# 5. LSTM extensions
# BLSTM and MDLSTM

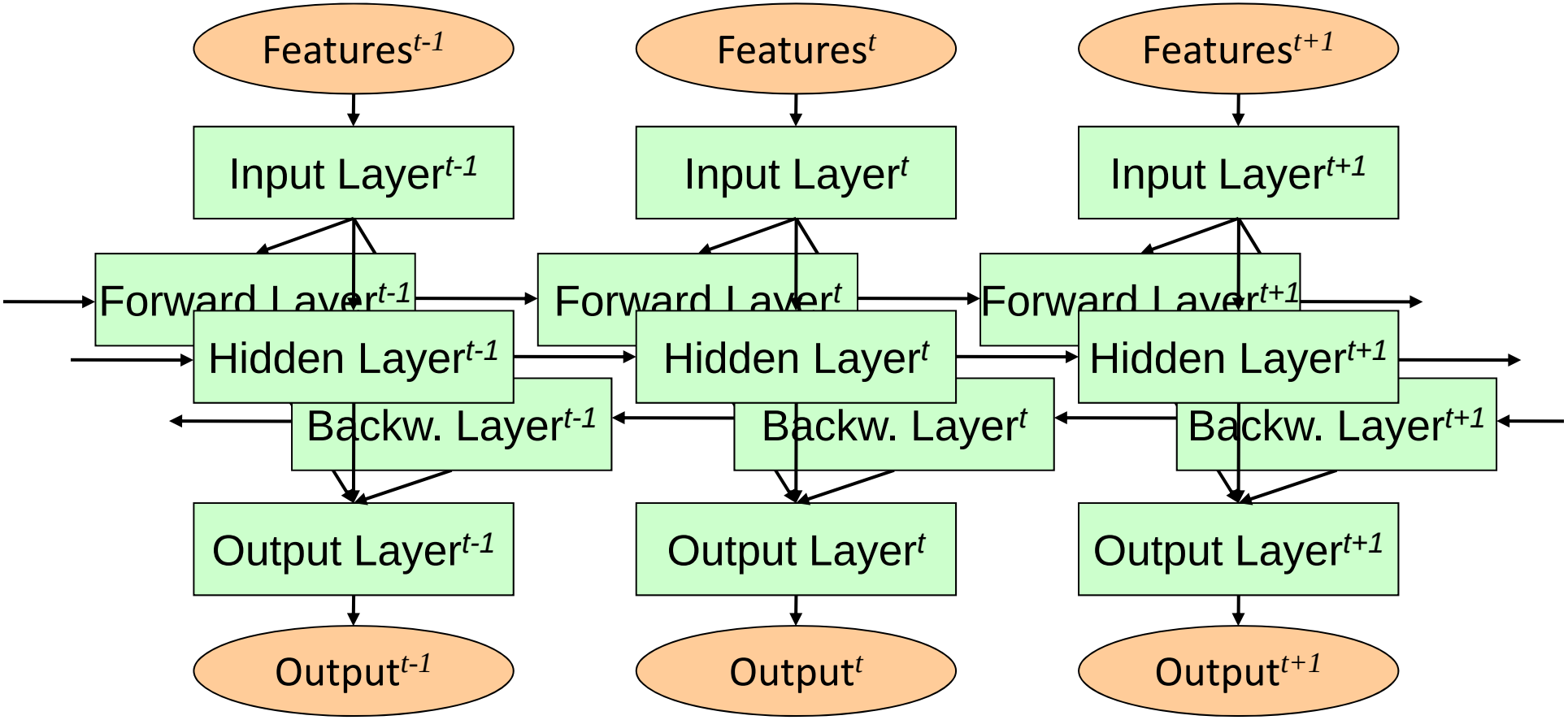
*Thomas Breuel, Volkmar Frinken,*
*Marcus Liwicki*




## LSTM is not Enough

- Several relevant, but more difficult problems exist
  - Complete sequence recognition
  - Sequence to sequence matching
- Complex applications domains
  - Speech recognition
  - Handwriting recognition
  - Protein localization
- Often the context from later is also interesting
  - Long or short ([a] or [a:])
  - Idea: Delayed output – but how long?
  - Better idea: use context from whole sequence

## Bidirectional RNN, resp. BLSTM




- Trained with backpropagation through time (forward path trough all time stamps for each hidden layer sequentially)

