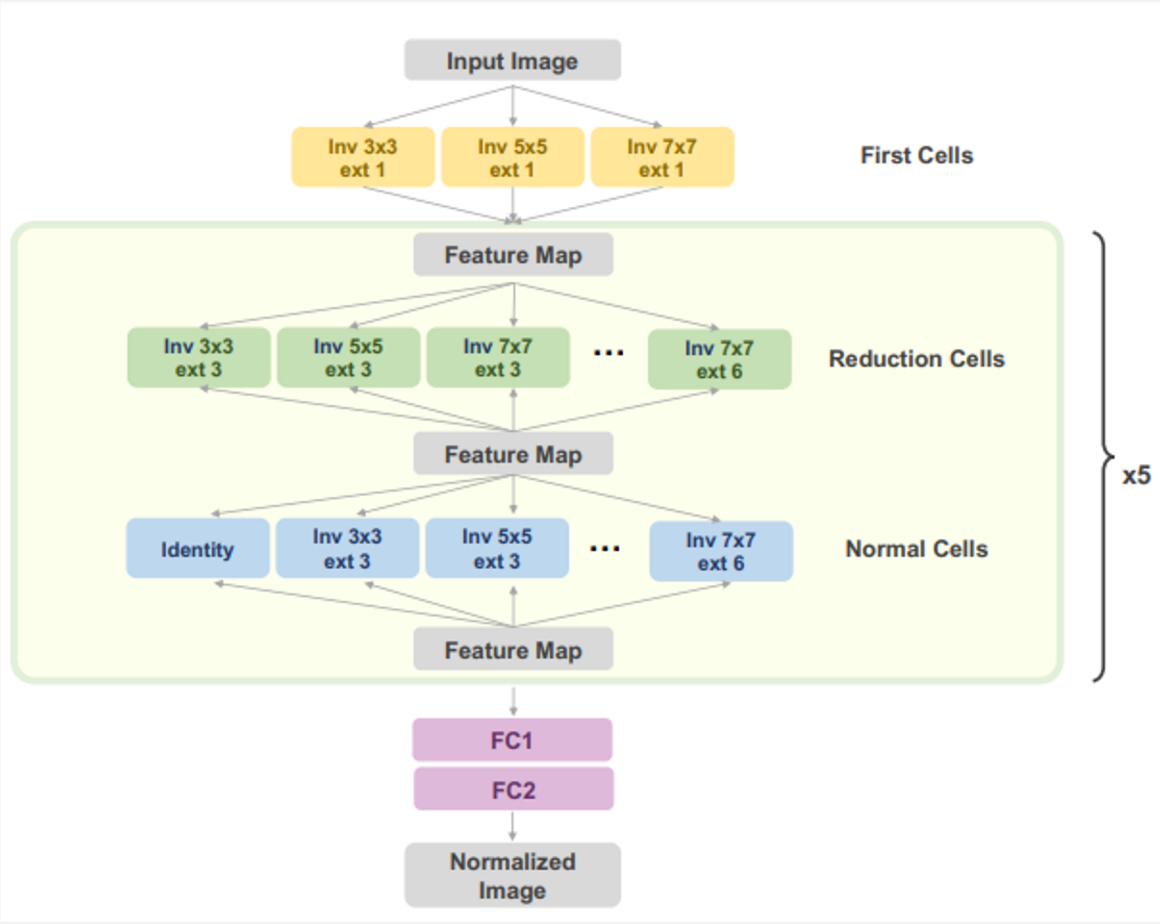


Rectifier Search:

