Dense Prediction on Sequences with Time-Dilated Convolutions for Speech Recognition

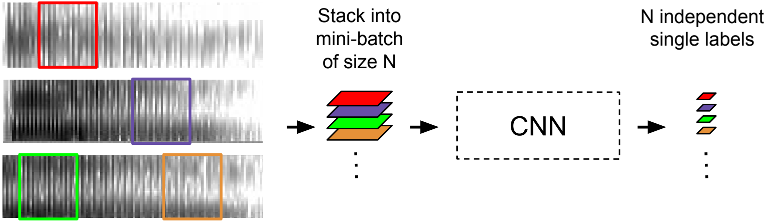
Tom Sercu, Vaibhava Goel

NIPS 2016
End-to-end Learning for Speech and Audio Processing Workshop

http://arxiv.org/abs/1611.09288

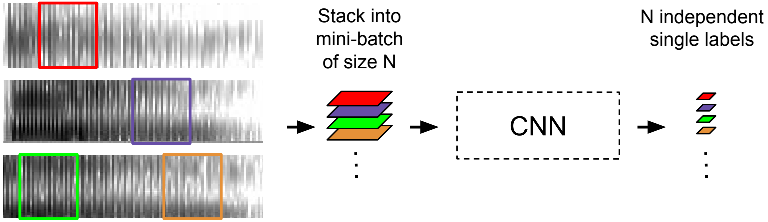
Convolutional Neural Networks for LVCSR

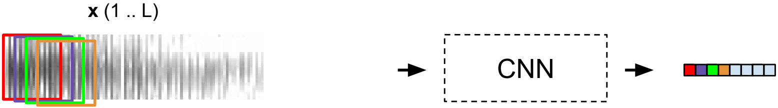
2-stage training scheme: Cross-Entropy (XE) vs Sequence Training (ST), test time



Convolutional Neural Networks for LVCSR

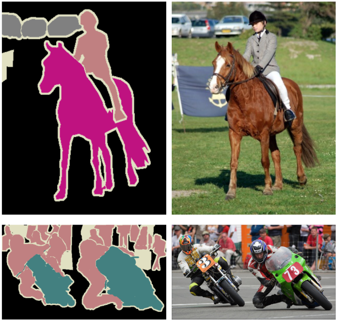
2-stage training scheme: Cross-Entropy (XE) vs Sequence Training (ST), test time





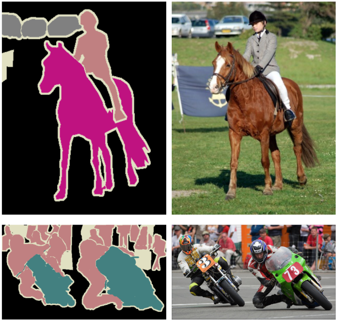
Dense Pixelwise Prediction in Computer Vision

