- Supplementary Material -
SEE: Towards Semi-Supervised End-to-End Scene Text Recognition

Christian Bartz and **Haojin Yang** and **Christoph Meinel**
Hasso Plattner Institute, University of Potsdam
Prof.-Dr.-Helmert Straße 2-3
14482 Potsdam, Germany
{christian.bartz, haojin.yang, meinel}@hpi.de

We created tools, which help us to gain insights into the training progress of our network. The most important tool runs every iteration and shows the network predictions for one given input image. In the following sections, we present some images showing the evolution of the network during training for different experiments.

SVHN Experiments

We first show the evolution of training on one of the SVHN-based datasets we generated. Here, we used the dataset with randomly positioned house numbers. Figure 1 shows how the predictions of the network evolve over the course of training. The top row shows the progress for the first training step. The first training step, is to train the network on images containing only one randomly positioned house number. The top-left image shows the prediction of the network after the first iteration and the top-right image shows the prediction of the network after the last iteration, using this dataset.

The second step, is to fine-tune the network on images that contain two randomly positioned house numbers. The image in the middle shows the prediction of the network after the first iteration with this dataset, and the image on the bottom shows the prediction after the last iteration.

All images are structured in the same way:

Top-left: The original image with the predicted text bounding box.

Top-right: The input image for the recognition network.

Top-right corner: The predicted textual content.

Bottom: The visualization of the regions which excite the convolutional parts of the network most (created with Visualbackprop (Bojarski et al. 2016)).

FSNS Experiments

Figure 2 shows how the predictions of the network evolve over the course of training on the FSNS dataset. The image on the top shows the prediction of the network after the first train iteration. The image in the middle shows intermediate results during the training. The image on the bottom shows the prediction of the network after the last iteration.

These images have the following structure:

Top row: original Input sample from the dataset with the predicted bounding boxes.

Top-right corner: The predicted textual content.

Bottom row: Result of applying Visualbackprop to the localization network of each individual view of the sample.

Insights

The visualization with Visualbackprop clearly shows that the network is able to learn features, which allow the localization network to locate text in each example image. The results on the FSNS dataset are the most interesting. Here it is possible to see the ability of the convolutional layers to learn that all images containing random noise do not contain any helpful information. It is also possible to see that, after a while, the network only concentrates on text in the image and ignores everything else that might still be there. All this is learned by the network itself, using only the guidance of the recognition network.