- TPS-based approach

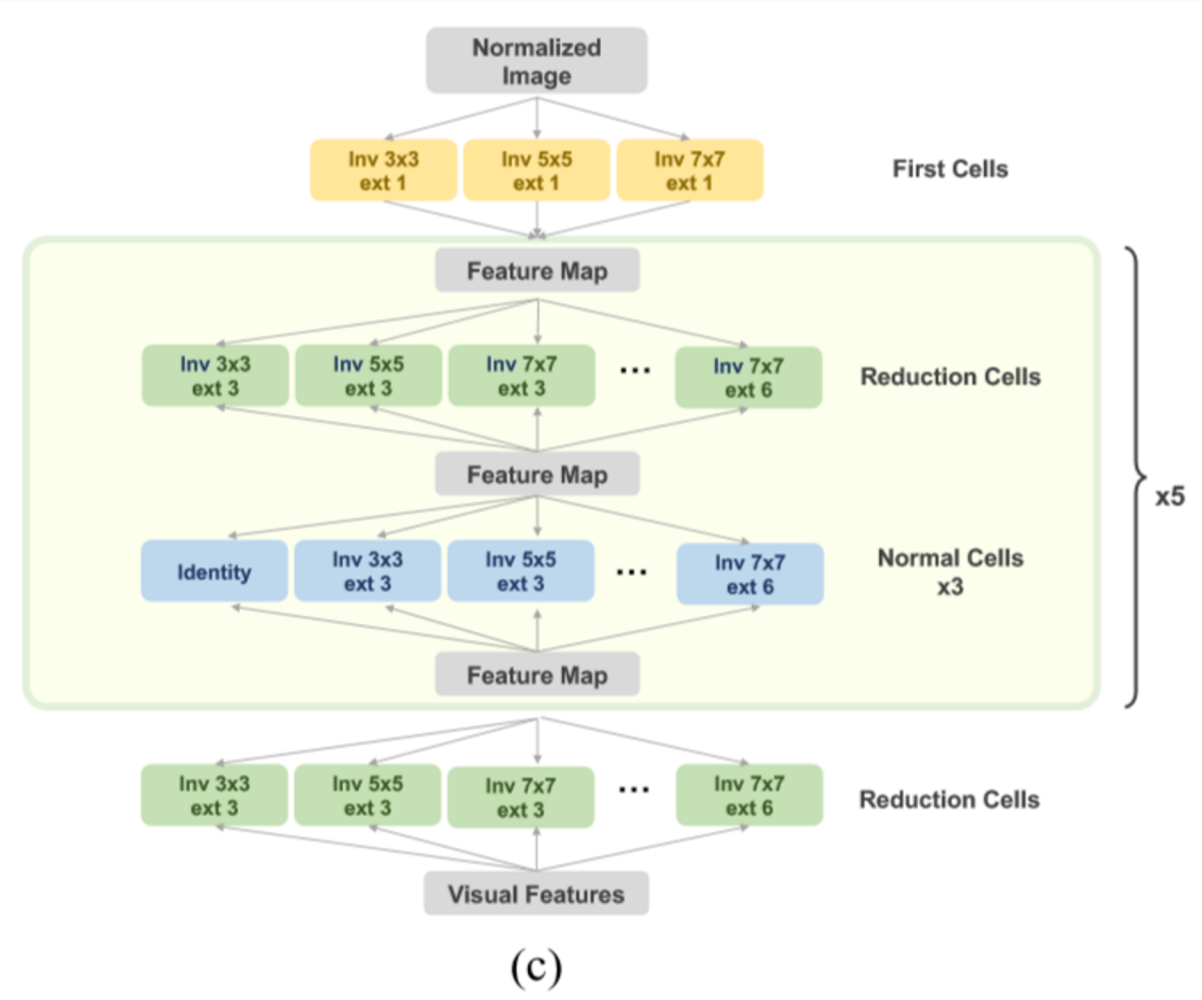


- Model Structure of Localization Network



