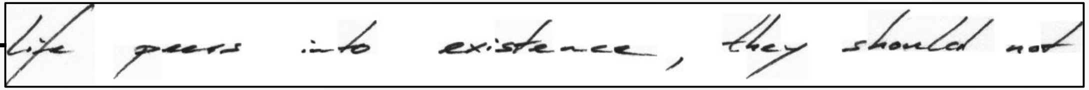
Boosting the deep multidimensional long short-term memory network for handwritten recognition systems

Dayvid Castro[1]
Byron L. D. Bezerra[1]
Mêuser Valença[1]

[1] Polytechnic School of Pernambuco
University of Pernambuco
Recife, Brazil

Handwriting Text Recognition (HTR)

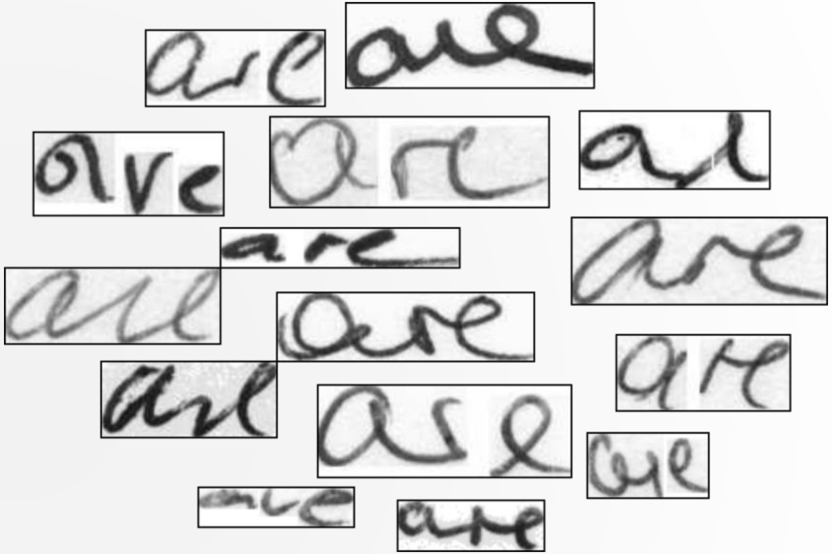
- Handwritten entry $\longmapsto$ digital representation
- **Offline Recognition**



$$\begin{bmatrix} 250 & 251 & 251 & \dots & 255 & 255 & 255 \\ 251 & 251 & 251 & \dots & 255 & 255 & 255 \\ 251 & 251 & 250 & \dots & 255 & 255 & 255 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots \\ 59 & 74 & 177 & \dots & 255 & 255 & 255 \\ 59 & 140 & 204 & \dots & 255 & 255 & 255 \\ 74 & 177 & 217 & \dots & 255 & 255 & 255 \end{bmatrix}$$

Offline HTR Challenges

- **Variability**
  - Different writing styles
  - Instrument (pen/pencil)
  - Paper type and quality
  - Space and time available
  - Vocabulary

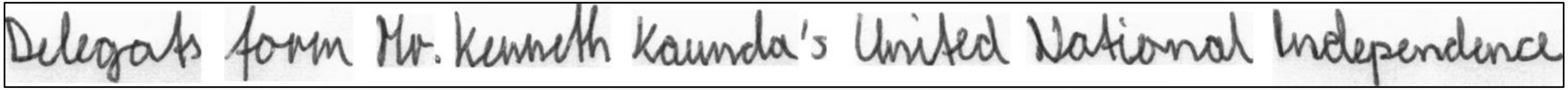


- **Similarity**
  - Similar shapes

Unconstrained Offline HTR

- Long text line sequences
- Cursive nature
- Different writing styles
- Large vocabulary

