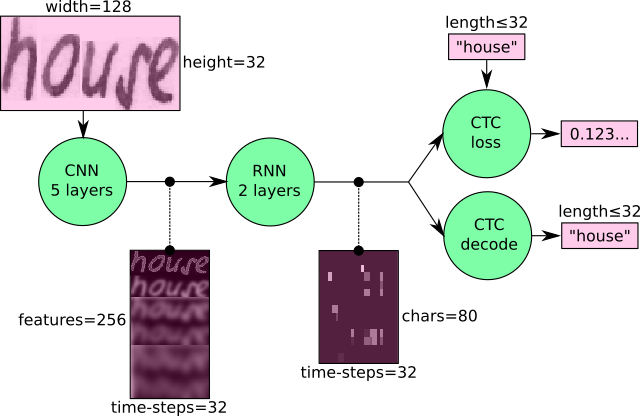
# Handwritten Text Recognition

Offline Handwritten Text Recognition (HTR) systems transcribe text contained in scanned images into digital text. We built a Neural Network which is trained on word-images from IAM dataset. As the input layer can be kept small for work-images, NN-training is feasible on the CPU. (GPU would be much better).

Our Neural Network (NN) consists of convolutional NN layers, recurrent NN layers and a final Connectionist Temporal Classification (CTC) layer.



In a more formal way the above Neural Network maps an image (matrix) of size W x H to a character sequence (c1, c2, ….). The text is recognized on character level, therefore words or texts not contained in the training data can be recognized too (as long as the individual characters get correctly classified).

**Convolution Neural Network Layer:**

The input image is fed into the CN layers. These layers are trained to extract relevant features from the image. Each layer consists of three operations.

1. Convolution operation, which applies filter kernel of size 5 x 5 in the first two layers and 3 x 3 in the last three layers to the input.
2. Then the non-linear RELU (Rectified Linear Unit) function is applied.
3. Finally, a pooling layer summarizes image regions and outputs a downsized version of the input.

After the image height is downsized, feature maps are added, so that the output feature map has a size of 32 x 256.

**Recurrent Neural Network Layer:**

The feature sequence contains 256 features per time-step, the RNN propagates relevant information through this sequence. The Long Short-Term Memory (LSTM) implementation of RNN’s is used. The RNN output sequence is mapped to a matrix of size 32 x 80.

**Connectionist Temporal Classification Layer:**

While training the Neural Network, the CTC is given the RNN output matrix and the ground truth text and it computes the loss value. While inferring, the CTC is only given the matrix and it decodes it into the final text.

**Pre-processing the input image:**

* The input image must be a grayscale image of size 128 x 32.
* The trained images have high contrast and Bold Writing Style
* So we used morphological transformations ( erosion and dilation ) on the input images before feeding them to the Neural Network.

The model predicted the handwritten text with an average accuracy of **70% (approximately).**