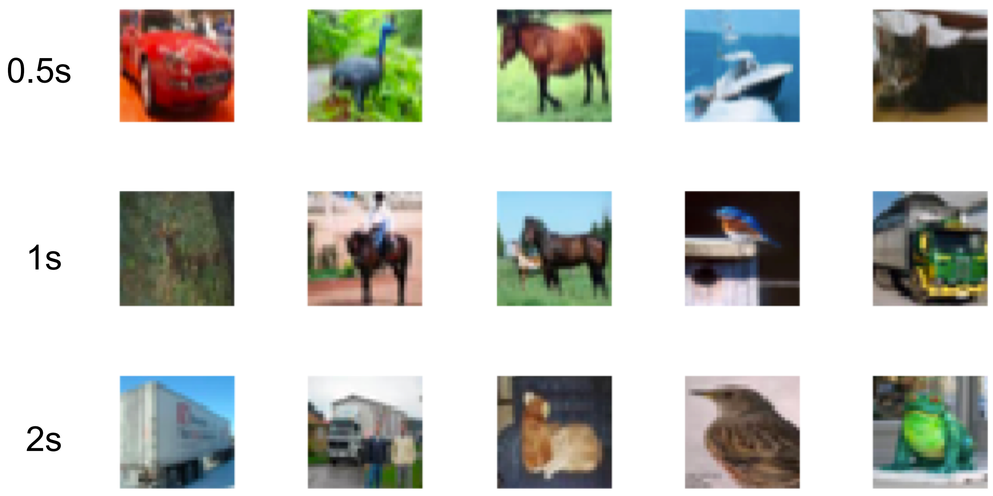
# Convolutions for Multiscale Dense Networks for Efficient Anytime Image Classification

Sven Peter[1], Nasim Rahaman[1], Ferran Diego[2], Fred A. Hamprecht[1]

[1]HCI, University Heidelberg, Germany [2]Telefonica I+D, Barcelona, Spain

## OBJECT RECOGNITION UNDER TEST-TIME CONSTRAINTS

What are the objects in the pictures below?




## AIM

- Construct a classifier that is both accurate and cheap to evaluate
- but being capable of using extra resources during prediction time, when available, to improve the performance
- Observation: Some instances can be accurately classified by computing a single cheap feature. Other instances require many / more expensives one.

## MOTIVATION

A test image:

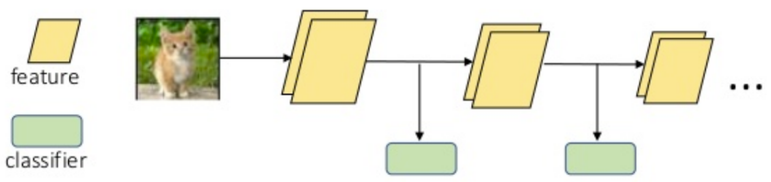
- **easy** vs **hard**
- **shallow model** vs **deep model**

Figure 1: Two images containing a *horse*. The left image is canonical and easy to detect even with a small model, whereas the right image requires a much more expensive network architecture. (Copyright Pixel Addict and Doyle (CC BY-ND 2.0).)

Training multiple classifiers with varying resource demands, which we adaptively apply during test time

**Develop CNNs:**

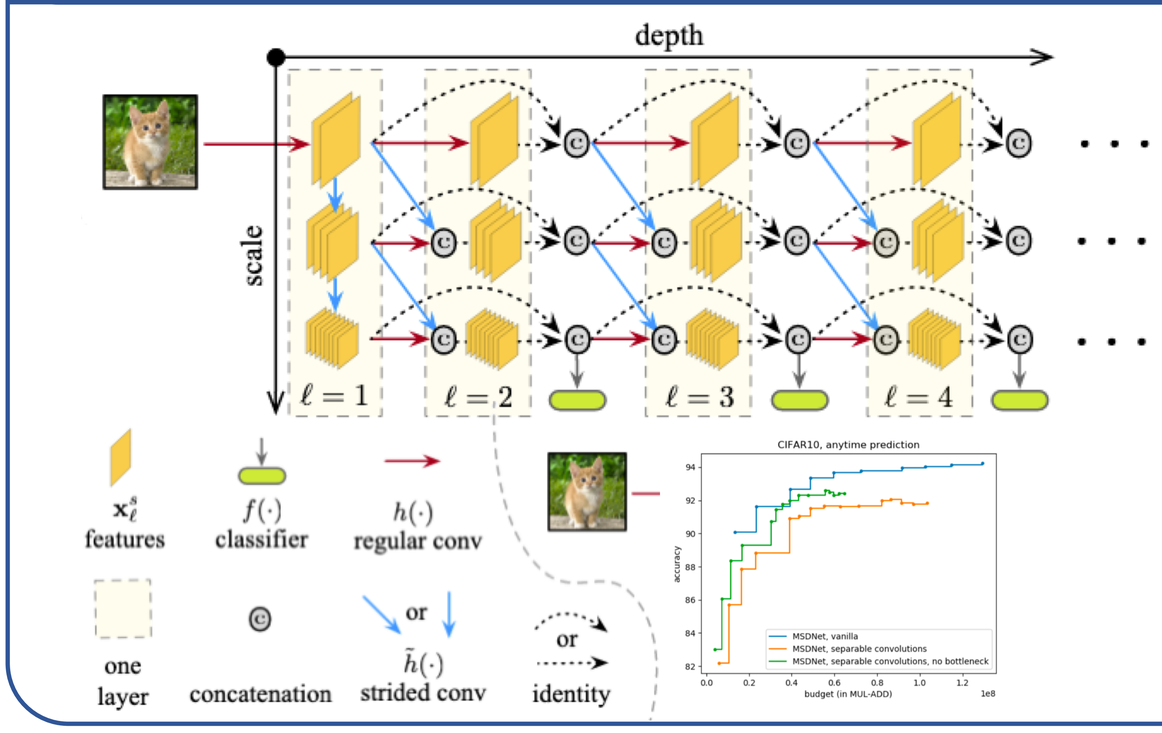
- "slice" the computation and process these slices one by one, stoppoing evaluation once the CPU time is depleted or the classification sufficiently certrain (through "early exits")