## Frame-Wise Phoneme Classification

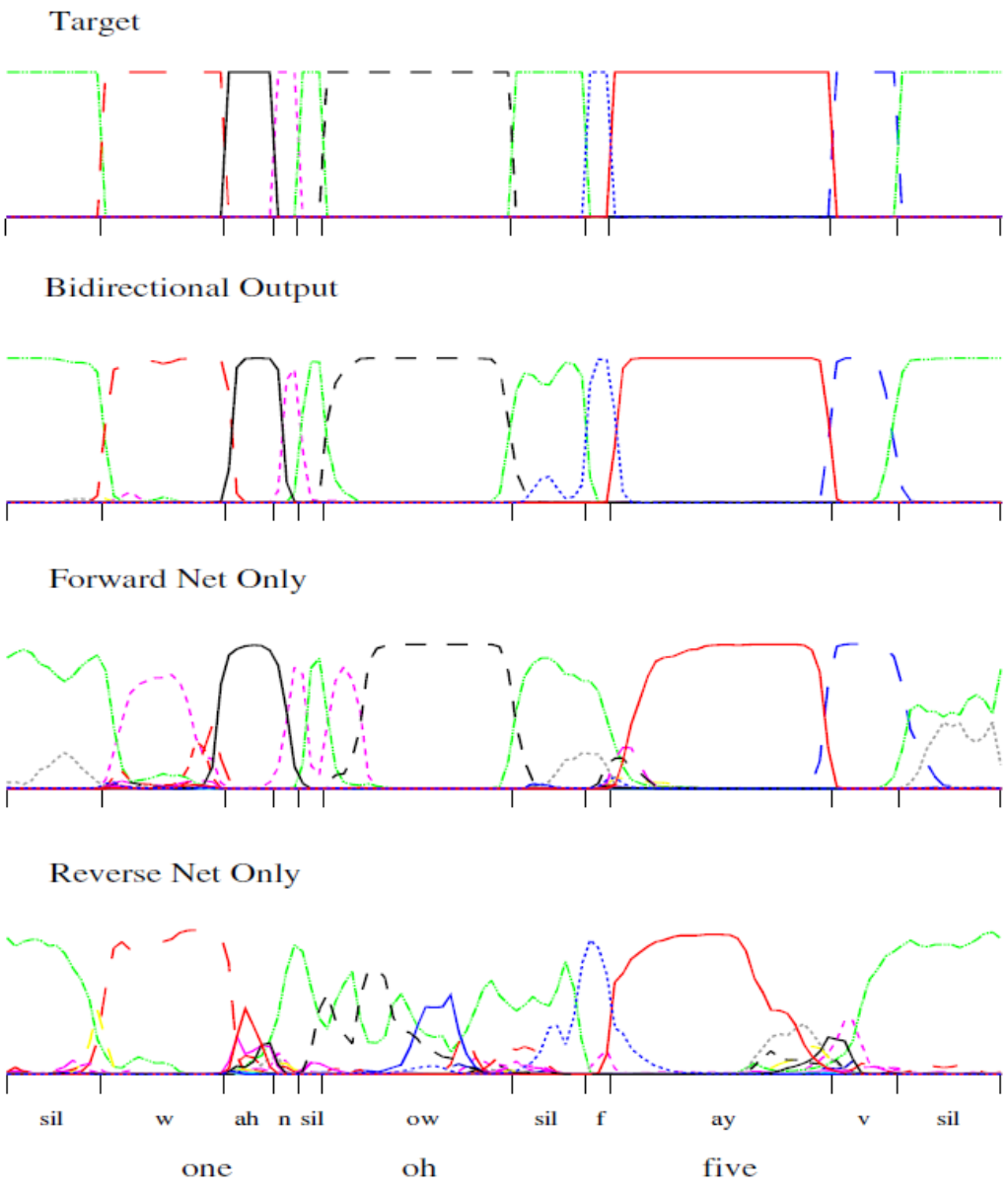
- TIMIT database
  - Texas Instruments and Massachusetts Institute of Technology
  - 3,696 training phonemes
  - 1,344 testing phonemes
  - Speaker independent
- Experiments with several comparable architectures
  - 26 input units (MFCC features)
  - 61 output units (one for each phoneme)
  - All networks had roughly the same number of weights (100,000)
  - Examples: BLSTM 93, LSTM 140, BRNN 185, RNN 275

## Results

| Network | Training set (%) | Test set (%) | Epochs |
|---|---|---|---|
| BLSTM (retrained) | 78.6 | 70.2 | 17 |
| BLSTM | 77.4 | 69.8 | 20.1 |
| BRNN | 76.0 | 69.0 | 170 |
| BLSTM (Weighted Error) | 75.7 | 68.9 | 15 |
| LSTM (5 frame delay) | 77.6 | 66.0 | 34 |
| RNN (3 frame delay) | 71.0 | 65.2 | 139 |
| LSTM (backwards, 0 frame delay) | 71.1 | 64.7 | 15 |
| LSTM (0 frame delay) | 70.9 | 64.6 | 15 |
| RNN (0 frame delay) | 69.9 | 64.5 | 120 |
| MLP (10 frame time-window) | 67.6 | 63.1 | 990 |
| MLP (no time-window) | 53.6 | 51.4 | 835 |
| RNN (Chen and Jamieson, 1996) | 69.9 | 74.2 | – |
| RNN (Robinson, 1994; Schuster, 1999) | 70.6 | 65.3 | – |
| BRNN (Schuster, 1999) | 72.1 | 65.1 | – |

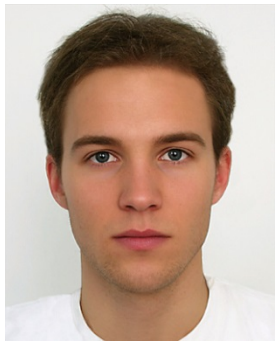
Retraining is done by increasing the target delay after each 5 epochs