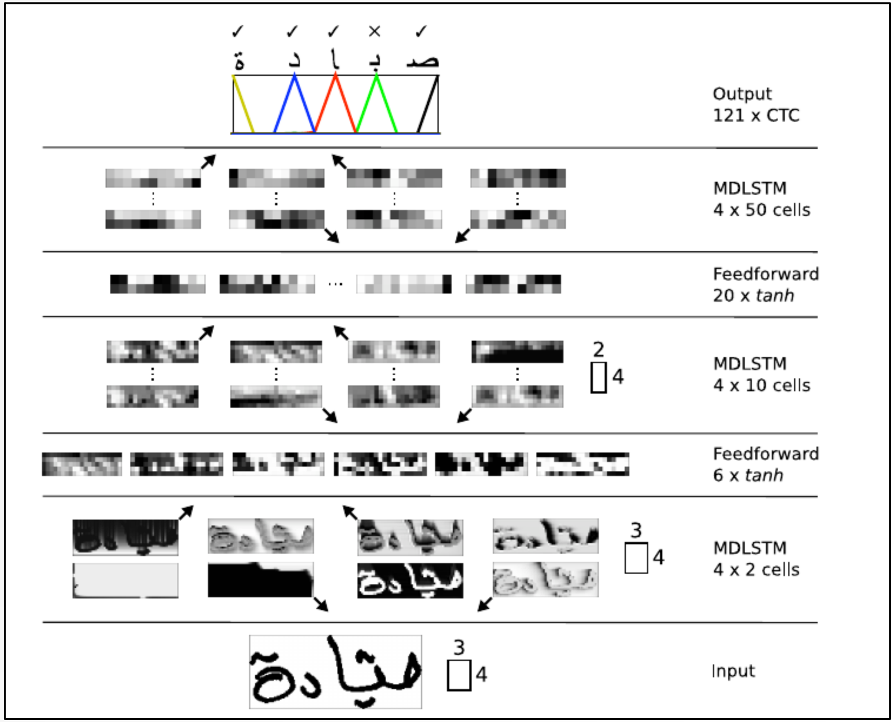
Open Problem



Segmentation-free approaches

Deep Neural Networks for Unconstrained HTR

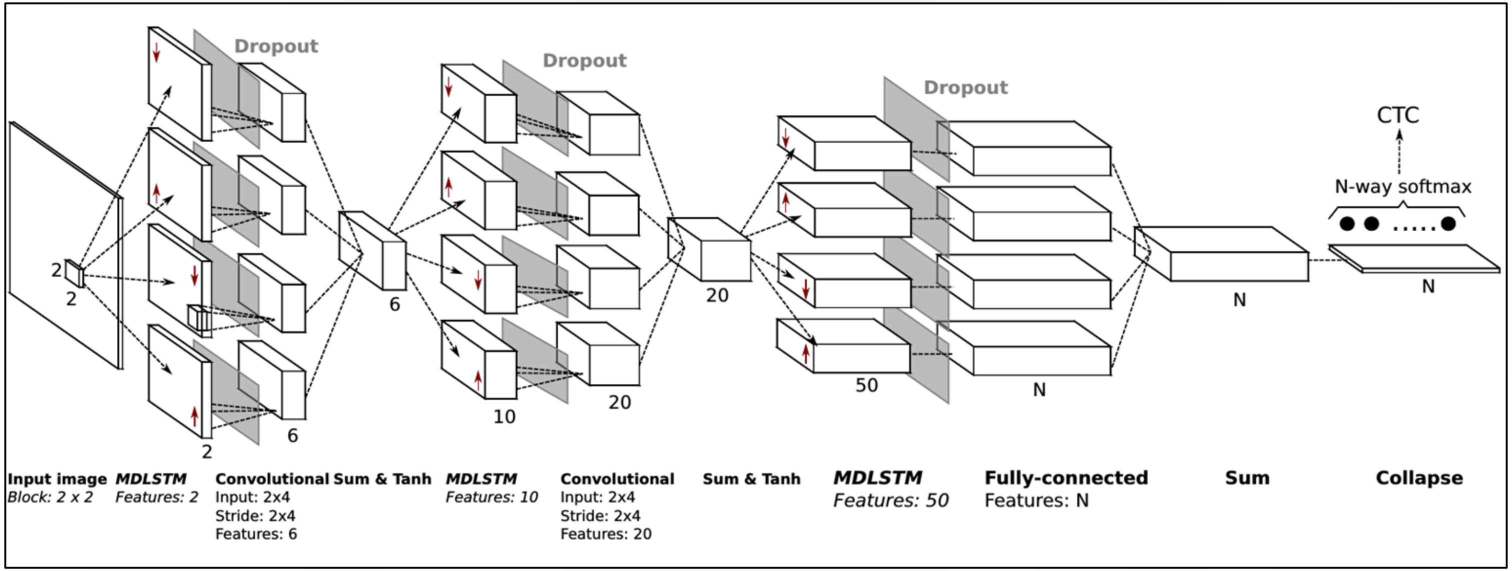
- Multiple Layers
- Representation Learning
- Building Blocks:
  - Convolutional and Pooling Layers
  - Recurrent Layers
  - Long Short-Term Memory (LSTM)
  - (Bi x Multi)dimentional flow
  - CTC