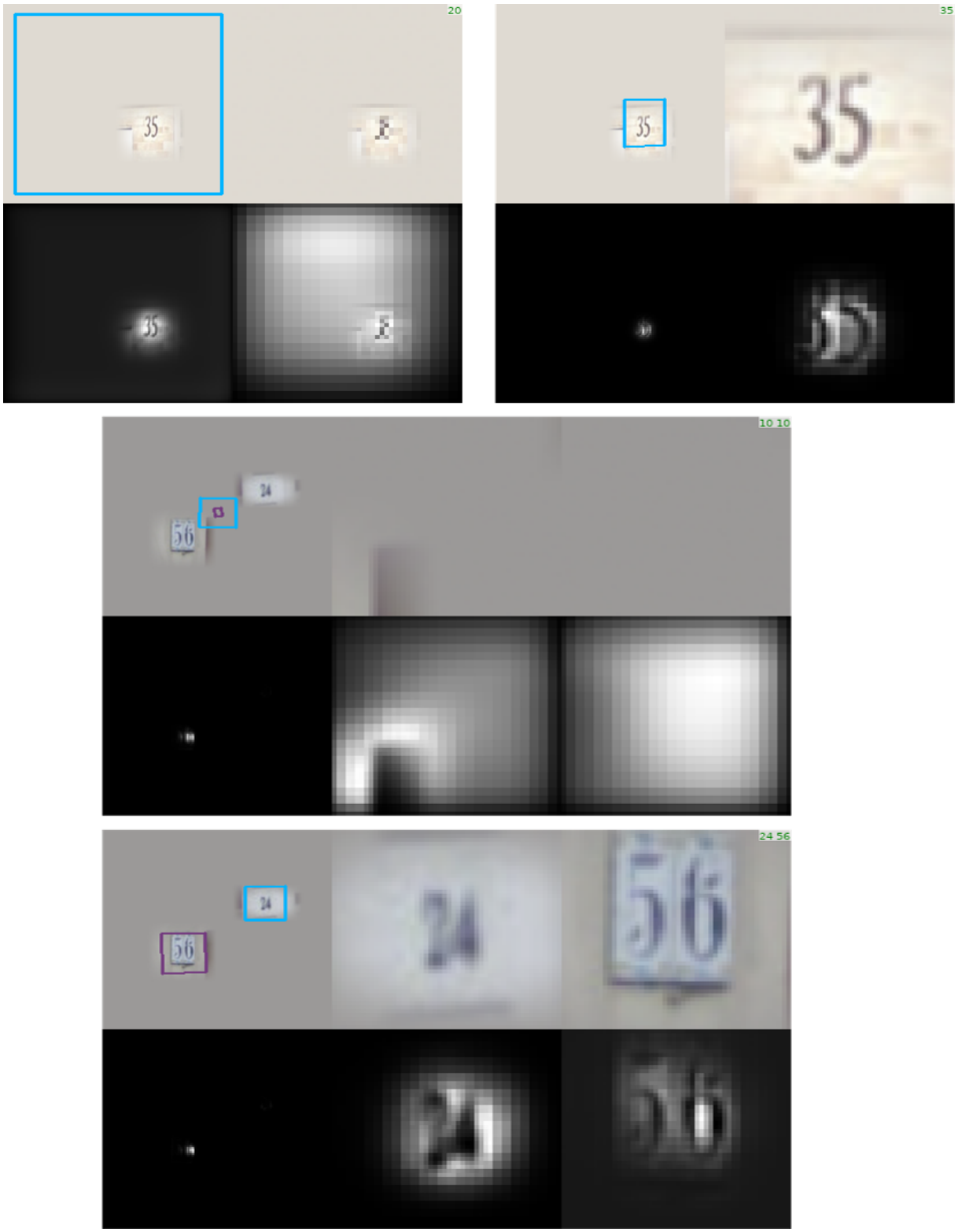

Figure 1: Visualization of training progress for experiment on dataset with randomly positioned house numbers.

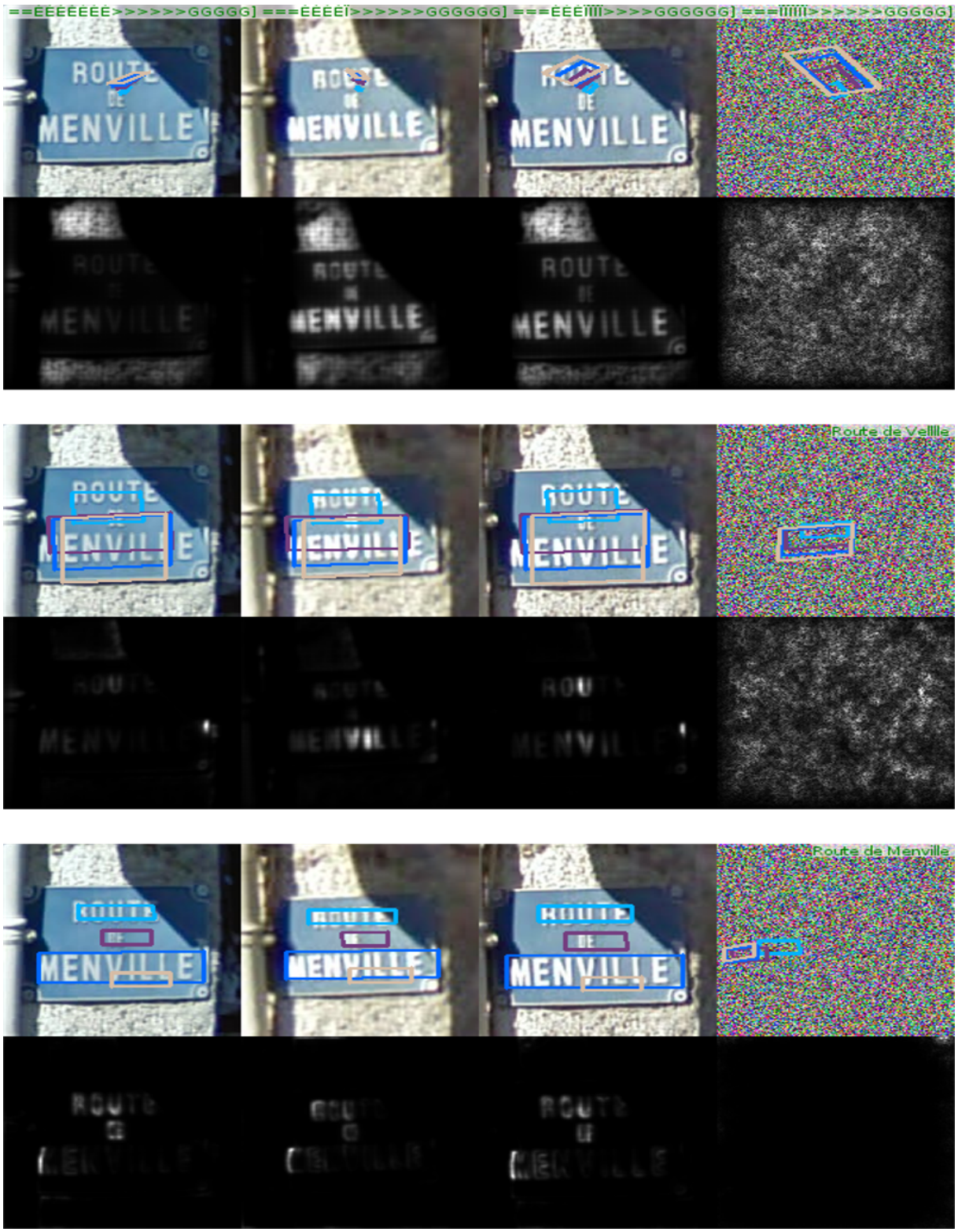

Figure 2: Visualization of training progress for experiments on the FSNS dataset.