# Supplementary material

Certainty-based Architecture Selection Framework for TinyML Devices

Joanna Komorniczak, Tobiasz Puślecki,
Paweł Ksieniewicz, Krzysztof Walkowiak

## 1 Data stream configuration

Table 1: The configuration of data streams for the desktop experiment

| Characteristics | Values |
|---|---|
| Origin dataset | MNIST, SVHN |
| Repeats | 5 |
| Chunk size | 50, 150, 300, 500 |
| Complexity cycles | 3, 5, 10, 25 |
| Domain change type | instant, linear, normal |

## 2 Neural Architecture Search

### Setup

The first experiment was intended to select a pool of architectures for the considered tasks. A total of 14 different network architectures were examined, from the simplest – containing only fully connected layers –to more complex – containing multiple convolutional layers. For the proposed framework, a specific set of considered architectures should be predefined, where the more time-demanding solution achieves a better classification quality. A two-criteria optimization task can be formulated, where one criterion is the inference time and the other is the recognition quality.

The experiment was carried out on static data, used for the training in the following experiments. Dataset was divided into 5 folds in stratified cross-validation. Architecture training was stopped at epoch 250 or when the MPS for the training set exceeded a given threshold. The average results of inference time and classification quality were analyzed. The aim of the experiment was to select five ordered architectures from the *Pareto Set* that meet the assumption that a model with higher accuracy will be more computationally expensive.

The operation of 14 neural network architectures was tested:

- *Fully Connected* (`FC`) – three architectures without convolutional layers with 2, 3 and 4 fully connected layers respectively.
- *One Convolutional Layer* (`CNN1`) – four architectures with one convolutional layer, with depths of 5, 10, 15 and 20 respectively.
- *Two Convolutional Layers* (`CNN2`) – five architectures with two convolutional layers with increasing filter depth.
- *Three Convolutional Layers* (`CNN3`) – two architectures with three convolutional layers and increasing filter depth.

## Results

Figure 1 shows the *Neural Architecture Search* (NAS) of the first experiment for the MNIST and SVHN datasets, respectively. The points indicate the average results for the considered architectures. Those that were finally selected for the pool used in the framework – belonging to the *Pareto set* – are marked in color. The problem considered in the article is two-criteria – a compromise is sought between the quality of classification and the inference time of the model. The average classification quality is marked on the horizontal axis, and the inference time for the full test set in seconds is marked on the vertical axis.

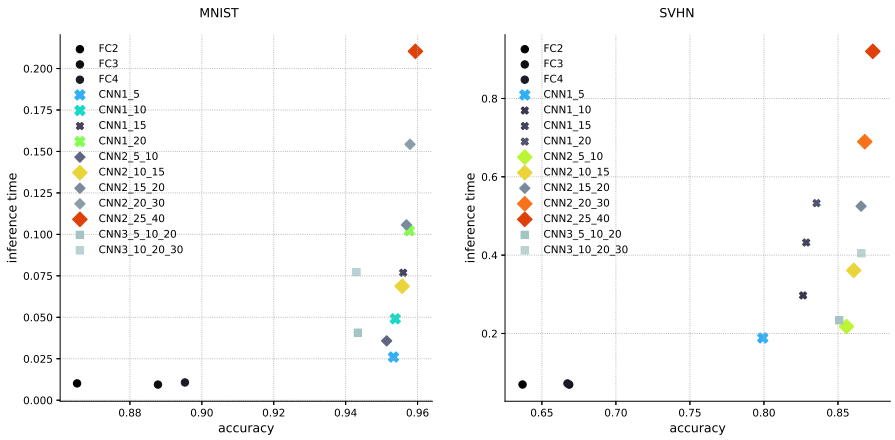


Figure 1: Neural architecture search results for MNIST and SVHN datasets

It can be easily noticed that the simplest architectures without convolutional layers (`FC`) are characterized by the lowest inference time, but achieve very low classification quality. Therefore, they were not taken into account when selecting architectures for further experiments.