# Closer Look into the Behaviour



## Going Into Multiple Dimensions

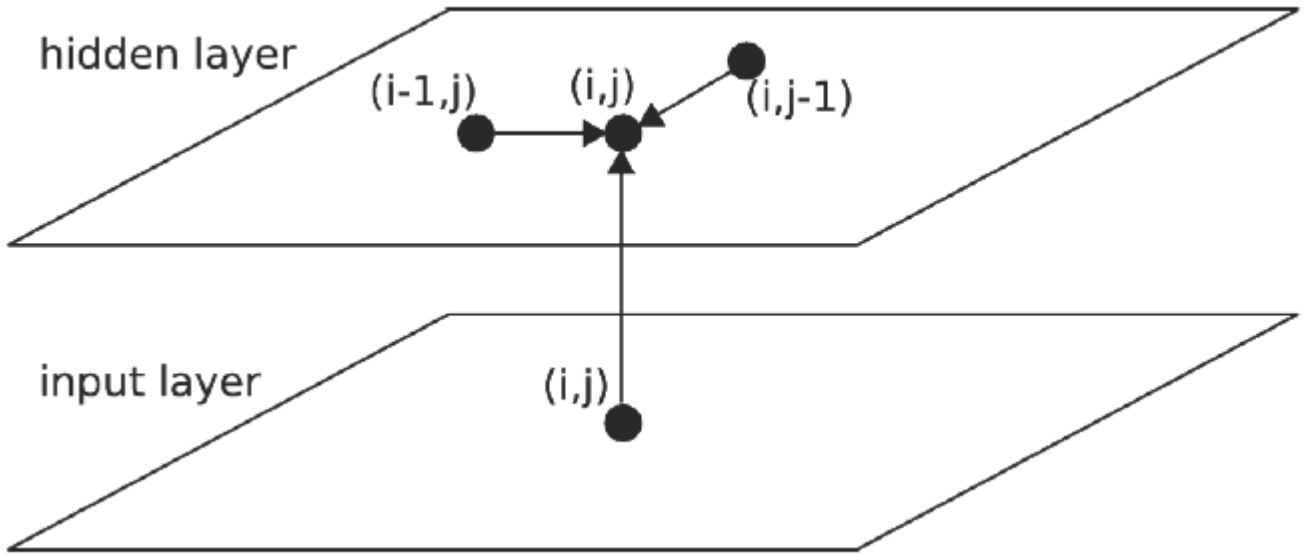
- If input size is fixed, MLP can be applied, but what if size is unknown?
- Face recognition



- Idea: Sliding window in multiple directions

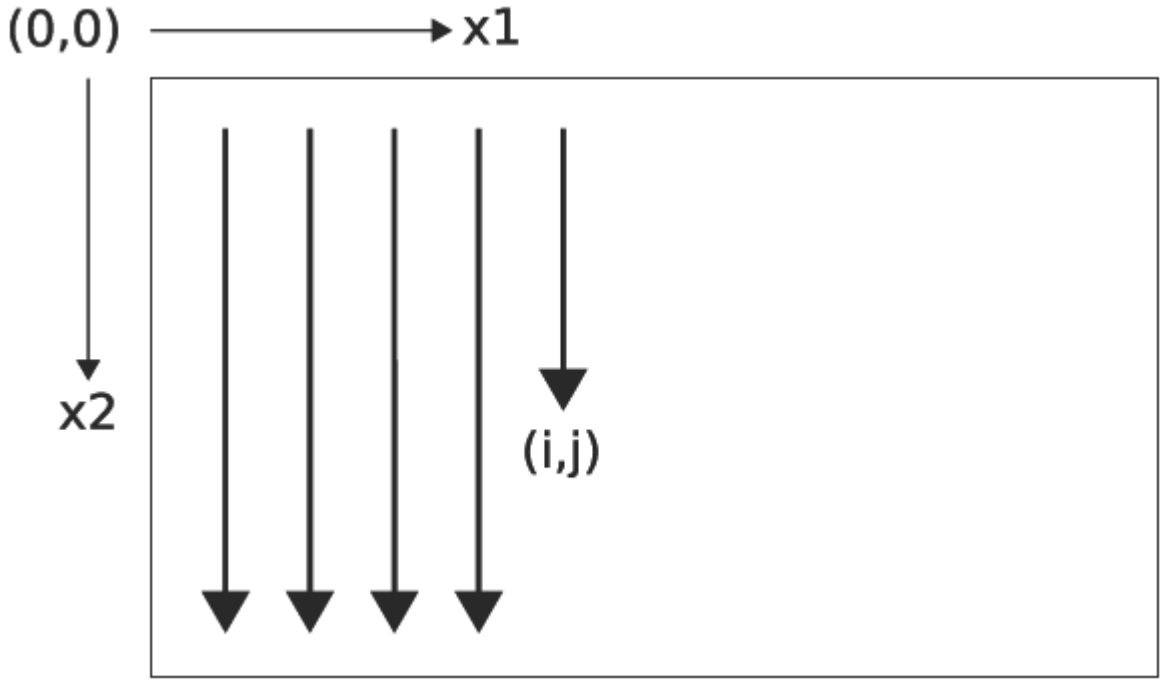
## Going Into Multiple Dimensions

- Concrete idea (like DAG-RNN):
  - Each neuron receives external input and its own activation from one step back along all dimensions
  - Can be applied to any dimensional sequences (img 2D, video 3D)



## Sequence Ordering Through Forward Pass

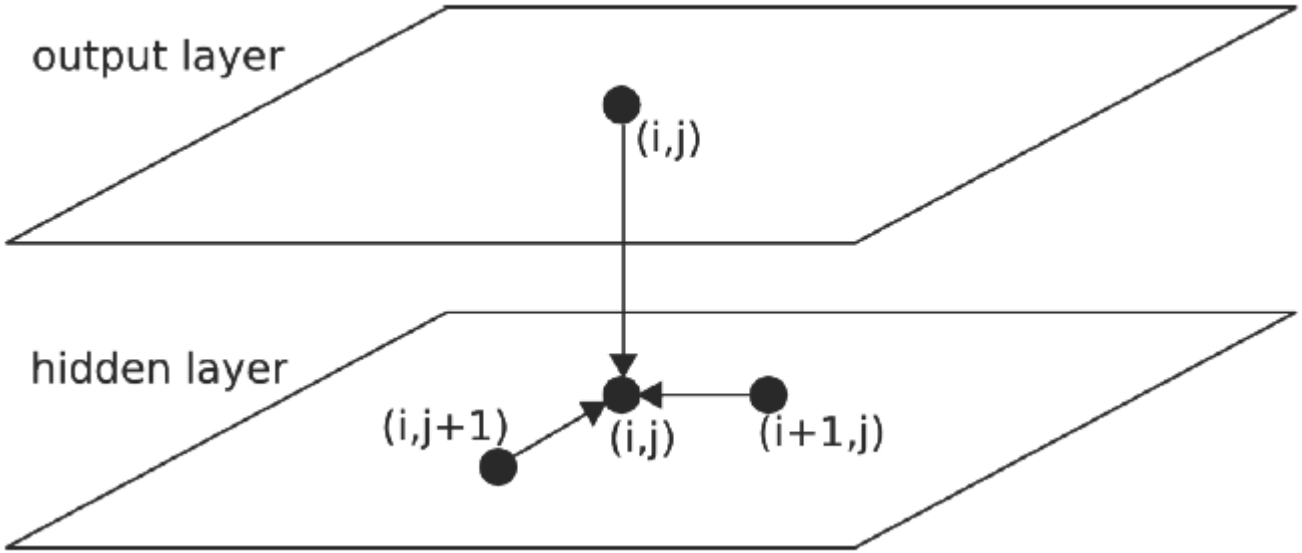
- Ensure that each previous' step output is already calculated
- Example:




- Note: Boundaries have to be omitted

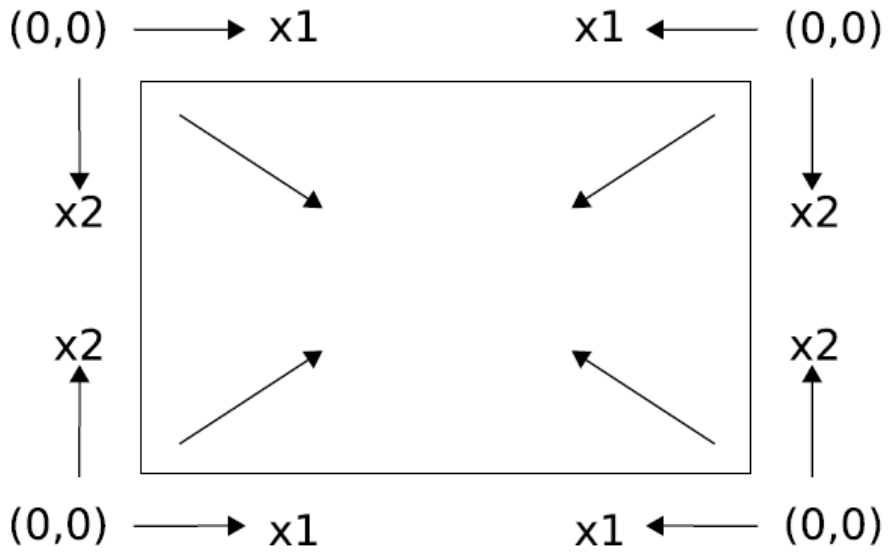
## Training

- N-dimensional backpropagation through time



## Multidirectional MDRNN

- Each neuron receives external input and its own activation from one step back along all dimensions
- Can be applied to any dimensional sequences (2D image, 3D movie, 4D with time, 6D for robot control)
- Idea: Use $2^D$ hidden layers

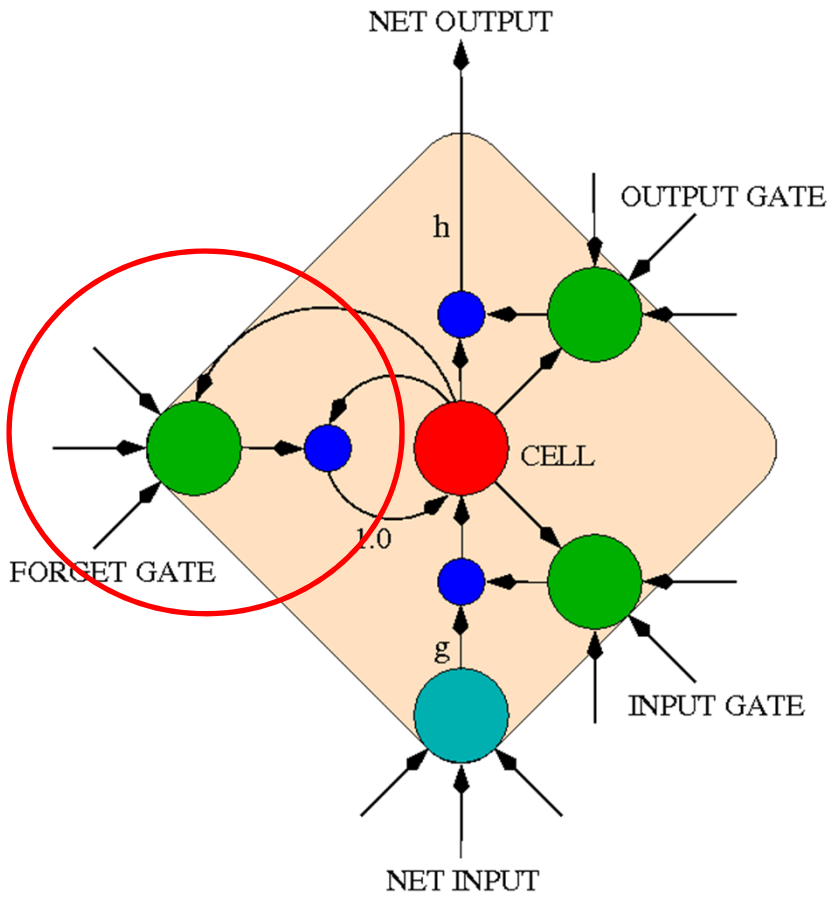


## Problem

- Does the complexity explode?
- $2^d$ seems to be quite large
- However
  - Number of weights has more influence
  - Several calculations can be shared
- Furthermore
  - Reduce the size of the hidden layers with increasing dimensionality
  - It has been found that for speech recognition the number of weights was reduced to half and MDRNN gave still better results
- Main scaling concern is the size of the data, i.e., the length of the sequence