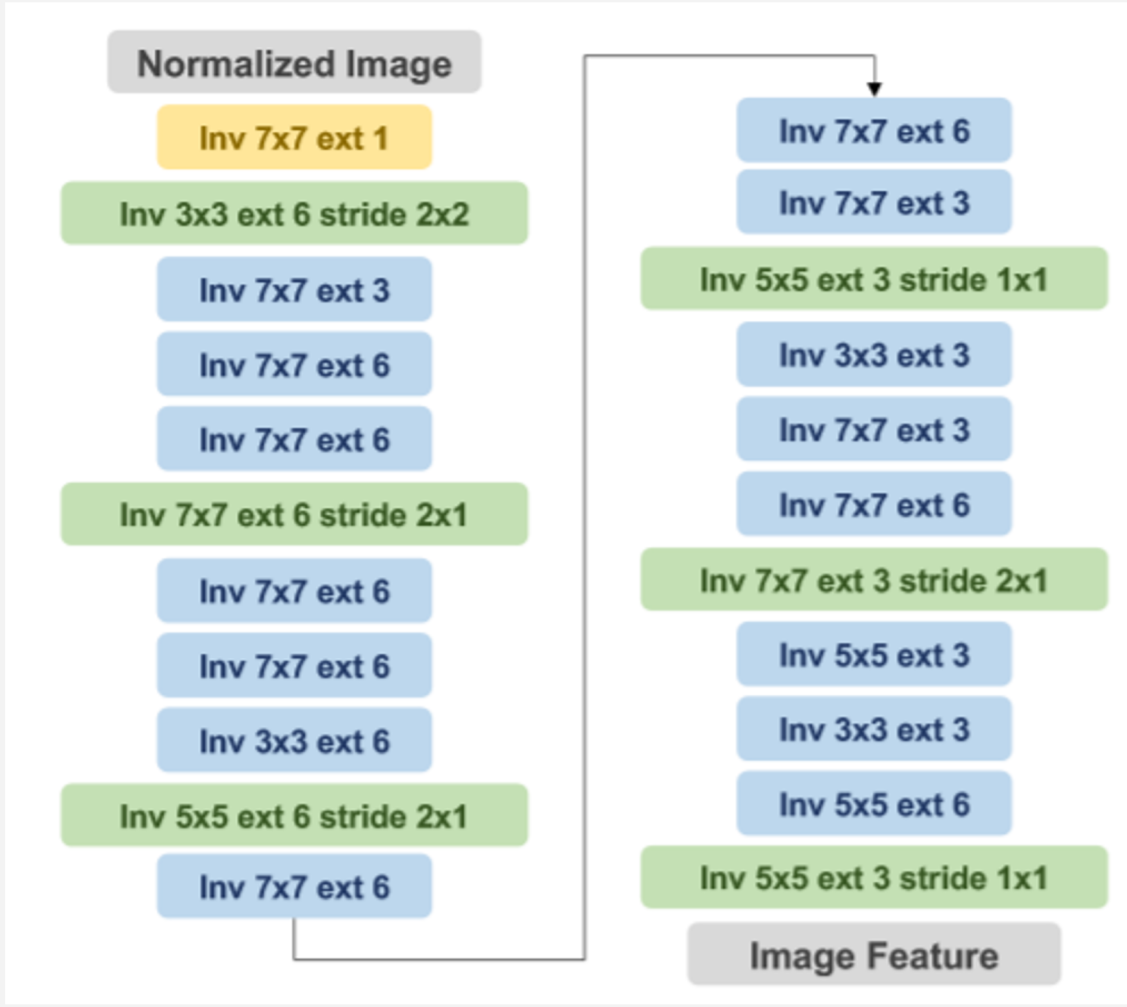
Feature Extractor Search:

- Model Structure of Feature Extractor



(c)