Architectures with one convolutional layer have performed relatively well in the MNIST problem, achieving satisfactory recognition quality with low inference time. Architectures with two convolutional layers performed best in terms of

quality, but their inference time was the highest. Architectures with 3 layers did not perform well in this problem, which may be related to the relative simplicity of the task being solved. In the SVHN problem, `CNN1` architectures achieved significantly lower quality results than `CNN2` and `CNN3`. Architectures with two convolutional layers dominate the *Pareto set* for this dataset.

As a result of experiment, the following architectures were selected for MNIST-based streams: `CNN1_5`, `CNN1_10`, `CNN2_10_15`, `CNN1_20`, `CNN2_25_40` and following for SVHN-based streams: `CNN1_5`, `CNN2_5_10`, `CNN2_10_15`, `CNN2_20_30`, `CNN2_25_40`. The models are ordered from the simplest to the most complex.

# 3 Desktop experiment visualizations

## Classification accuracy

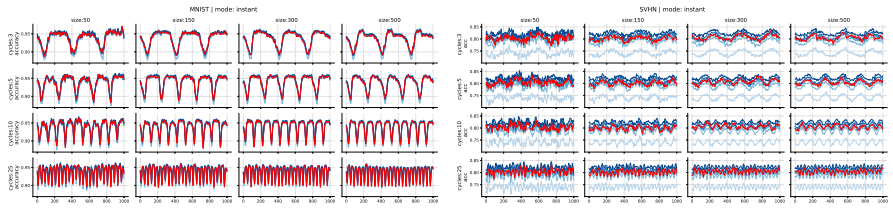


Figure 2: Accuracy of MNIST-based (left) and SVHN-based (right) stream with instant difficulty change

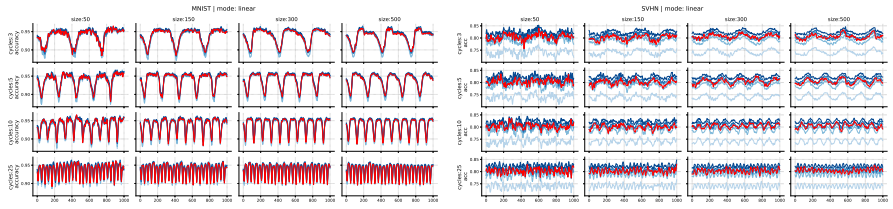


Figure 3: Accuracy of MNIST-based (left) and SVHN-based (right) stream with linear difficulty change

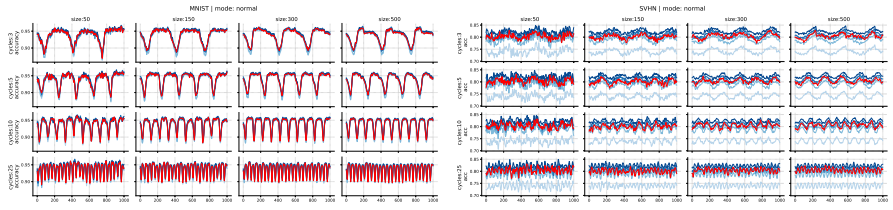

Figure 4: Accuracy of MNIST-based (left) and SVHN-based (right) stream with normal difficulty change

## Selected architecture

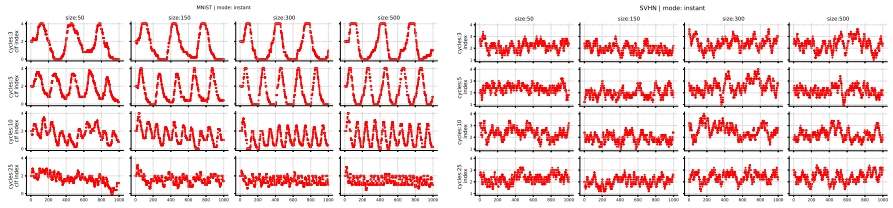

Figure 5: Selected architecture of MNIST-based (left) and SVHN-based (right) stream with instant difficulty change