Graves et al. 2009

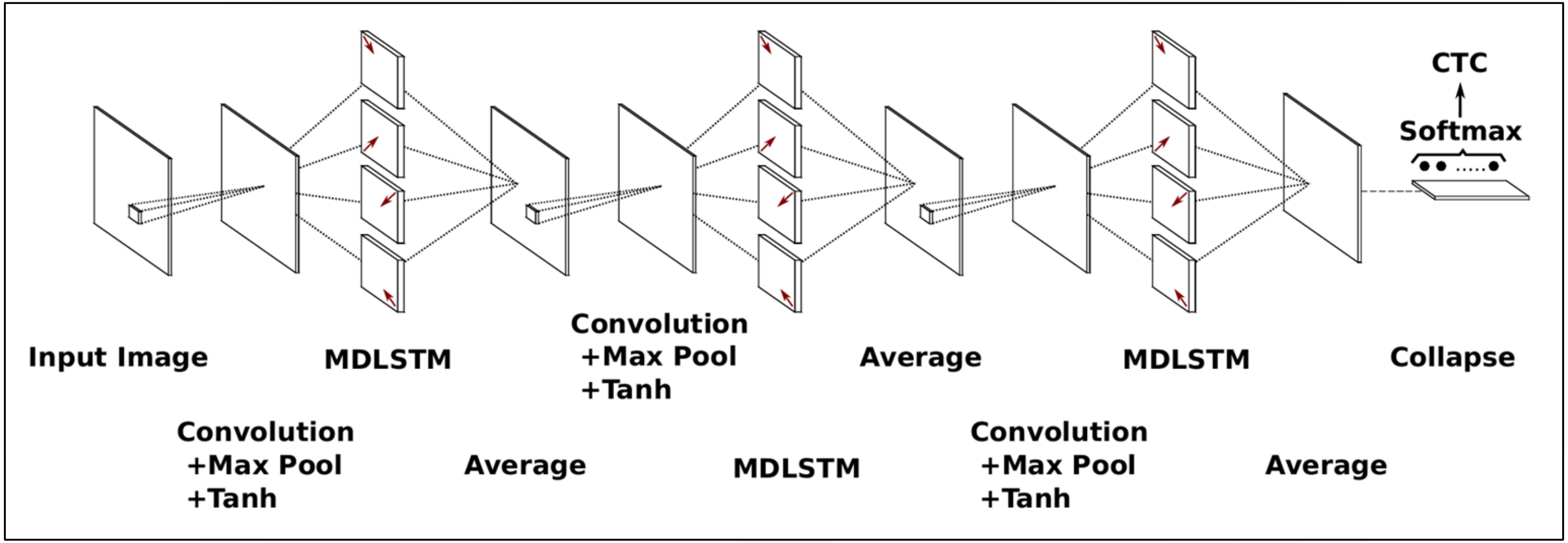
MDLSTM Network Hierarchy in HTR

Pham et al. 2014



MDLSTM Network Hierarchy in HTR

Voigtlaender et al. 2016



GPU implementation of MDLSTM (RETURNN tool) -> Deeper configurations

Hypothesis and Proposal

- ❖ **The goal**
- ❖ **Optical model proposal**

Main Goal

The main goal of this work was to investigate alternative optical modeling approaches that can contribute to the optimization of offline and unconstrained HTR systems.

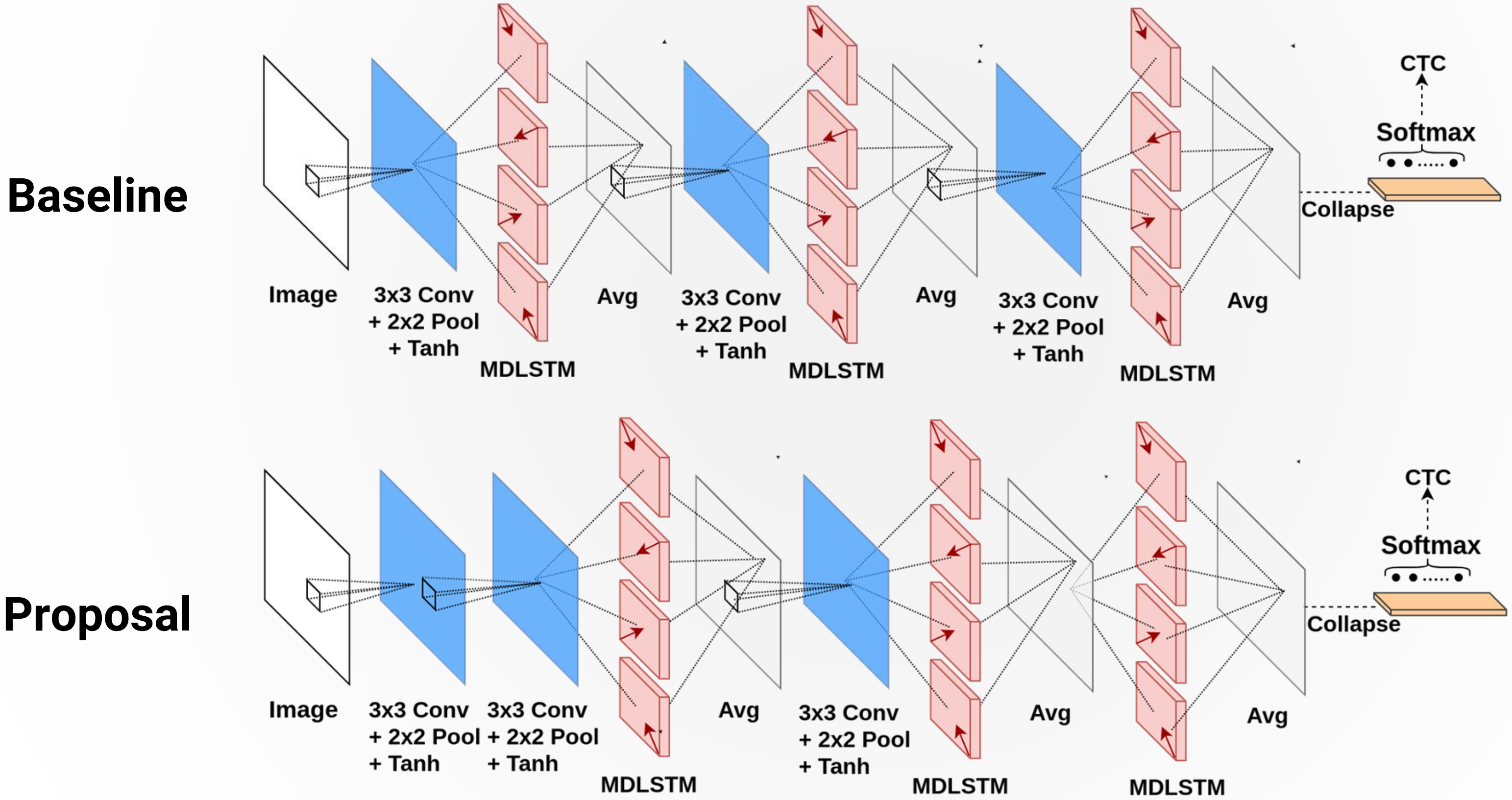
- New hierarchical representations for a MDLSTM optical model
- Speed-ups the training and inference time at the hierarchical-level