Dense Pixelwise Prediction in Computer Vision

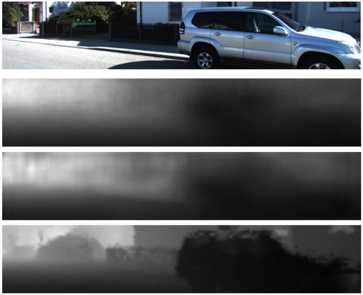


Semantic segmentation

Dense Pixelwise Prediction in Computer Vision

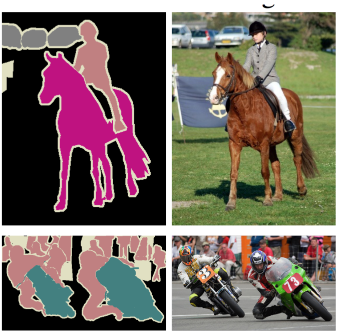


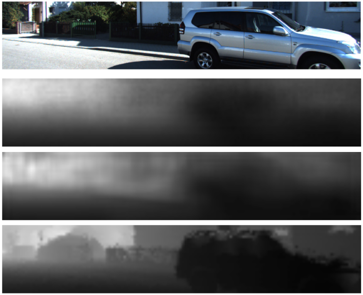
Semantic segmentation



Depth map prediction

Dense Pixelwise Prediction in Computer Vision





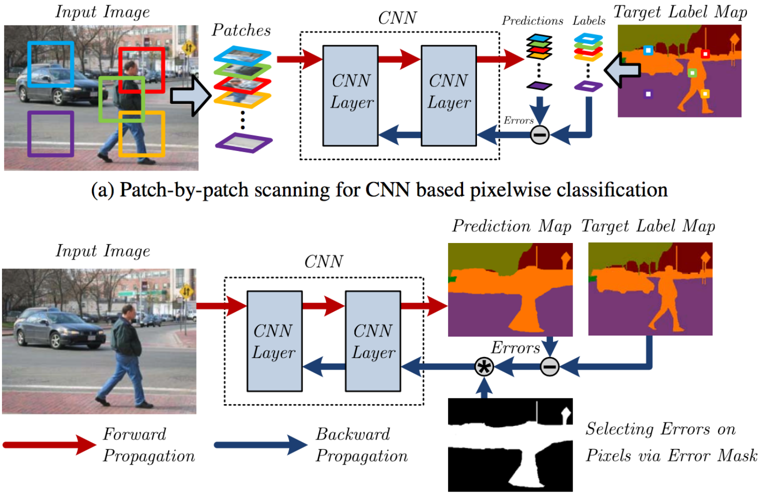
Depth map prediction

Semantic segmentation

Framewise classification = dense pixelwise prediction!

Dense Pixelwise Prediction with Convnets

Patch-by-patch vs efficient

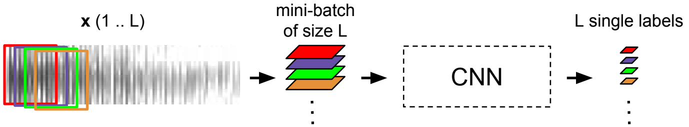


(a) Patch-by-patch scanning for CNN based pixelwise classification

(b) Our approach

Sequential: Framewise classification

Spliced (bad) vs efficient (good)



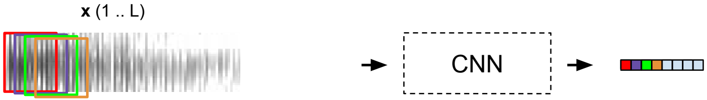
x (1 .. L)

mini-batch of size L

CNN

L single labels

A Spliced, inefficient. **Classification** viewpoint



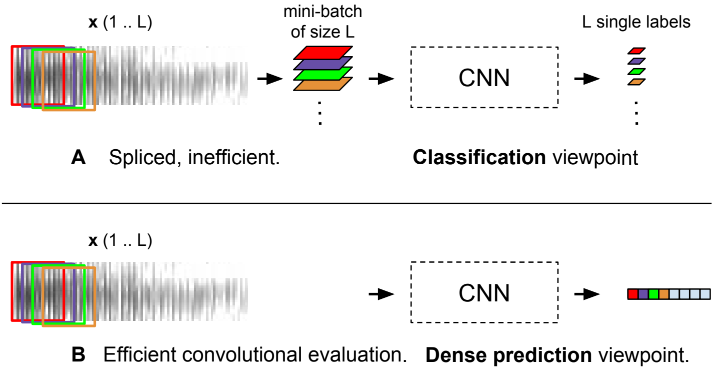
x (1 .. L)

CNN

B Efficient convolutional evaluation. **Dense prediction** viewpoint.

Sequential: Framewise classification

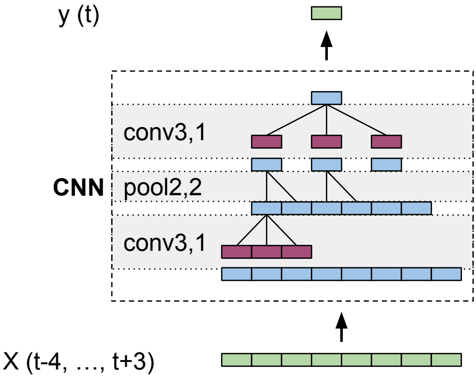
Spliced (bad) vs efficient (good)


x (1 .. L)
mini-batch
of size L
CNN
L single labels
A Spliced, inefficient.
Classification viewpoint
x (1 .. L)
CNN
B Efficient convolutional evaluation.
Dense prediction viewpoint.

Can we have time-pooling into the CNN?