## Combining idea with LSTM Unit

- Introduce $2^d$ self-connections, i.e., $2^d$ forget gates, each connected along one dimension
- However, only one input gate (connected to all dimensions)
- Also, only one output gate, since only the cell state is considered



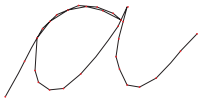
# 6. Applications
# a. Handwriting Recognition

*Thomas Breuel, Volkmar Frinken,*
*Marcus Liwicki*



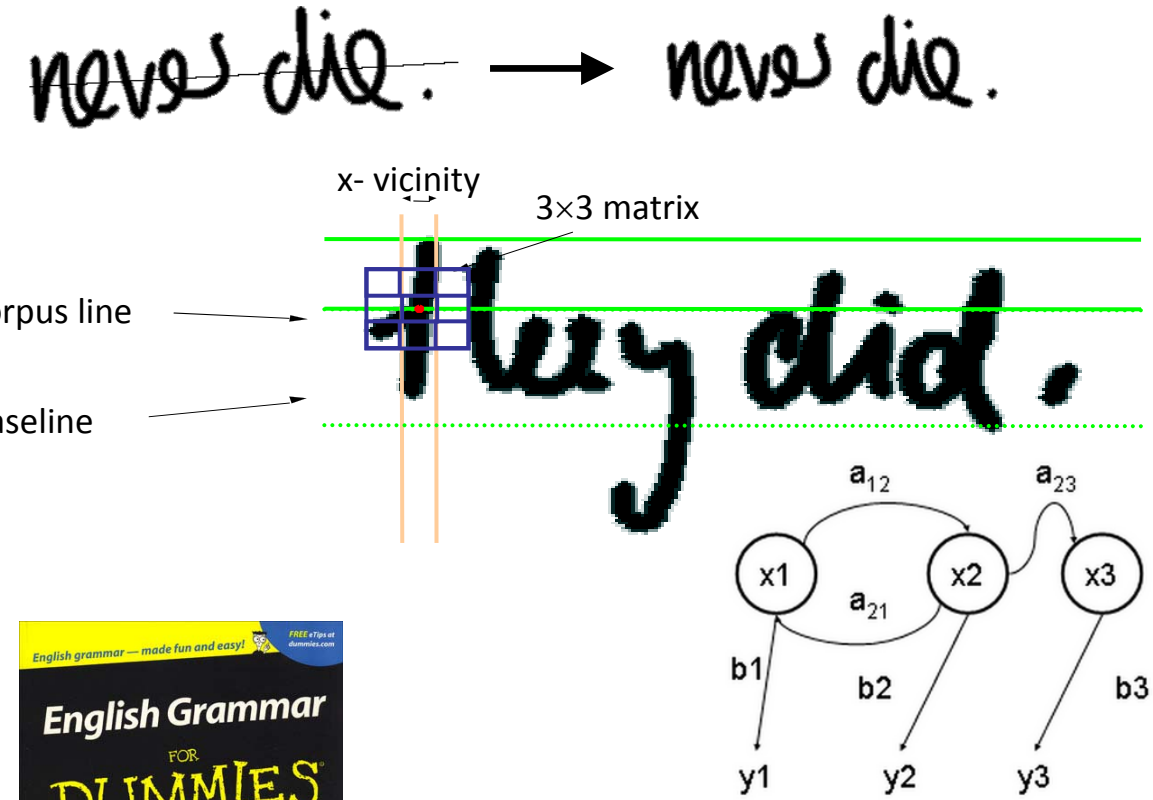

# First DAR application: Handwriting Recognition

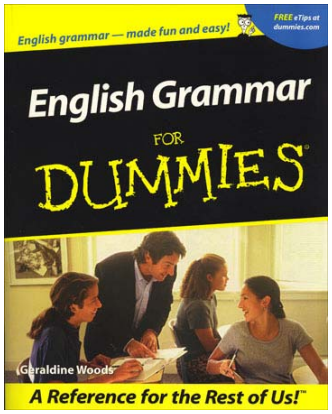
- eBeam system used
  - Pen in special casing
  - Sends signals
  - Receiver in one corner
- Result: sequence of time-stamped points
  - On-line format