Proposal and hypothesis

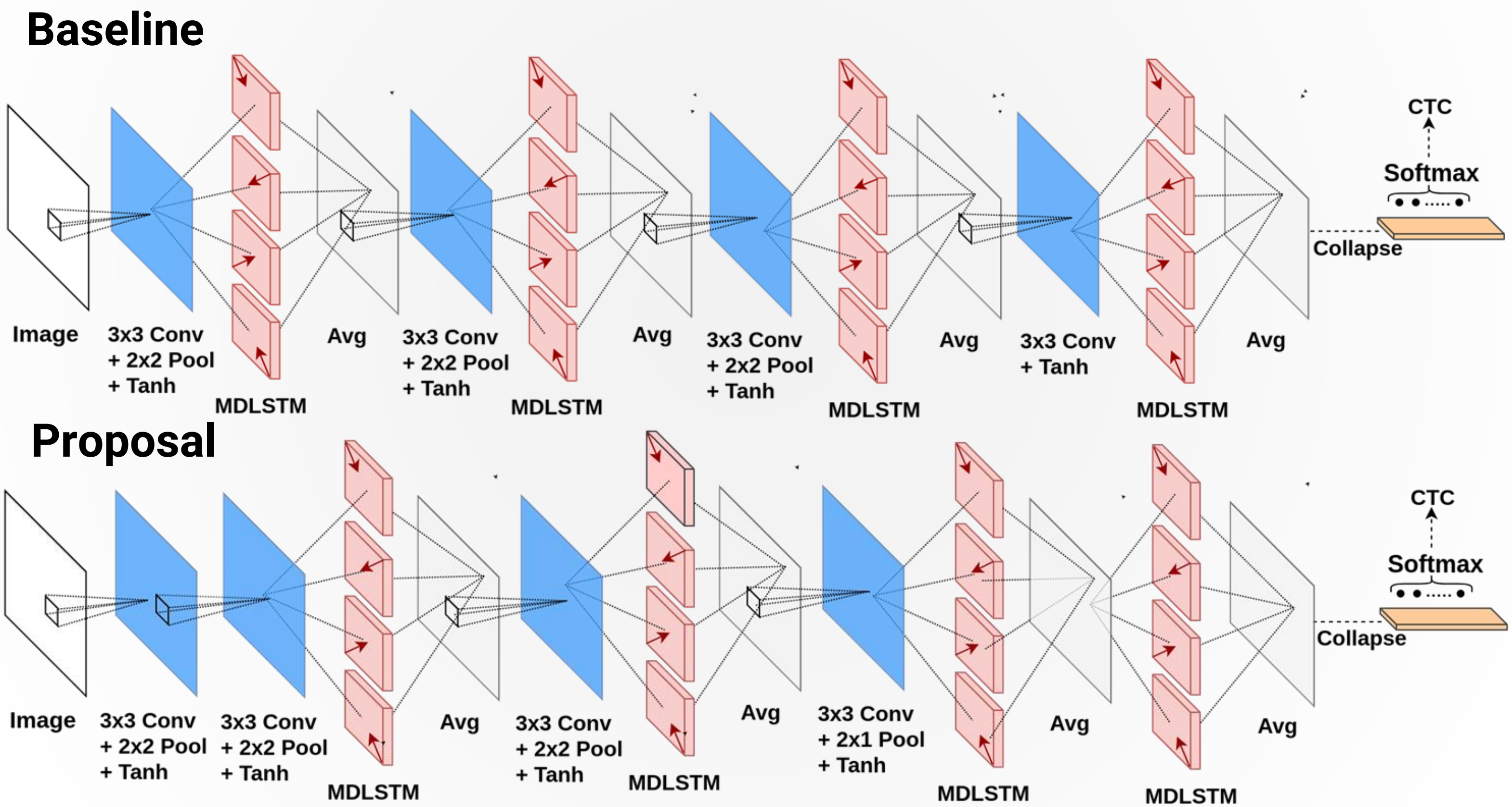
1. Repositioning convolutional and recurrent aspects of the state-of-the-art MDLSTM Voigtlaender model may be useful to discard low-frequency features and send to the MDLSTM layers a richer representation of the input data

2. Adding an extra max pooling to decrease computational time and improve the invariance to small shifts and distortions

Optical Model (six hidden layers)