Our Searched



How to Train

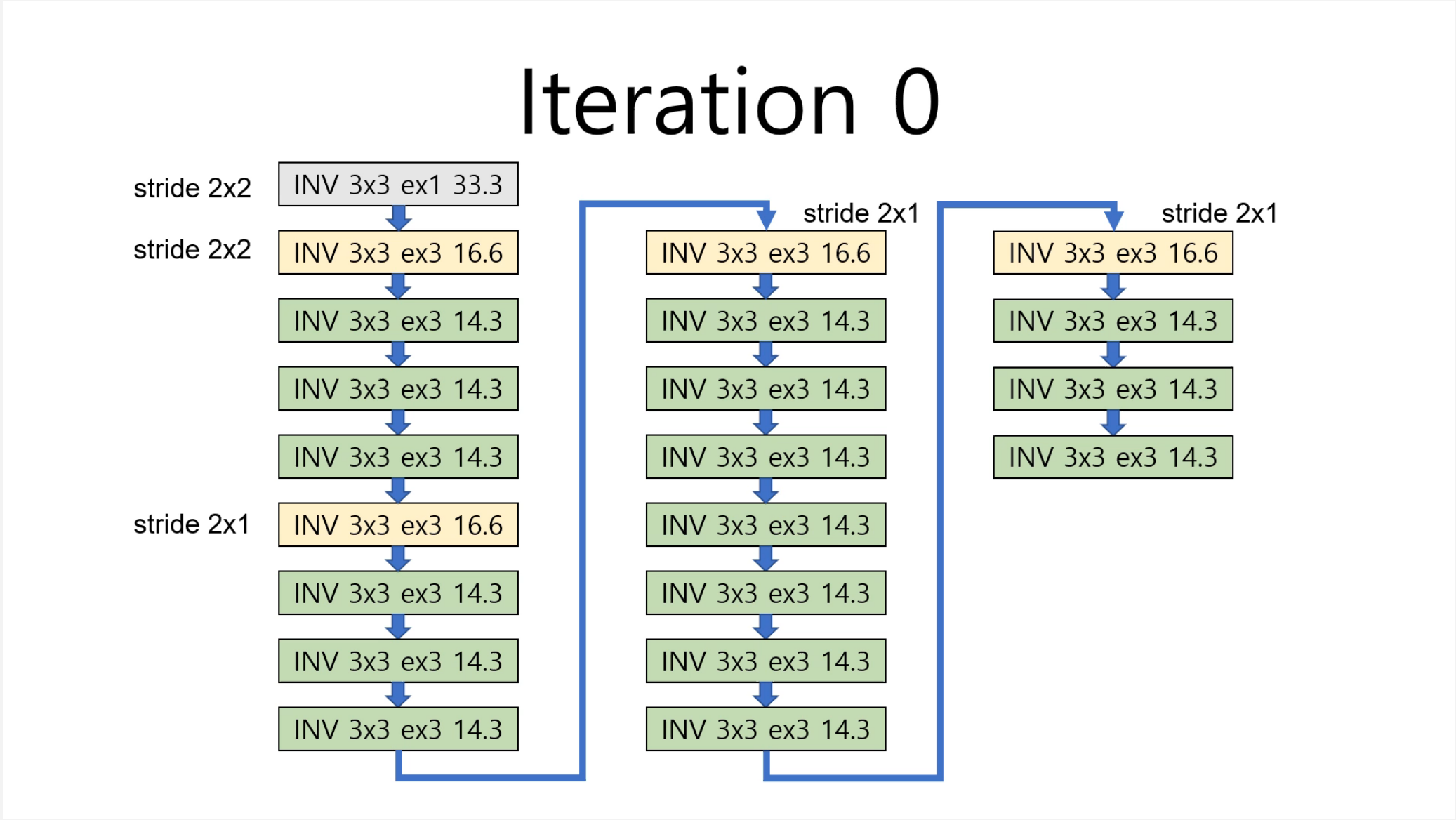
- **Respectively** update **architecture** parameters & **model** parameters
- Loss for updating architecture parameters
  - Cross Entropy Loss

$$\frac{\partial L}{\partial \alpha_i} \approx \sum_{j=1}^{N} \frac{\partial L}{\partial g_j} \frac{\partial \frac{exp(\alpha_j)}{\sum_k exp(\alpha_k)}}{\partial \alpha_i} = \sum_{j=1}^{N} \frac{\partial L}{\partial g_j} p_j(\delta_{ij} - p_i)$$

  - **Latency** : previously measured for every possible dimensions

$$L_{total} = L_{CE} + \lambda_1 * L_{latency}$$

DEMO


Iteration 0
stride 2x2
stride 2x2
stride 2x1
INV 3x3 ex1 33.3
INV 3x3 ex3 16.6
INV 3x3 ex3 14.3
INV 3x3 ex3 14.3
INV 3x3 ex3 14.3
INV 3x3 ex3 16.6
INV 3x3 ex3 14.3
INV 3x3 ex3 14.3
INV 3x3 ex3 14.3
stride 2x1
INV 3x3 ex3 16.6
INV 3x3 ex3 14.3
INV 3x3 ex3 14.3
INV 3x3 ex3 14.3
INV 3x3 ex3 14.3
INV 3x3 ex3 14.3
INV 3x3 ex3 14.3
stride 2x1
INV 3x3 ex3 16.6
INV 3x3 ex3 14.3
INV 3x3 ex3 14.3
INV 3x3 ex3 14.3