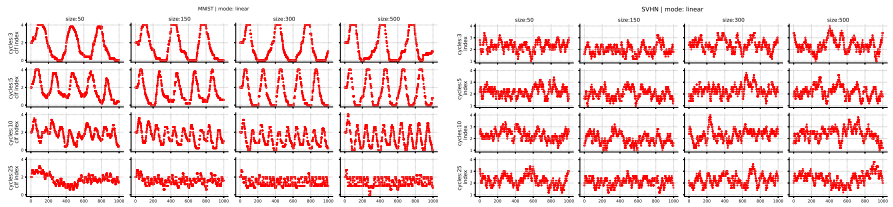

Figure 6: Selected architecture of MNIST-based (left) and SVHN-based (right) stream with linear difficulty change

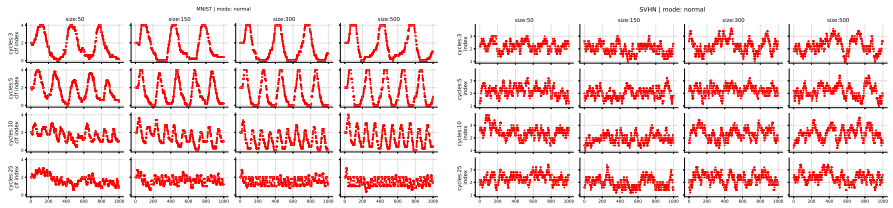

Figure 7: Selected architecture of MNIST-based (left) and SVHN-based (right) stream with normal difficulty change

## Number of switches between architectures

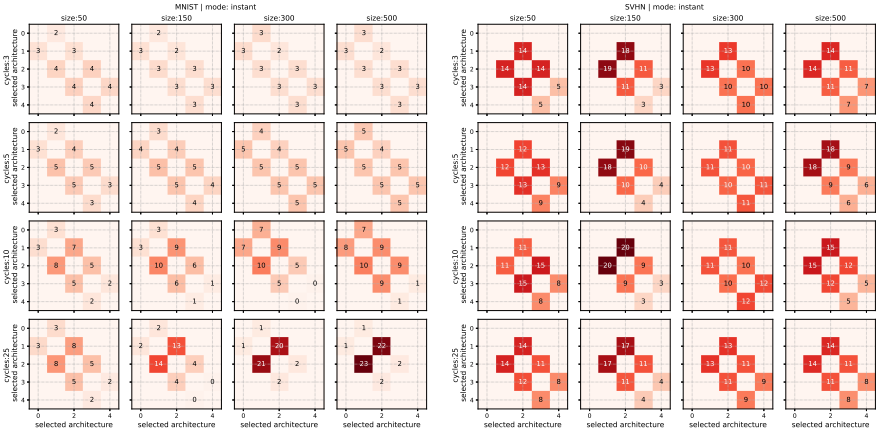


Figure 8: Number of architecture switches of MNIST-based (left) and SVHN-based (right) stream with instant difficulty change

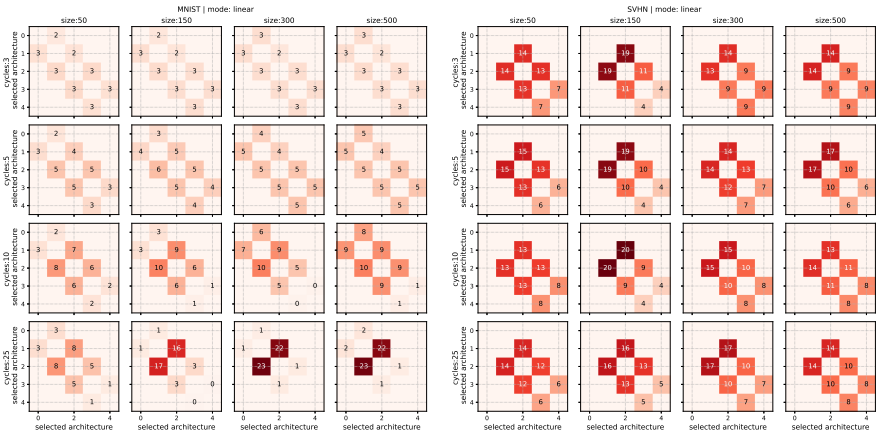


Figure 9: Number of architecture switches of MNIST-based (left) and SVHN-based (right) stream with linear difficulty change