**Problems:**

- The lack of coarse-level features of early-exit classifiers
- Early classifiers interfere with later classifier

## PREVIOUS WORK: MSDNeT




## OBJECT RECOGNITION UNDER TEST-TIME CONSTRAINTS

Our architecture contribution is therefore required to overcome this limitation

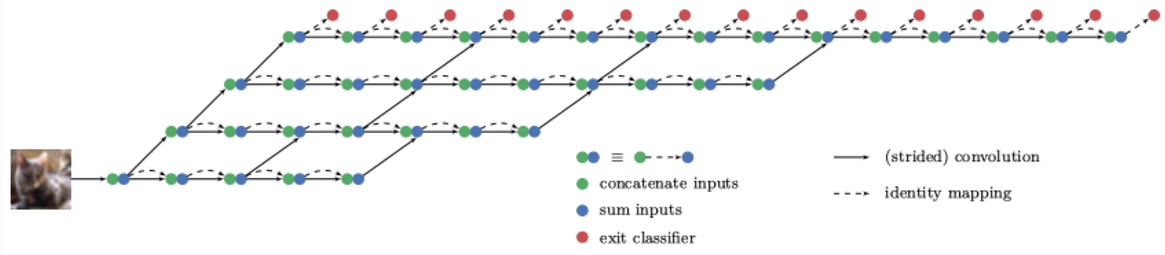
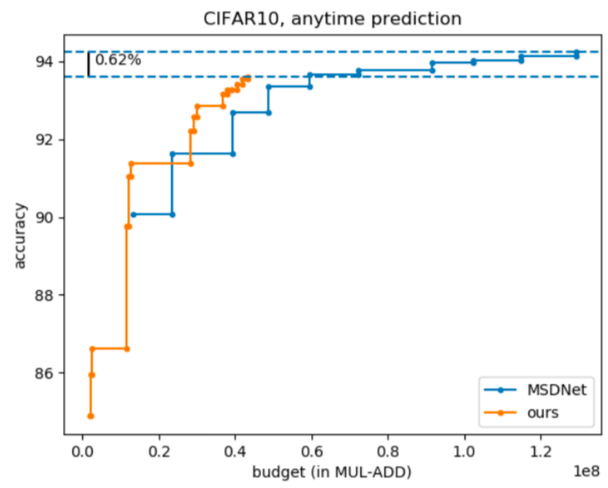
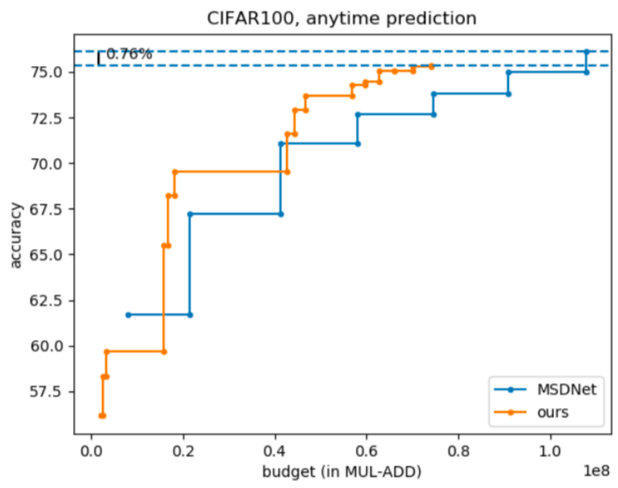



Fig. 1: **Proposed architecture for fast anytime prediction with separable convolutions**: Each row corresponds to a different scale at which the feature maps are computed (in this case 32x32, 16x16, 8x8, and 4x4 from bottom to top). Each edge represents a convolution operation or an identity mapping. Each node represents a summation or a concatenation of all inputs along the channel axis. Connections across scales are aggregated using summation while connections within the same scale are concatenated. The input image in this figure is taken from the cat class of the CIFAR10 dataset [7].

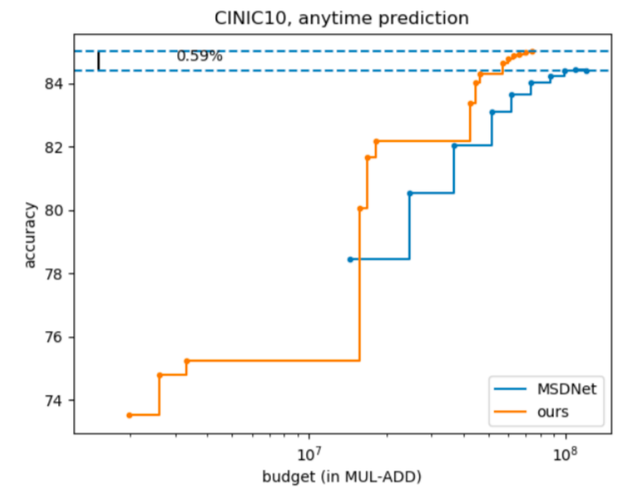
## RESULTS




(a) CIFAR10




(b) CIFAR100




(c) CINIC10

## CONCLUSIONS

- New neural network architecture for fast anytime prediction
- Dense connectivity and simultaneous representations across multiple scales could be adapted to allow using depthwise and spatially separable convolutions
- Our CNN architecture is useful for low budget settings where first predictions have to be available as soon as possible.