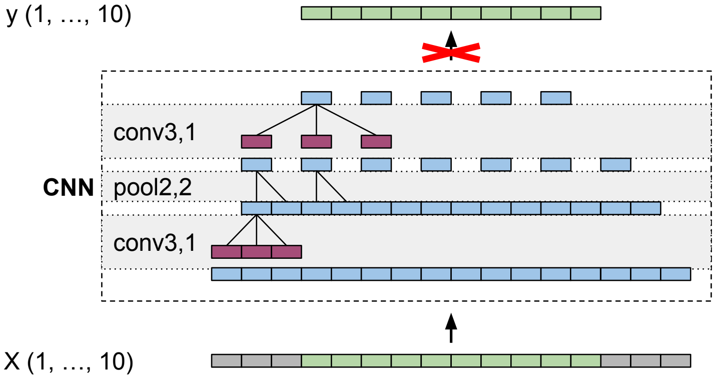
Time-pooling

XE: Toy CNN with pooling in time



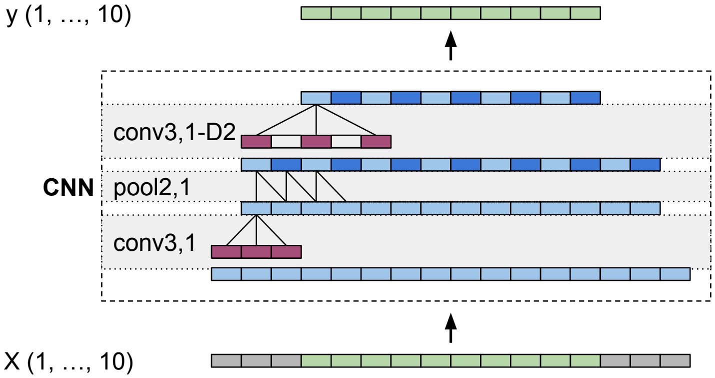
Time-pooling

ST: Downsampling is a problem



Time-pooling → Time-dilated convolutions

ST: Solution to downsampling



Based on Spatial dilated convolution [Li et al., 2014, Yu and Koltun, 2016] or OverFeat [Sermanet et al., 2013]

What did we gain?

With dense prediction viewpoint for ST

- CNNs with strided pooling in time

What did we gain?

With dense prediction viewpoint for ST

- CNNs with strided pooling in time
    - Better performance [Sercu et al., 2016, Sercu and Goel, 2016]

What did we gain?

With dense prediction viewpoint for ST

- CNNs with strided pooling in time
    - Better performance [Sercu et al., 2016, Sercu and Goel, 2016]
- While maintaining efficient dense prediction, enabling:

What did we gain?

With dense prediction viewpoint for ST

- CNNs with strided pooling in time
    - Better performance [Sercu et al., 2016, Sercu and Goel, 2016]
- While maintaining efficient dense prediction, enabling:
    - Efficient convolutional evaluation

What did we gain?

With dense prediction viewpoint for ST

- CNNs with strided pooling in time
  - Better performance [Sercu et al., 2016, Sercu and Goel, 2016]
- While maintaining efficient dense prediction, enabling:
  - Efficient convolutional evaluation
  - Batch Normalization

What did we gain?

With dense prediction viewpoint for ST

- CNNs with strided pooling in time
    - Better performance [Sercu et al., 2016, Sercu and Goel, 2016]
- While maintaining efficient dense prediction, enabling:
    - Efficient convolutional evaluation
    - Batch Normalization
- End-to-end models with CNNs
    - Can accept downsampling
    - But this allows to pool more than acceptable amount of downsampling

Result on switchboard

Big n-gram LM

	SWB	CH
IBM 2015 DNN+RNN+CNN	8.8 †	15.3 †
IBM 2016 RNN+VGG+LSTM	7.6 †	13.7 †
MSR 2016 ResNet *	8.6	14.8
MSR 2016 LACE *	8.3	14.8
MSR 2016 BLSTM *	8.7	16.2
VGG + BN	8.1	15.9
VGG + BN + pool	**7.7**	**14.5**

- † model combination / $*$ smaller LM
- Note: simple language model. Followed by: LM rescoring

Figures sources and references

- Slide 3 (Semantic Segmentation) [Long et al., 2015] Figure 6
- Slide 3 (Depth map prediction) [Eigen et al., 2014] Figure 4
- Slide 4 (Patches) [Li et al., 2014] Figure 1

- Eigen, D., Puhrsch, C., and Fergus, R. (2014).
  Depth map prediction from a single image using a multi-scale deep network.
  In *Advances in Neural Information Processing Systems*, pages 2366–2374.
- Li, H., Zhao, R., and Wang, X. (2014).
  Highly efficient forward and backward propagation of convolutional neural networks for pixelwise classification.
  arXiv:1412.4526.
- Long, J., Shelhamer, E., and Darrell, T. (2015).
  Fully convolutional networks for semantic segmentation.
  CVPR.
- Sercu, T. and Goel, V. (2016).
  Advances in very deep convolutional neural networks for lvcsr.
  Proc. Interspeech.
- Sercu, T., Puhrsch, C., Kingsbury, B., and LeCun, Y. (2016).
  Very deep multilingual convolutional neural networks for lvcsr.
  Proc. ICASSP.
- Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2013).
  Overfeat: Integrated recognition, localization and detection using convolutional networks.
  arXiv:1312.6229.
- Yu, F. and Koltun, V. (2016).
  Multi-scale context aggregation by dilated convolutions.
  proc ICLR.

Thank you! Questions?

Here's a cake which doesn't have anything to do with the talk