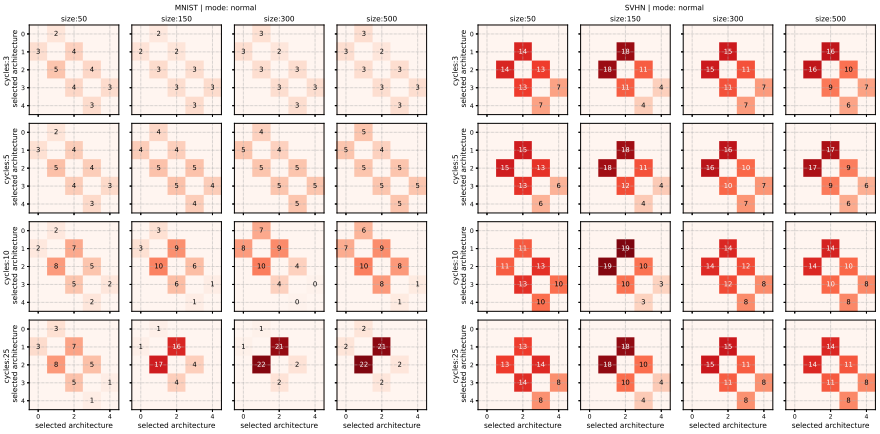


Figure 10: Number of architecture switches of MNIST-based (left) and SVHN-based (right) stream with normal difficulty change