Experiment Results:

Dataset:

Train - MJSynth + SynthText

Test - IIIT, SVT, IC03, IC13, IC15, SVT Perspective, CUTE80

3 Different Settings:

1. same hyperparameters as ProxylessNAS settings & total loss
2. w/o latency
3. w/o latency + lower learning rates

Experiment Results

- Architecture Search: **Rectifier** vs **Feature Extractor**

First Cell

3x3	5x5	7x7
0.334	0.334	0.321

First Cell

3x3	5x5	7x7
0.033	0.95	0.016

Normal cell

identity	3x3 ex3	5x5 ex3	7x7 ex3	3x3 ex6	5x5 ex6	7x7 ex6
0.140	0.142	0.147	0.138	0.137	0.150	0.146
0.145	0.152	0.144	0.143	0.140	0.137	0.138
0.140	0.145	0.142	0.143	0.143	0.144	0.143
0.143	0.143	0.144	0.145	0.144	0.140	0.141
0.156	0.140	0.147	0.135	0.140	0.135	0.147

Normal cell

identity	3x3 ex3	5x5 ex3	7x7 ex3	3x3 ex6	5x5 ex6	7x7 ex6
0.009	0.009	0.036	0.379	0.011	0.022	0.534
0.011	0.014	0.015	0.414	0.019	0.018	0.51
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.082	0.154	0.19	0.16	0.104	0.099	0.211
0.09	0.152	0.199	0.097	0.11	0.201	0.15

Reduction Cell

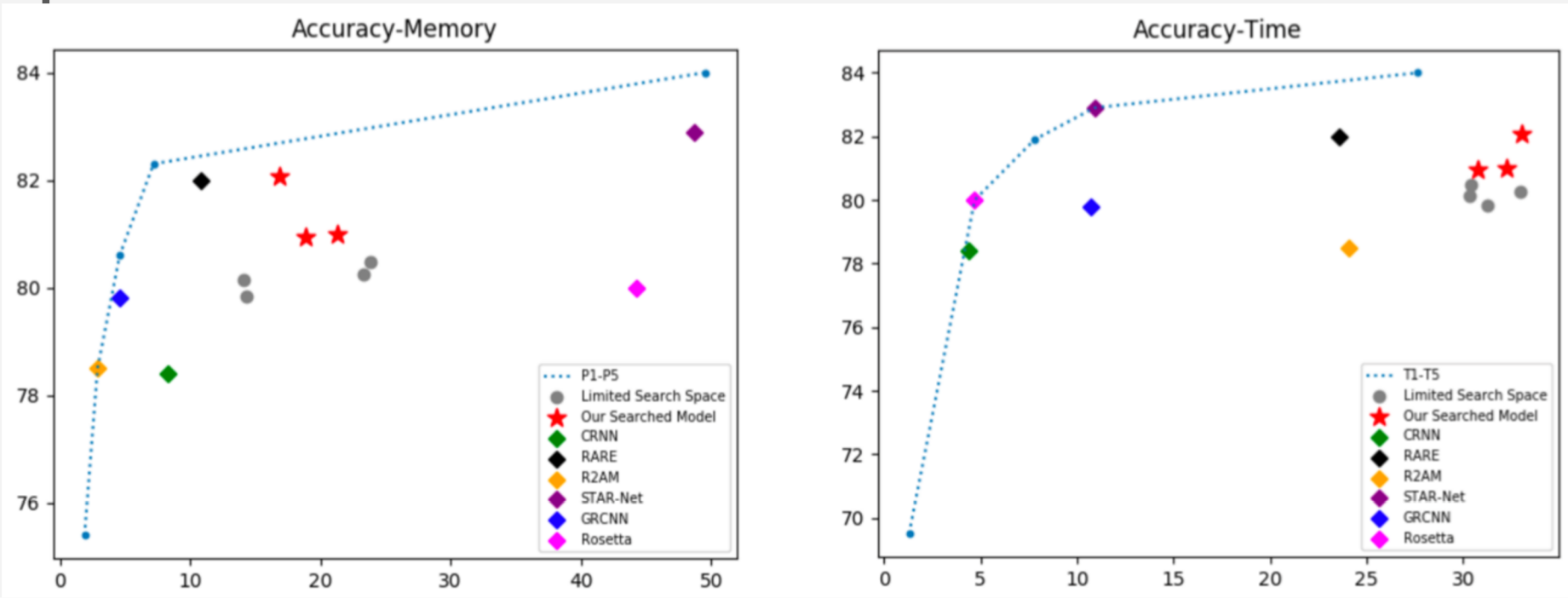
3x3 ex3	5x5 ex3	7x7 ex3	3x3 ex6	5x5 ex6	7x7 ex6
0.176	0.166	0.174	0.152	0.166	0.166
0.169	0.163	0.164	0.173	0.175	0.155
0.158	0.162	0.170	0.165	0.175	0.170
0.173	0.166	0.163	0.163	0.164	0.171
0.154	0.161	0.172	0.165	0.176	0.172