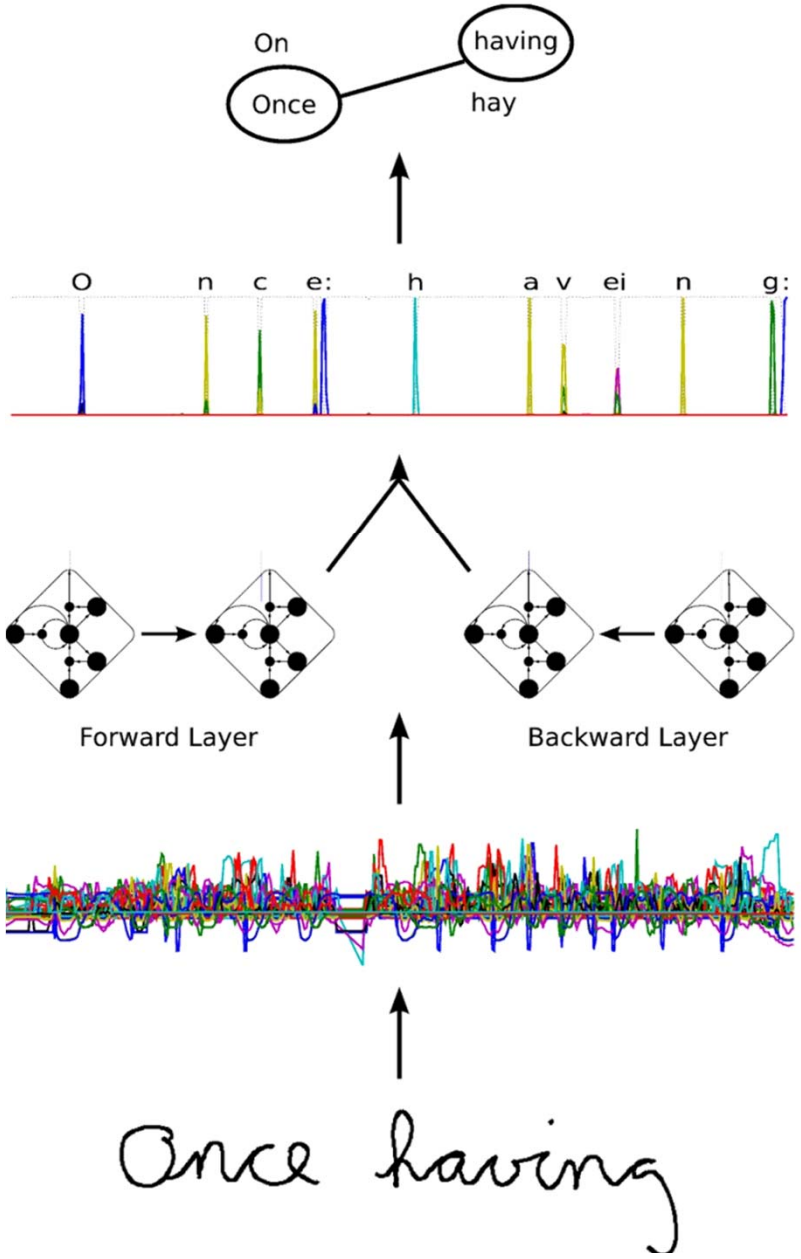
# Handwriting Recognition Sub-Tasks

- Handwriting
- Preprocessing
- Improved data
- Feature extraction
- Features
- Classification
- Alternates
- Post-processing
- Text



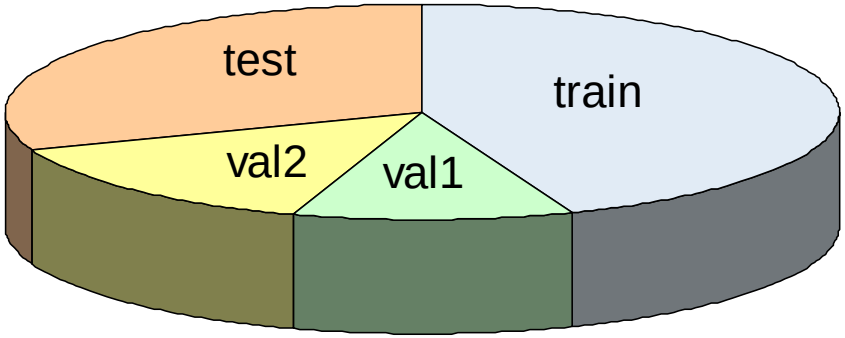



# Overall System



## Experimental Setup

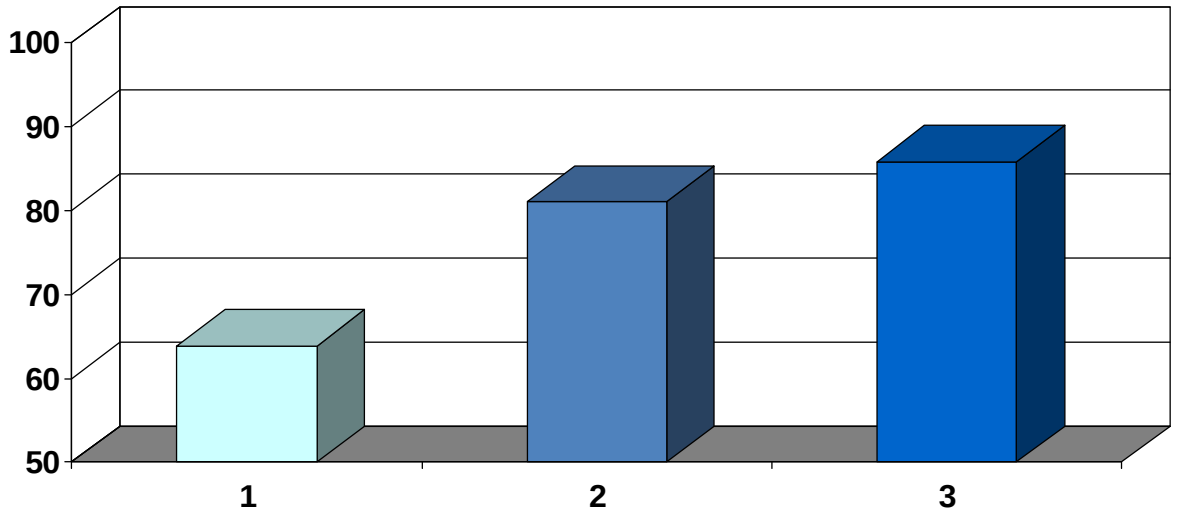
- IAM-OnDB-t2 benchmark[1]:
  - Training set, two validation sets, test set
  - Open vocabulary, 82 characters, 5.8% OOVs




- LM trained on three corpora (Brown, LOB, Wellington)
- Accuracy measured on the word level