# 4 Tables

## MNIST

| stream | R acc | CAS acc | CAS Latency (*1e3) | CAS MACC(*1e9) | TTAG |
|---|---|---|---|---|---|
| instant cs50 c3 | 0.940 | 0.938 (-0.002) | 1.545 (-2.637) | 0.029 (-0.071) | 180.281 |
| instant cs50 c5 | 0.940 | 0.938 (-0.002) | 1.394 (-2.788) | 0.024 (-0.075) | 196.680 |
| instant cs50 c10 | 0.940 | 0.938 (-0.003) | 1.120 (-3.062) | 0.017 (-0.083) | 210.163 |
| instant cs50 c25 | 0.940 | 0.936 (-0.004) | 1.140 (-3.042) | 0.017 (-0.082) | 152.951 |
| instant cs150 c3 | 0.940 | 0.938 (-0.002) | 4.512 (-8.035) | 0.084 (-0.215) | 224.077 |
| instant cs150 c5 | 0.943 | 0.940 (-0.002) | 3.472 (-9.075) | 0.055 (-0.243) | 241.693 |
| instant cs150 c10 | 0.941 | 0.938 (-0.003) | 3.106 (-9.441) | 0.044 (-0.255) | 215.492 |
| instant cs150 c25 | 0.939 | 0.936 (-0.004) | 3.141 (-9.406) | 0.044 (-0.254) | 170.423 |
| instant cs300 c3 | 0.940 | 0.938 (-0.002) | 8.578 (-16.516) | 0.157 (-0.439) | 187.806 |
| instant cs300 c5 | 0.942 | 0.940 (-0.002) | 7.862 (-17.232) | 0.137 (-0.460) | 225.327 |
| instant cs300 c10 | 0.942 | 0.939 (-0.003) | 5.407 (-19.687) | 0.069 (-0.527) | 200.847 |
| instant cs300 c25 | 0.939 | 0.935 (-0.004) | 5.945 (-19.149) | 0.081 (-0.515) | 162.393 |
| instant cs500 c3 | 0.938 | 0.936 (-0.002) | 15.184 (-26.639) | 0.285 (-0.709) | 175.249 |
| instant cs500 c5 | 0.940 | 0.938 (-0.002) | 14.144 (-27.679) | 0.255 (-0.739) | 214.620 |
| instant cs500 c10 | 0.940 | 0.937 (-0.003) | 9.842 (-31.981) | 0.133 (-0.861) | 213.547 |
| instant cs500 c25 | 0.938 | 0.934 (-0.004) | 9.884 (-31.939) | 0.134 (-0.860) | 152.603 |
| linear cs50 c3 | 0.940 | 0.938 (-0.002) | 1.497 (-2.685) | 0.027 (-0.072) | 203.507 |
| linear cs50 c5 | 0.940 | 0.938 (-0.002) | 1.470 (-2.713) | 0.026 (-0.073) | 189.081 |
| linear cs50 c10 | 0.940 | 0.938 (-0.002) | 1.241 (-2.941) | 0.020 (-0.080) | 258.983 |
| linear cs50 c25 | 0.939 | 0.936 (-0.003) | 1.164 (-3.018) | 0.018 (-0.082) | 183.636 |
| linear cs150 c3 | 0.940 | 0.938 (-0.002) | 4.481 (-8.066) | 0.083 (-0.215) | 220.977 |
| linear cs150 c5 | 0.943 | 0.941 (-0.002) | 3.639 (-8.908) | 0.059 (-0.239) | 255.219 |
| linear cs150 c10 | 0.941 | 0.938 (-0.003) | 3.074 (-9.473) | 0.042 (-0.256) | 234.464 |
| linear cs150 c25 | 0.939 | 0.936 (-0.003) | 3.120 (-9.427) | 0.044 (-0.255) | 174.711 |
| linear cs300 c3 | 0.940 | 0.938 (-0.002) | 8.467 (-16.627) | 0.154 (-0.443) | 200.468 |
| linear cs300 c5 | 0.942 | 0.940 (-0.002) | 7.963 (-17.131) | 0.139 (-0.457) | 234.060 |
| linear cs300 c10 | 0.942 | 0.940 (-0.003) | 5.441 (-19.653) | 0.070 (-0.527) | 221.728 |
| linear cs300 c25 | 0.939 | 0.936 (-0.004) | 5.901 (-19.193) | 0.080 (-0.516) | 163.837 |
| linear cs500 c3 | 0.939 | 0.936 (-0.002) | 15.007 (-26.816) | 0.280 (-0.714) | 193.112 |
| linear cs500 c5 | 0.940 | 0.938 (-0.002) | 14.264 (-27.559) | 0.258 (-0.736) | 199.600 |
| linear cs500 c10 | 0.940 | 0.937 (-0.003) | 9.567 (-32.256) | 0.127 (-0.867) | 211.779 |
| linear cs500 c25 | 0.938 | 0.934 (-0.004) | 9.821 (-32.002) | 0.133 (-0.861) | 152.490 |
| normal cs50 c3 | 0.940 | 0.938 (-0.002) | 1.515 (-2.667) | 0.028 (-0.072) | 197.426 |
| normal cs50 c5 | 0.940 | 0.937 (-0.003) | 1.362 (-2.820) | 0.023 (-0.076) | 188.023 |
| normal cs50 c10 | 0.940 | 0.937 (-0.003) | 1.158 (-3.024) | 0.018 (-0.082) | 207.315 |
| normal cs50 c25 | 0.940 | 0.936 (-0.003) | 1.116 (-3.067) | 0.016 (-0.083) | 166.036 |
| normal cs150 c3 | 0.940 | 0.938 (-0.002) | 4.498 (-8.048) | 0.083 (-0.215) | 221.754 |
| normal cs150 c5 | 0.943 | 0.941 (-0.002) | 3.727 (-8.820) | 0.062 (-0.236) | 247.973 |
| normal cs150 c10 | 0.941 | 0.938 (-0.003) | 3.069 (-9.478) | 0.043 (-0.256) | 228.711 |
| normal cs150 c25 | 0.939 | 0.936 (-0.003) | 3.171 (-9.376) | 0.045 (-0.254) | 185.673 |
| normal cs300 c3 | 0.940 | 0.938 (-0.002) | 8.496 (-16.598) | 0.155 (-0.442) | 189.809 |
| normal cs300 c5 | 0.942 | 0.940 (-0.002) | 7.811 (-17.283) | 0.135 (-0.461) | 222.045 |
| normal cs300 c10 | 0.943 | 0.939 (-0.003) | 5.312 (-19.781) | 0.067 (-0.529) | 198.701 |
| normal cs300 c25 | 0.939 | 0.935 (-0.004) | 6.120 (-18.974) | 0.085 (-0.512) | 156.508 |
| normal cs500 c3 | 0.938 | 0.936 (-0.002) | 15.110 (-26.713) | 0.282 (-0.712) | 197.701 |
| normal cs500 c5 | 0.940 | 0.938 (-0.002) | 14.253 (-27.570) | 0.258 (-0.736) | 204.746 |
| normal cs500 c10 | 0.940 | 0.937 (-0.003) | 9.832 (-31.991) | 0.133 (-0.861) | 222.464 |
| normal cs500 c25 | 0.938 | 0.934 (-0.004) | 9.846 (-31.977) | 0.134 (-0.860) | 158.669 |