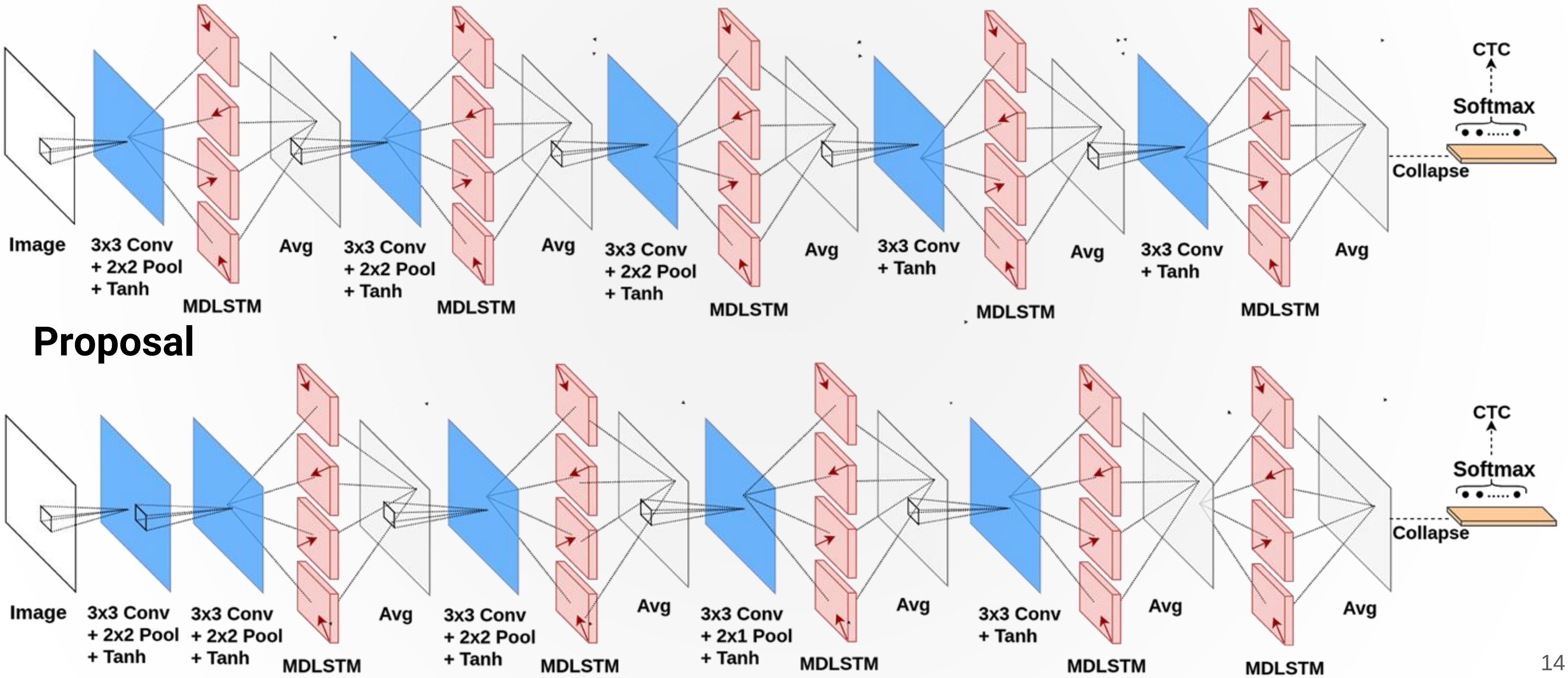
Optical Model (eight hidden layers)



Optical Model (ten hidden layers)

Baseline

Proposal



Experiments

- ❖ Evaluating the MDLSTM optical model
- ❖ Including Linguistic Knowledge
- ❖ Comparison with the state-of-the-art

Experiments

Dataset detailed information

Dataset	Language	Partition: Training	Partition: Validation	Partition: Test	# Symbols	Train. Width (Avg)	Train. Height (Avg)
IAM	English	6.161 (747)	976 (116)	2.781 (336)	79	1.751	124
RIMES	French	10.203 (1351)	1.130 (149)	778 (100)	99	1.658	113

Network Training

Tool: RETURNN

Batch size: 600.000 pixels

Weight Initialization: Glorot or Xavier Initialization

Gradient Descent: Nadam optimizer

Learning Rates Schedule: 0.0005 (1-24), 0.0003 (25-34), 0.0001 (35-Early Stopping)

Training Duration: Early Stopping with patience=20

Optimizing network topologies on the IAM dataset

C = single conv. layer
LP = conv with pooling followed by MDLSTM
L = conv without pooling followed by MDLSTM
M = single MDLSTM Layer