Reduction Cell

3x3 ex3	5x5 ex3	7x7 ex3	3x3 ex6	5x5 ex6	7x7 ex6
0.012	0.012	0.256	0.016	0.018	0.686
0.013	0.024	0.025	0.013	0.232	0.694
⋮	⋮	⋮	⋮	⋮	⋮
0.006	0.201	0.157	0.285	0.154	0.144
0.132	0.153	0.158	0.187	0.116	0.255

Experiment Results

- **Comparisons** with other SOTA models

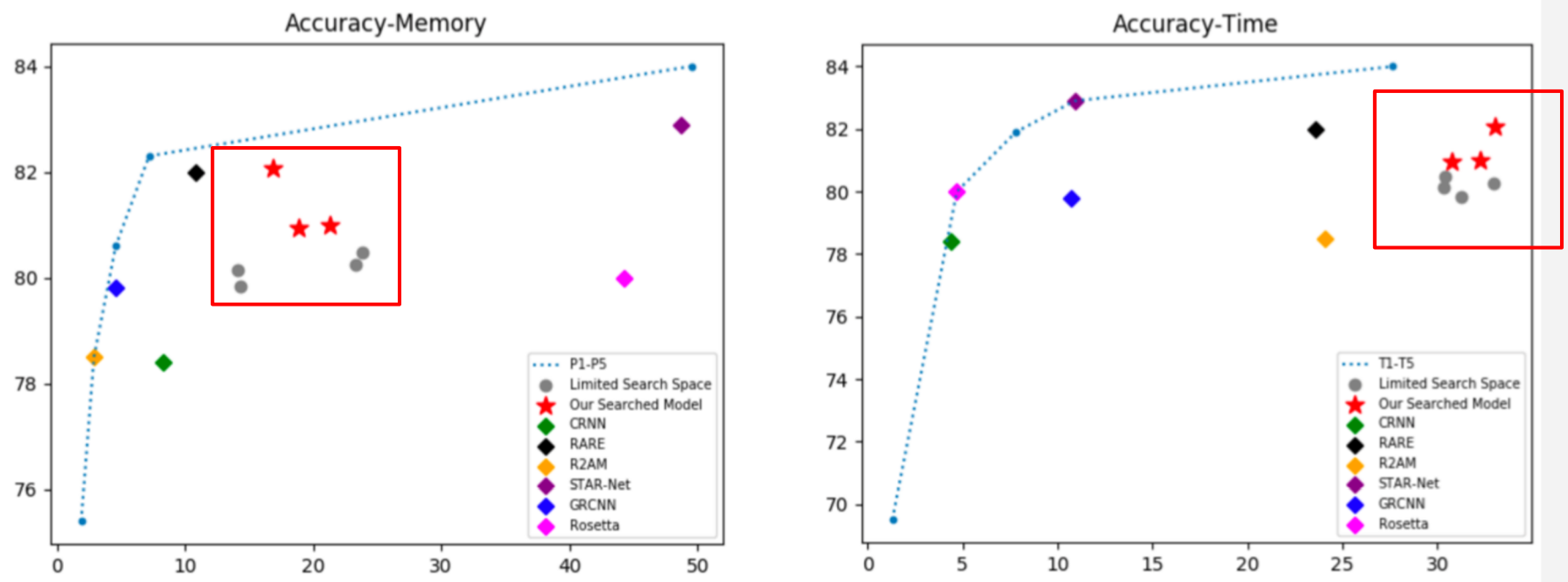


Model	IIIT 3000	SVT 647	IC03 860	IC03 867	IC13 857	IC13 1015	IC15 1811	IC15 2077	SVTP	CUTE80	Time ms/image	params $\times 10^6$
CRNN [29]	82.9	82.380	81.6	93.1	92.6	91.1	89.2	69.4	70.0	65.5	**4.4**	8.3
RARE [30]	86.2	85.8	93.9	93.7	92.6	91.1	74.5	68.9	76.2	70.4	23.6	10.8
R2AM [16]	83.4	82.4	92.2	92.0	90.2	88.1	68.9	63.6	72.1	64.9	24.1	2.9
STAR-Net [20]	**87.0**	**86.9**	**94.4**	**94.0**	**92.8**	**91.5**	**76.1**	**70.3**	**77.5**	71.7	10.9	48.7
GRCNN [33]	84.2	83.7	93.5	93.0	90.9	88.8	71.4	65.8	73.6	68.1	10.7	**4.6**
Rosetta [2]	84.3	84.7	93.4	92.9	90.9	89.0	71.2	66.0	73.8	69.2	4.7	44.3
STR-NAS3	85.7	85.9	93.5	93.1	91.6	90.5	75.6	69.9	76.9	**72.2**	33.0	16.8
Best combination [1]	87.9	87.5	94.9	94.4	93.6	92.3	77.6	71.8	79.2	74.0	27.6	49.6

Experiment Results

- **Verify** the effect of NAS

Model	IIIT 3000	SVT 647	IC03 860	IC03 867	IC13 857	IC13 1015	IC15 1811	IC15 2077	SVTP	CUTE80	Time ms/image	params $\times 10^6$