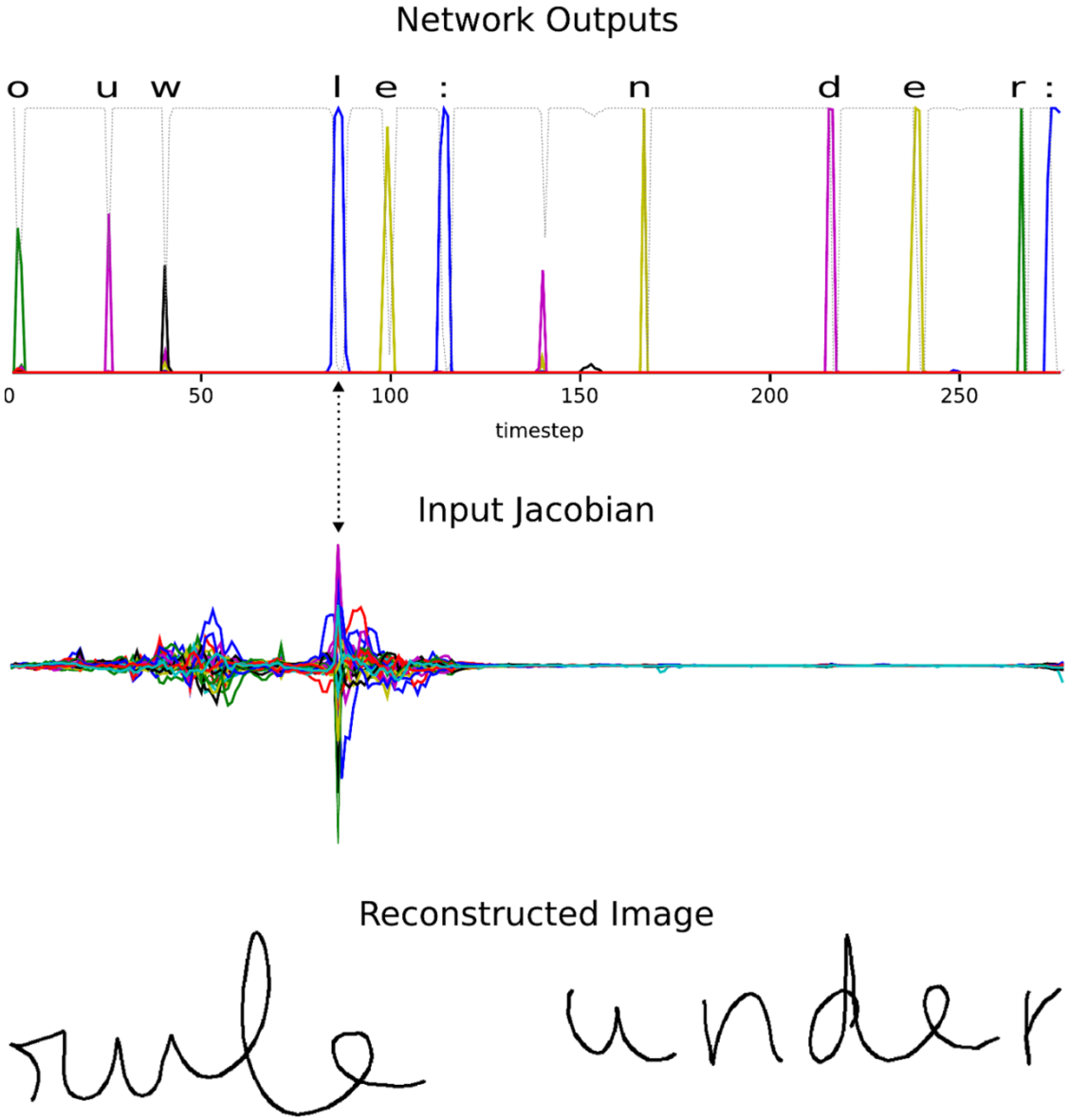
## Handwriting Recognition Experiments

- Experiments
  1. 63.86% with Hidden Markov Model (HMM)
  2. 81.05% with BLSTM (100 cells) and CTC
  3. 86 % after combination of several classifiers



## Information Preservation Experiment

- Example
  - Output at l
  - Amount of information for each cell derived from each time stamp
  - Called input Jacobian
  - Estimated by Backprop.