Table 2: Results for all streams generated based on MNIST

## SVHN

| stream | R acc | CAS acc | CAS Latency(*1e3) | CAS MACC(*1e9) | TTAG |
|---|---|---|---|---|---|
| instant cs50 c3 | 0.823 | 0.803 (-0.020) | 3.231 (-4.093) | 0.076 (-0.125) | 14.144 |
| instant cs50 c5 | 0.823 | 0.804 (-0.019) | 3.661 (-3.663) | 0.089 (-0.111) | 12.021 |
| instant cs50 c10 | 0.823 | 0.804 (-0.019) | 3.641 (-3.683) | 0.089 (-0.112) | 12.131 |
| instant cs50 c25 | 0.822 | 0.803 (-0.019) | 3.507 (-3.817) | 0.084 (-0.116) | 13.320 |
| instant cs150 c3 | 0.824 | 0.801 (-0.023) | 8.640 (-13.331) | 0.194 (-0.407) | 14.905 |
| instant cs150 c5 | 0.823 | 0.800 (-0.023) | 8.683 (-13.288) | 0.196 (-0.406) | 14.856 |
| instant cs150 c10 | 0.823 | 0.799 (-0.024) | 8.307 (-13.664) | 0.183 (-0.418) | 14.771 |
| instant cs150 c25 | 0.823 | 0.801 (-0.022) | 8.980 (-12.991) | 0.205 (-0.397) | 14.566 |
| instant cs300 c3 | 0.825 | 0.807 (-0.018) | 23.199 (-20.743) | 0.575 (-0.629) | 11.406 |
| instant cs300 c5 | 0.824 | 0.807 (-0.016) | 24.531 (-19.412) | 0.618 (-0.586) | 10.769 |
| instant cs300 c10 | 0.823 | 0.807 (-0.016) | 25.020 (-18.922) | 0.632 (-0.571) | 10.429 |
| instant cs300 c25 | 0.822 | 0.805 (-0.018) | 23.007 (-20.935) | 0.569 (-0.634) | 11.748 |
| instant cs500 c3 | 0.824 | 0.805 (-0.019) | 36.474 (-36.764) | 0.890 (-1.116) | 12.102 |
| instant cs500 c5 | 0.825 | 0.804 (-0.020) | 32.798 (-40.440) | 0.772 (-1.234) | 13.785 |
| instant cs500 c10 | 0.824 | 0.804 (-0.019) | 34.290 (-38.947) | 0.821 (-1.185) | 13.353 |
| instant cs500 c25 | 0.823 | 0.804 (-0.019) | 37.349 (-35.889) | 0.917 (-1.088) | 11.662 |
| linear cs50 c3 | 0.824 | 0.804 (-0.019) | 3.490 (-3.833) | 0.084 (-0.117) | 12.971 |
| linear cs50 c5 | 0.823 | 0.803 (-0.019) | 3.296 (-4.028) | 0.078 (-0.123) | 14.201 |
| linear cs50 c10 | 0.822 | 0.803 (-0.019) | 3.519 (-3.805) | 0.085 (-0.116) | 13.038 |
| linear cs50 c25 | 0.823 | 0.802 (-0.020) | 3.360 (-3.963) | 0.080 (-0.121) | 13.216 |
| linear cs150 c3 | 0.823 | 0.800 (-0.023) | 8.690 (-13.282) | 0.196 (-0.406) | 14.568 |
| linear cs150 c5 | 0.823 | 0.801 (-0.022) | 8.898 (-13.074) | 0.202 (-0.399) | 14.543 |
| linear cs150 c10 | 0.823 | 0.800 (-0.023) | 8.492 (-13.479) | 0.189 (-0.412) | 14.943 |
