Réseaux récurrents bi-directionnels pour la reconnaissance de l'écriture

Laurence Likforman-Sulem

Département Traitement du Signal et des Images – TSI

http://perso.telecom-paristech.fr/~lauli/

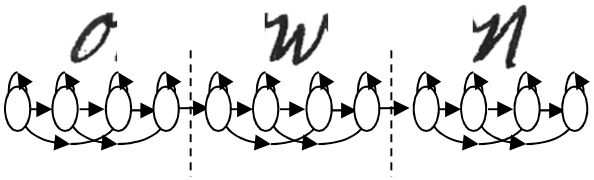
Septembre 2014

Overview

- Introduction
  - handwriting recognition
  - sliding window approaches
- Recurrent Neural Networks
  - RNNs
  - Bi-directional RNNs
  - LSTM nets
    - RNN-LSTMs
    - BLSTMs
  - Connectionist Temporal Classification: CTC
- Conclusions

Reconnaissance d'écriture

Noms de ville
Noms de rue



Lignes de texte

Approches stochastiques
(modèles Markov Cachés)
réseaux neurones récurrents

Mots isolés

Approches stochastiques
(modèles Markov Cachés)
hybrides Markov/MLP

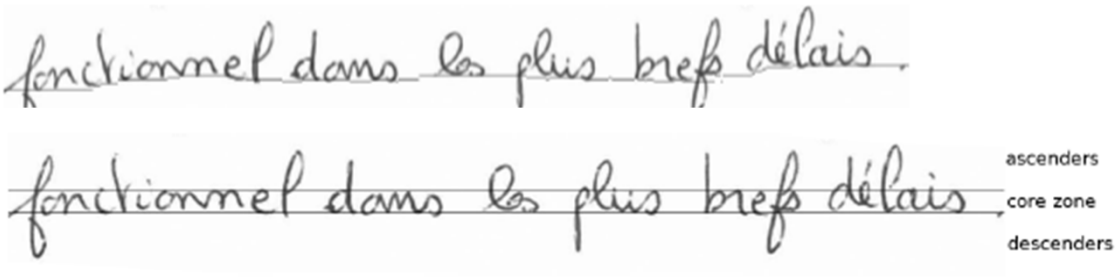
Caractères isolés

Réseaux neurones
template matching

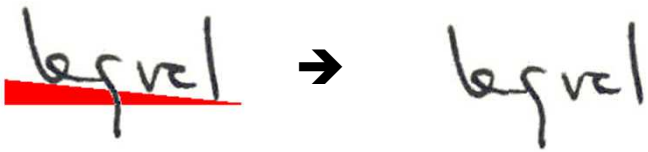
1990 → 2000 → 2009

Word/Text-line recognition with HMMs or RNNs

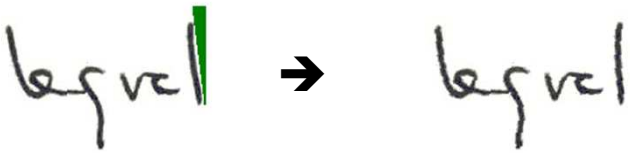
preprocessing: background noise removal, baseline correction, deslant



deskew

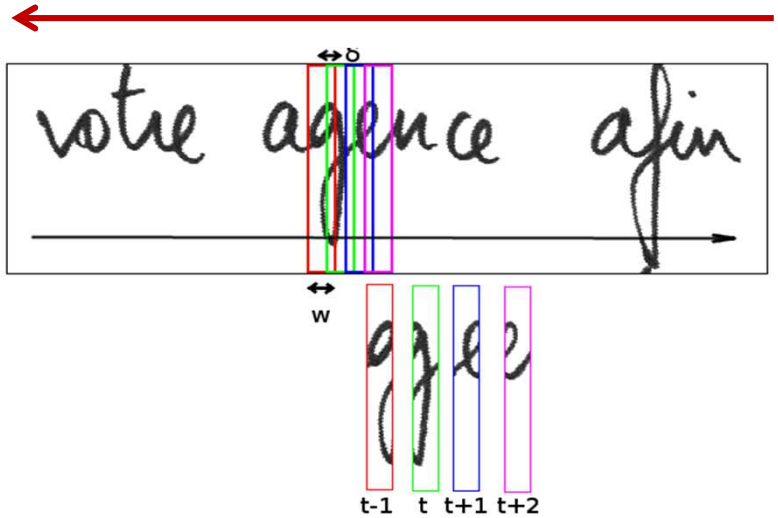


deslant



Word/Text-line recognition with RNNs/HMMs

- feature extraction: sliding window left-right (+ right-left for RNNs)