3*3 ex3	84.433	82.380	92.442	92.272	92.065	90.148	72.170	66.731	74.109	69.097	**30.361**	**14.101**
5*5 ex3	83.667	83.308	92.791	92.964	91.715	90.443	71.121	65.527	75.039	69.097	31.251	14.292
5*5 ex6	84.333	81.917	93.140	93.541	91.599	90.049	72.612	66.827	75.194	67.361	32.994	23.283
7*7 ex6	84.567	83.153	92.907	92.388	91.249	89.655	73.771	67.646	75.194	68.056	30.415	23.857
STR-NAS1	84.967	83.771	92.907	92.849	92.182	91.034	73.771	67.935	74.729	70.139	32.238	21.332
STR-NAS2	85.000	83.771	94.070	93.772	92.532	91.034	73.440	67.935	73.488	68.750	30.738	18.923
STR-NAS3	**85.667**	**85.935**	**93.488**	**93.080**	**91.599**	**90.542**	**75.594**	**69.860**	**76.899**	**72.222**	33.017	16.821



Conclusion:

- We are the pioneers of the **NAS on STR** with results suggesting positive potential development.

Future Work:

- We can apply NAS to other **pipeline of STR modules**,
  or design the improved algorithm to handle with sensitive localization network
- We can make improvements on **many other real-world applications** by applying NAS

THANK YOU!

Memory-Efficient Models for Scene Text Recognition via Neural Architecture Search

SeulGi Hong DongHyun Kim Min-Kook Choi

hutom

NASFW 2020

WACV 2020

March 1-5, Aspen, CO