| linear cs150 c25 | 0.823 | 0.802 (-0.021) | 9.377 (-12.594) | 0.218 (-0.384) | 14.260 |
| linear cs300 c3 | 0.825 | 0.807 (-0.018) | 22.755 (-21.187) | 0.560 (-0.643) | 11.782 |
| linear cs300 c5 | 0.823 | 0.805 (-0.019) | 21.638 (-22.305) | 0.526 (-0.677) | 12.517 |
| linear cs300 c10 | 0.823 | 0.804 (-0.019) | 21.705 (-22.238) | 0.528 (-0.676) | 12.555 |
| linear cs300 c25 | 0.823 | 0.802 (-0.020) | 20.938 (-23.005) | 0.503 (-0.700) | 12.487 |
| linear cs500 c3 | 0.824 | 0.805 (-0.019) | 36.955 (-36.283) | 0.904 (-1.102) | 11.949 |
| linear cs500 c5 | 0.825 | 0.804 (-0.020) | 33.569 (-39.669) | 0.797 (-1.209) | 13.197 |
| linear cs500 c10 | 0.824 | 0.805 (-0.018) | 37.794 (-35.444) | 0.932 (-1.074) | 11.684 |
| linear cs500 c25 | 0.823 | 0.805 (-0.018) | 37.719 (-35.518) | 0.928 (-1.078) | 11.926 |
| normal cs50 c3 | 0.822 | 0.803 (-0.019) | 3.399 (-3.925) | 0.081 (-0.119) | 13.601 |
| normal cs50 c5 | 0.823 | 0.804 (-0.019) | 3.313 (-4.010) | 0.078 (-0.122) | 14.349 |
| normal cs50 c10 | 0.822 | 0.804 (-0.018) | 3.820 (-3.504) | 0.095 (-0.106) | 11.503 |
| normal cs50 c25 | 0.823 | 0.804 (-0.018) | 3.574 (-3.749) | 0.087 (-0.114) | 13.056 |
| normal cs150 c3 | 0.823 | 0.801 (-0.023) | 8.671 (-13.301) | 0.195 (-0.407) | 14.922 |
| normal cs150 c5 | 0.823 | 0.800 (-0.023) | 8.894 (-13.078) | 0.203 (-0.399) | 14.096 |
| normal cs150 c10 | 0.823 | 0.800 (-0.023) | 8.551 (-13.421) | 0.191 (-0.410) | 14.877 |
| normal cs150 c25 | 0.823 | 0.800 (-0.023) | 8.674 (-13.297) | 0.195 (-0.407) | 14.731 |
| normal cs300 c3 | 0.824 | 0.805 (-0.020) | 21.048 (-22.895) | 0.507 (-0.696) | 12.563 |
| normal cs300 c5 | 0.823 | 0.804 (-0.019) | 21.111 (-22.832) | 0.509 (-0.694) | 12.738 |
| normal cs300 c10 | 0.823 | 0.804 (-0.019) | 22.105 (-21.837) | 0.541 (-0.663) | 12.087 |
| normal cs300 c25 | 0.823 | 0.804 (-0.019) | 21.738 (-22.204) | 0.529 (-0.675) | 12.154 |
| normal cs500 c3 | 0.824 | 0.804 (-0.020) | 34.619 (-38.618) | 0.829 (-1.176) | 12.876 |
| normal cs500 c5 | 0.825 | 0.804 (-0.020) | 33.074 (-40.163) | 0.781 (-1.225) | 13.477 |
| normal cs500 c10 | 0.824 | 0.805 (-0.018) | 36.878 (-36.359) | 0.902 (-1.103) | 12.198 |
| normal cs500 c25 | 0.823 | 0.805 (-0.018) | 37.391 (-35.847) | 0.919 (-1.087) | 12.086 |

Table 3: Results for all streams generated based on SVHN