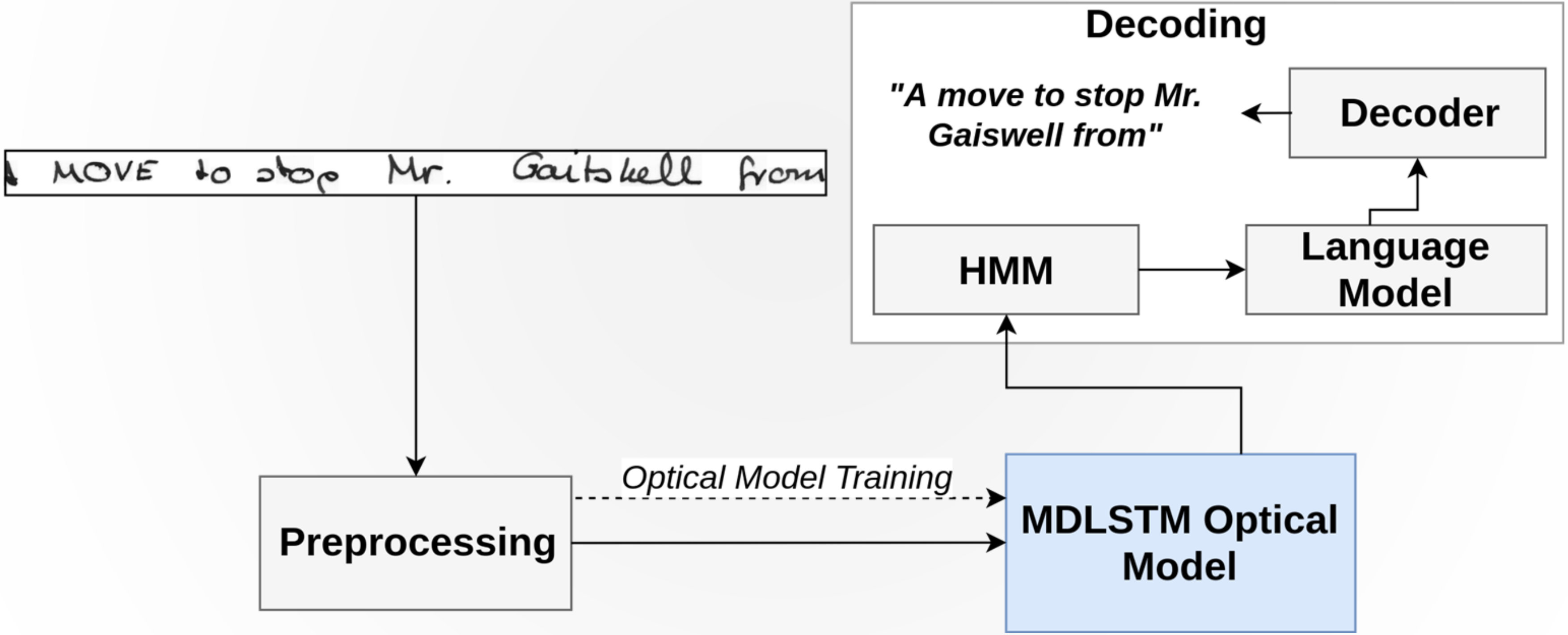
#ID	Architecture	Hidden Layers	Width	Params	Epoch	WER (%)		CER (%)		Train. Time	Valid. Time	Test. Time
						Val.	Test.	Val.	Test.			
01	C-LP-LP-M	6	$15n$	922.070	76	19.58	25.04	4.71	7.1	31.4	1.5	7.5
02			$20n$	1.634.000	37	19.37	24.54	4.66	6.9	39.0	2.0	9.3
03			$25n$	2.548.230	52	18.71	23.77	4.57	6.7	44.3	2.5	11.5
04	LP-LP-LP		$15n$	765.800	35	21.89	27.99	5.35	7.85	81.1	2.9	13.9
05	C-LP-LP-LP-M	8	$15n$	1.987.010	86	18.65	24.35	4.53	6.91	34.8	1.7	7.9
06			$20n$	3.524.920	46	18.7	23.87	4.51	6.67	44.5	2.1	10.2
07			$25n$	5.500.630	55	**17.71**	**22.82**	**4.31**	**6.39**	49.6	2.7	12.8
08	LP-LP-LP-L		$15n$	1.683.755	46	19.99	25.26	4.75	7.06	89.4	3.4	16.1
09	C-LP-LP-LP-LP-M	10	$15n$	3.636.230	57	19.32	24.47	4.79	6.99	41.6	1.9	9.3
10			$20n$	6.454.280	81	18.42	23.04	4.52	6.51	53.9	2.6	12.0
11			$25n$	10.075.330	67	18.83	23.38	4.56	6.64	61.1	3.3	15.3
12	LP-LP-LP-L-L		$15n$	2.627.660	40	18.65	24	4.42	6.64	96.2	3.8	17.9

Experimental Results

Summary

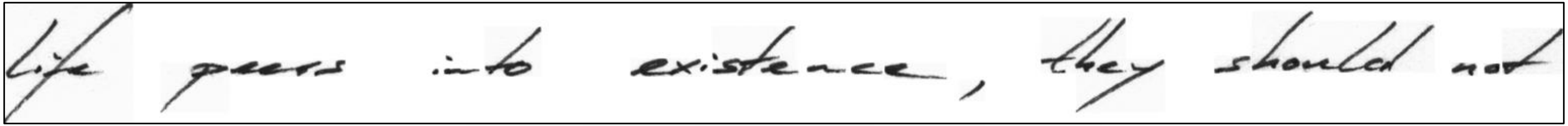
- The modifications did not hurt the recognition performance (hypothesis test confirmed this results)
- Faster model
  - Reduction of roughly 50% and 30% in training and classification times respectively.
- Optimal configuration obtained with eight-layers while the baseline presents ten-layers.
- The proposal presents generalization benefits on larger models.

The complete HTR system

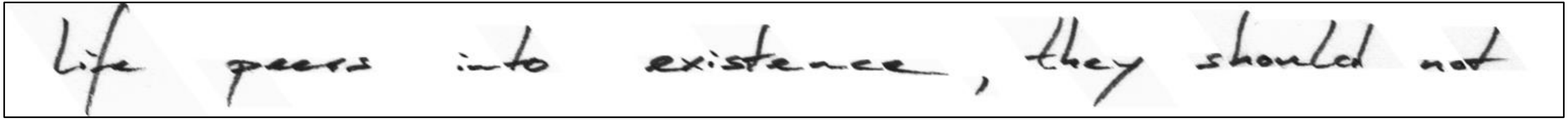


Preprocessing

- **No preprocessing**



- **Dislanting**



- **Inversion of pixel values**

Linguistic knowledge-based decoding

Hybrid ANN/HMM scheme

Finite-state transducers (FST):

- **HMM transducers** (H): each character is represented by an HMM.
- **Lexicon FST** (L): maps a sequence of characters to a valid word.
- **Grammar FST** (G): represents the *n*-gram language model on computing the probability of word sequences.