## MDLSTM

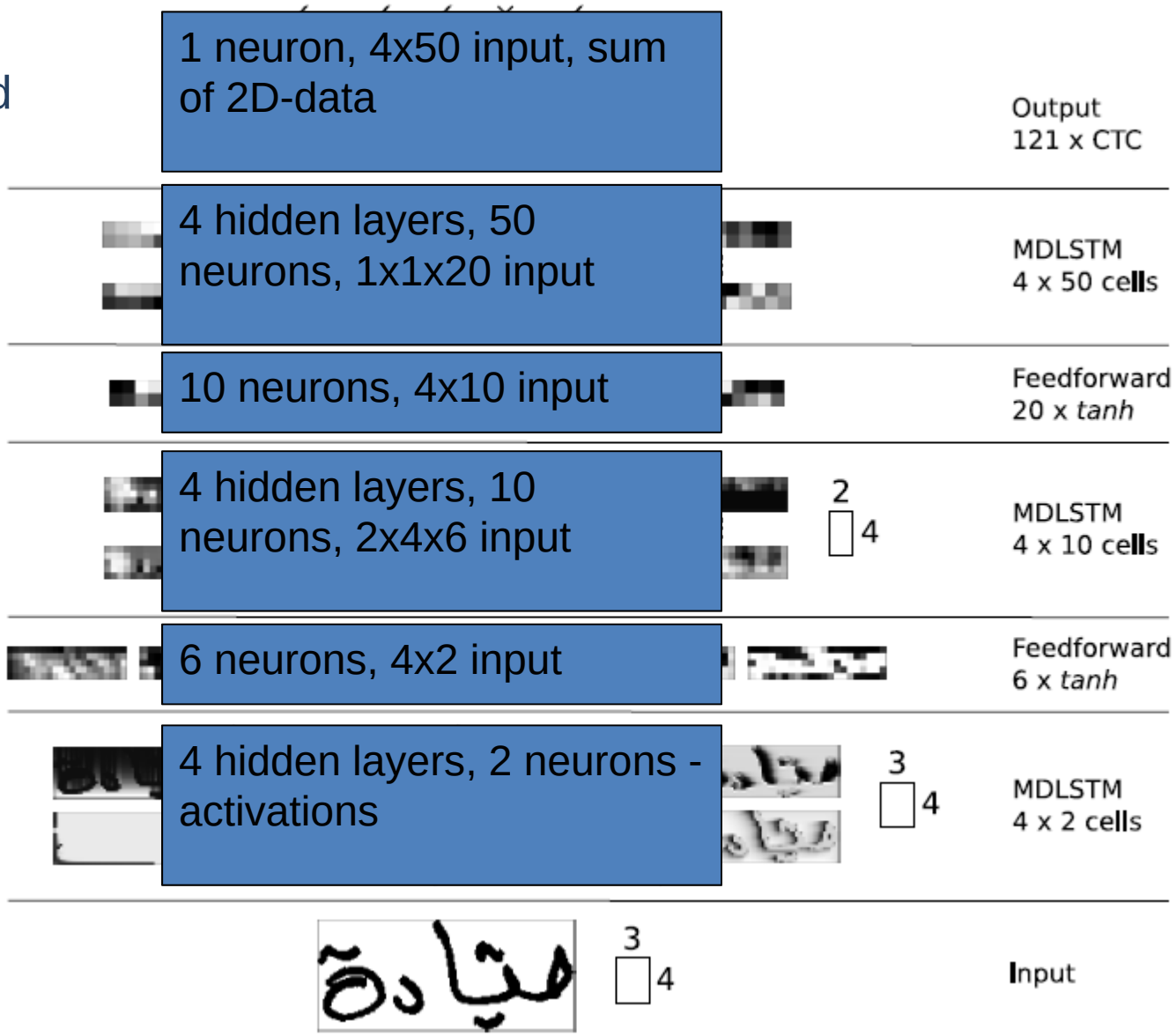
MDLSTM layers and feed-forward layers

4x3=>12-dim vector

Small at bottom large at top

159,369 weights but most at top

1D at top by summing up



# ICDAR 2007 Arabic handwriting recognition contest

| System | SET f | | | SET s | | |
|---|---|---|---|---|---|---|
| | top 1 | top 5 | top 10 | top 1 | top 5 | top 10 |
| CACI-3 | 14.28 | 29.88 | 37.91 | 10.68 | 21.74 | 30.20 |
| CACI-2 | 15.79 | 21.34 | 22.33 | 14.24 | 19.39 | 20.53 |
| CEDAR | 59.01 | 78.76 | 83.70 | 41.32 | 61.98 | 69.87 |
| MITRE | 61.70 | 81.61 | 85.69 | 49.91 | 70.50 | 76.48 |
| UOB-ENST-1 | 79.10 | 87.69 | 90.21 | 64.97 | 78.39 | 82.20 |
| PARIS V | 80.18 | 91.09 | 92.98 | 64.38 | 78.12 | 82.13 |
| ICRA | 81.47 | 90.07 | 92.15 | 72.22 | 82.84 | 86.27 |
| UOB-ENST-2 | 81.65 | 90.81 | 92.35 | 69.61 | 83.79 | 85.89 |
| UOB-ENST-4 | 81.81 | 88.71 | 90.40 | 70.57 | 79.85 | 83.34 |
| UOB-ENST-3 | 81.93 | 91.20 | 92.76 | 69.93 | 84.11 | 87.03 |
| SIEMENS-1 | 82.77 | 92.37 | 93.92 | 68.09 | 81.70 | 85.19 |
| MIE | 83.34 | 91.67 | 93.48 | 68.40 | 80.93 | 83.73 |
| SIEMENS-2 | 87.22 | 94.05 | 95.42 | 73.94 | 85.44 | 88.18 |
| **MDLSTM** | **91.43** | **96.12** | **96.75** | **78.83** | **88.00** | **91.05** |

A. Graves and J. Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. In Advances in Neural Information Processing Systems 21, 2009.

# Results of the ICDAR 2009 Arabic HWR Contest

| System | Word Accuracy | Time/Image |
|---|---|---|
| CTC | 81.06% | 371.61 *ms* |
| Arab-Reader HMM | 76.66% | 2583.64 *ms* |
| Multi-Stream HMM | 74.51% | 143,269.81 *ms* |

4 Summarized results from the offline Arabic handwriting recognition competition

Volker Märgner, Haikal El Abed, "ICDAR 2009 Arabic Handwriting Recognition Competition," Document Analysis and Recognition, International Conference on, pp. 1383-1387, 2009 10th International Conference on Document Analysis and Recognition, 2009

# Results of the ICDAR 2009 French HWR Contest

| System | Word Accuracy |
|---|---|
| CTC | 93.17% |
| HMM+MLP Combination | 83.17% |
| Non-Symmetric HMM | 76.34% |

Summarized results from the offline (French) handwriting recognition competition

Grosicki, E.; El Abed, H.; , "ICDAR 2009 Handwriting Recognition Competition," *Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on* , vol., no., pp.1398-1402, 26-29 July 2009