Compose the H, L, and, G in a decoding graph and search for the most likely transcription using a beam search algorithm.

Language Model Experimental Setup

- **Tool:** SRILM
- **Language model:** 3-gram language model
- **Smoothing technique**: modified Kneser-Ney
- **Text source:** Brown, LOB, and Wellington corpus.
- **Vocabulary**: 50.000 words
- **Perplexity and OOV on the valid set**: 270 (3.1% OOV)
- **Perplexity and OOV rate on test set**: 304 (2.9% OOV)

Decoding Experimental Setup

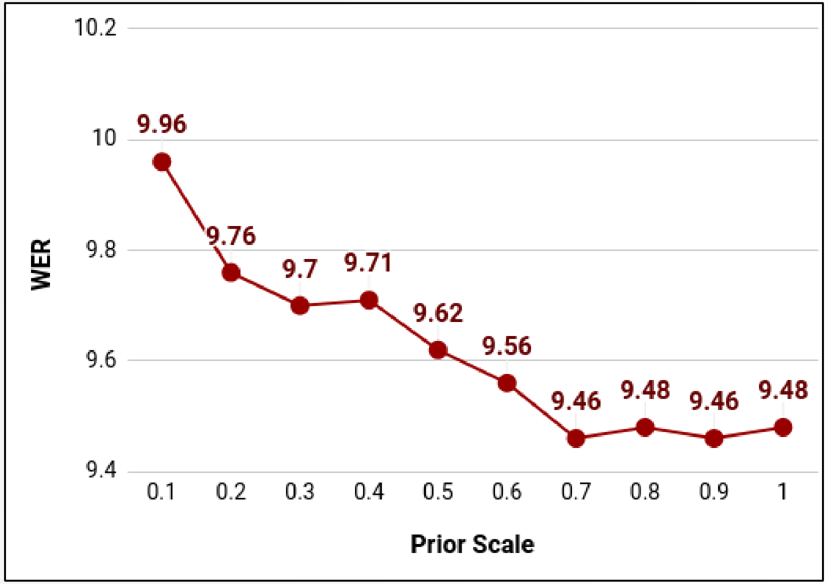
- **Decoders**:
    - Best path decoding for tuning the network topology
    - Linguistic knowledge-based decoding for final results
        - The HMM, lexicon, and language models are represented as Finite-state transducers (FST)
            - Tool: Kaldi toolkit

Experimental Results

Including Linguistic Knowledge - Prior scale tuning

Optical scale fixed at 1.0

Optimal value: 0.7

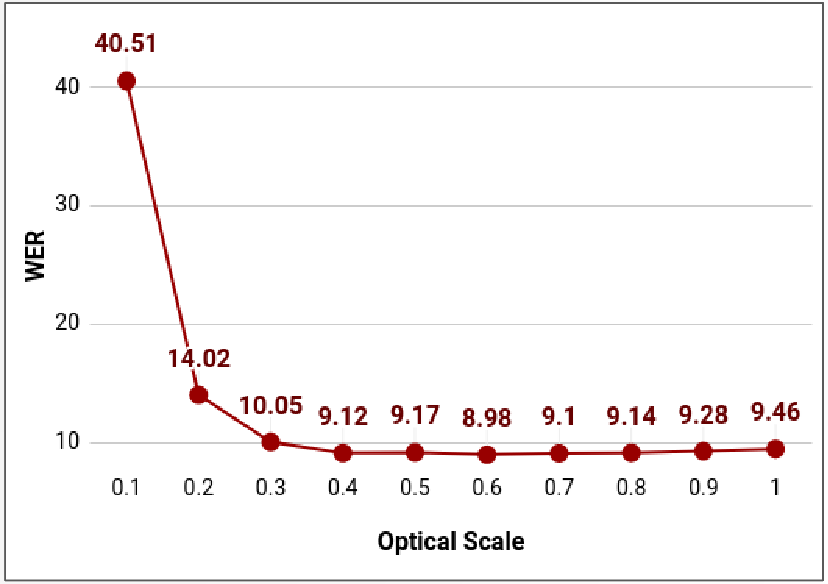


Experimental Results

Including Linguistic Knowledge - Optical scale tuning

Prior scale fixed at 0.7

Optimal value: 0.6

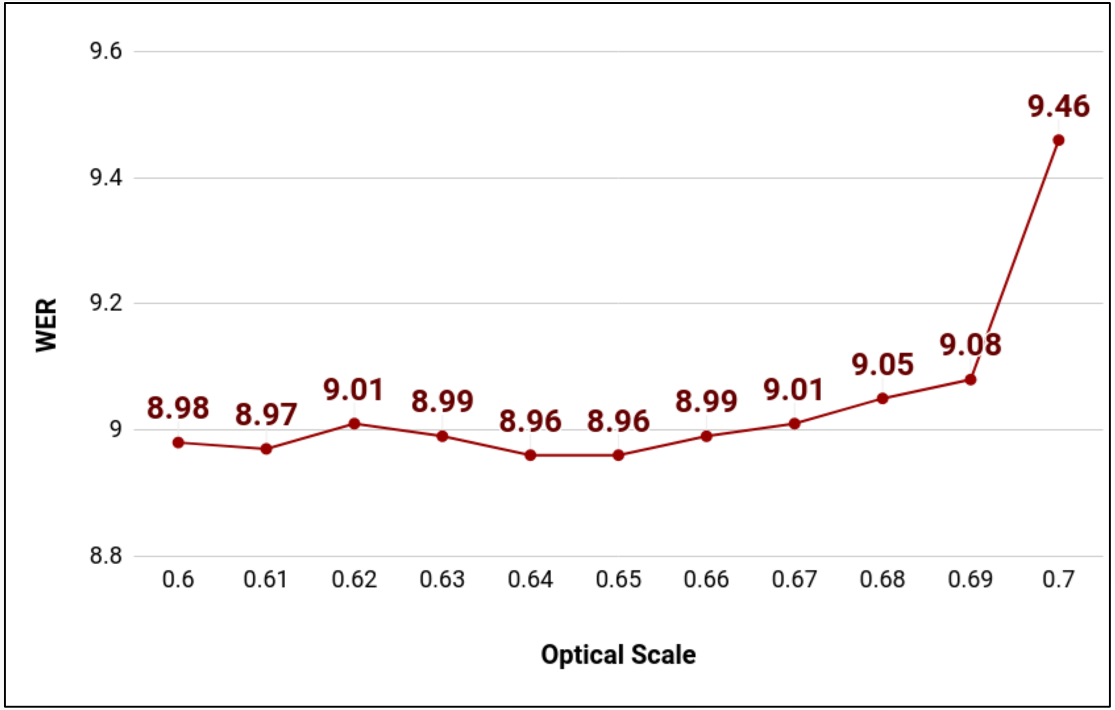


Experimental Results

Second fine-tuning for the optical scale

Prior scale fixed at 0.70

Optimal result: 0.65



Experimental Results

Including Linguistic Knowledge

Decoding Param.		WER (%)		CER (%)	
Prior Scale	Optical Scale	Valid.	Test.	Valid.	Test.
0.7	0.65	8.96	10.52	2.57	3.58

Baseline system (without Ling. Know.) 24 6.64

Comparison with the state-of-the-art - IAM

System	Vocabulary Type	WER (%)		CER (%)	
		Valid.	Test	Valid.	Test
Voigtlaender et al. [23]	Open	**7.1**	**9.3**	**2.4**	3.5
Bluche et al. [39]	Open	-	10.5	-	**3.2**
Proposed System	Closed	9.0	10.5	2.6	3.6
Bluche et al. [25]	Closed	9.6	10.9	3.3	4.4
Voigtlaender et al. [23]	Closed	10.1	11.7	-	-
Puigcerver [40]	Closed	9.2	12.2	2.9	4.4
Doetsch et al. [54]	Open	8.4	12.2	2.5	4.7
Voigtlaender et al. [90]	Open	8.7	12.7	2.6	4.8
Kozielski et al. [26]	Open	9.5	13.3	2.7	5.1
Kozielski et al. [26]	Closed	11.9	-	3.2	-
Pham et al. [22]	Closed	11.2	13.6	3.7	5.1

Brand new results

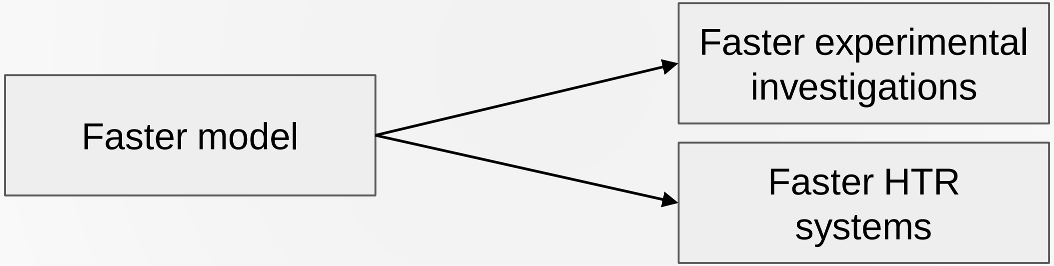
- According with the published results of the ICFHR2018 Competition on Automated Text Recognition on a READ Dataset, our approach achieved the best rate when using only the general dataset provided in the first round of this competition!!!
- We have verified our proposed optical model architecture outperforms the baseline system in the Rimes dataset with a confidence level of 95%.

Optical Model	WER (%) Valid.	WER (%) Test	CER (%) Valid.	CER (%) Test	Train. Time	Valid. Time	Test Time
Proposed	**11.69** [10.39 – 13.09]	**13.21** [11.36 – 15.17]	**2.43** [2.07 – 2.81]	**2.89** [2.33 – 3.53]	76.15	3.37	3.37
Baseline	13.53 [12.06 – 15.07]	15.11 [13.01 – 17.21]	2.76 [2.39 – 3.16]	3.16 [2.57 – 3.81]	150.48	4.65	4.75

Conclusion

Main Contributions

- New MDLSTM hierarchical representation able to reduce the training and classification times without affecting the recognition quality.



- Important tradeoff information between the depth and width of the proposed MDLSTM model.
- Evaluation of the MDLSTM variant in a hybrid ANN/HMM scheme with linguistic knowledge.

Future Works

- Apply the convolutional layer repositioning strategy with the (1D,B)LSTM HTR system, taking advantage of the recent results presented by Puigcerver et al. (2017) in ICDAR.
- Explore the Open-vocabulary scenario
- Evaluate the model with data augmentation

Boosting the deep multidimensional long short-term memory network for handwritten recognition systems

Prof. Byron L. D. Bezerra
byronleite@ecomp.poli.br