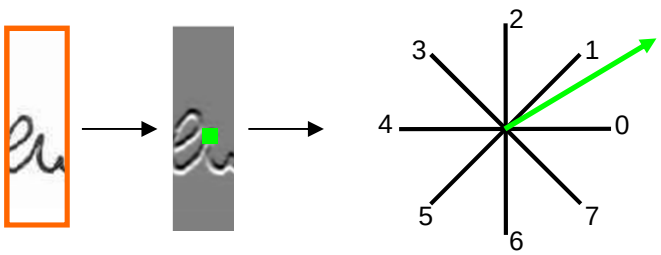


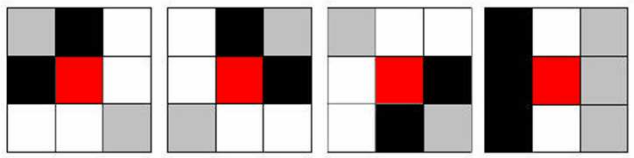
- concavity configurations, pixel distribution, B/W transitions, histograms of gradient (HoG): 8 directions
- pixel values

Word/Text-line recognition with RNNs/HMMs

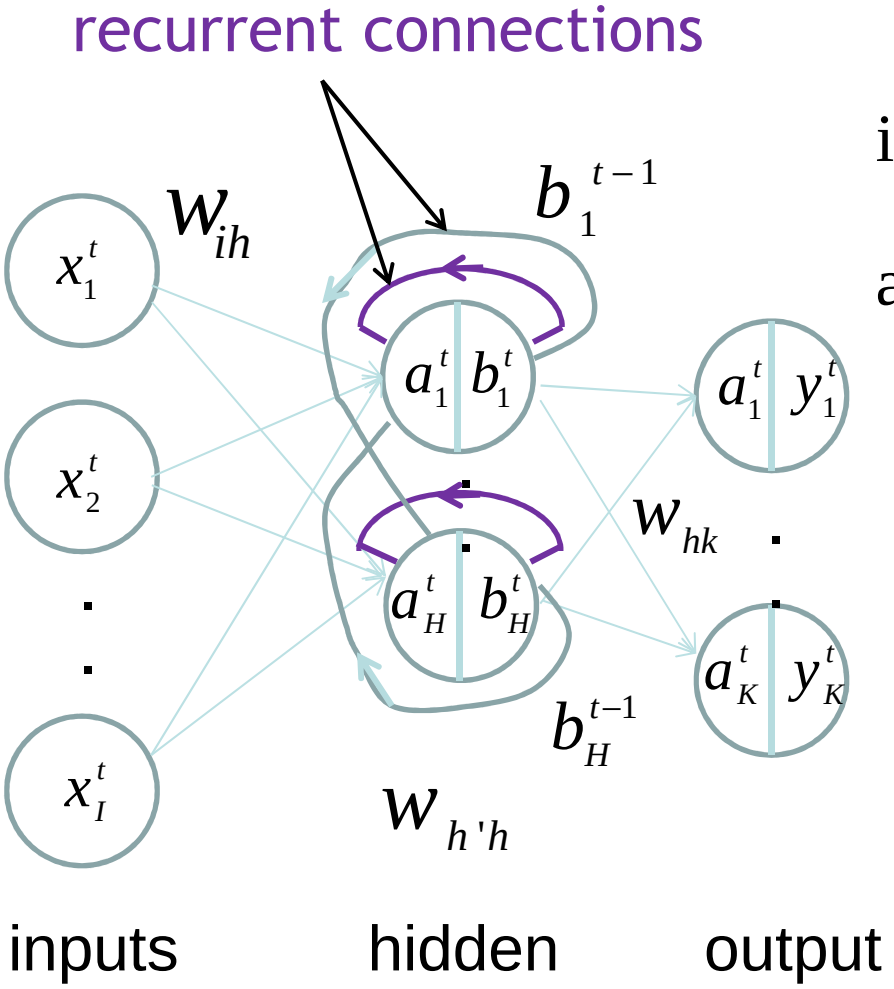
- feature extraction
  - concavity configurations, pixel distribution, B/W transitions, histograms of gradient (HoG): 8 directions
  - pixel values

o





RNN recurrent neural network



hidden unit h:

$$\text{input} : a_h^t = \sum_{i=1}^{I} w_{ih} x_i^t + \sum_{h'=1}^{H} w_{h'h} b_{h'}^{t-1}$$

$$\text{activation} : b_h^t = \theta(a_h^t)$$

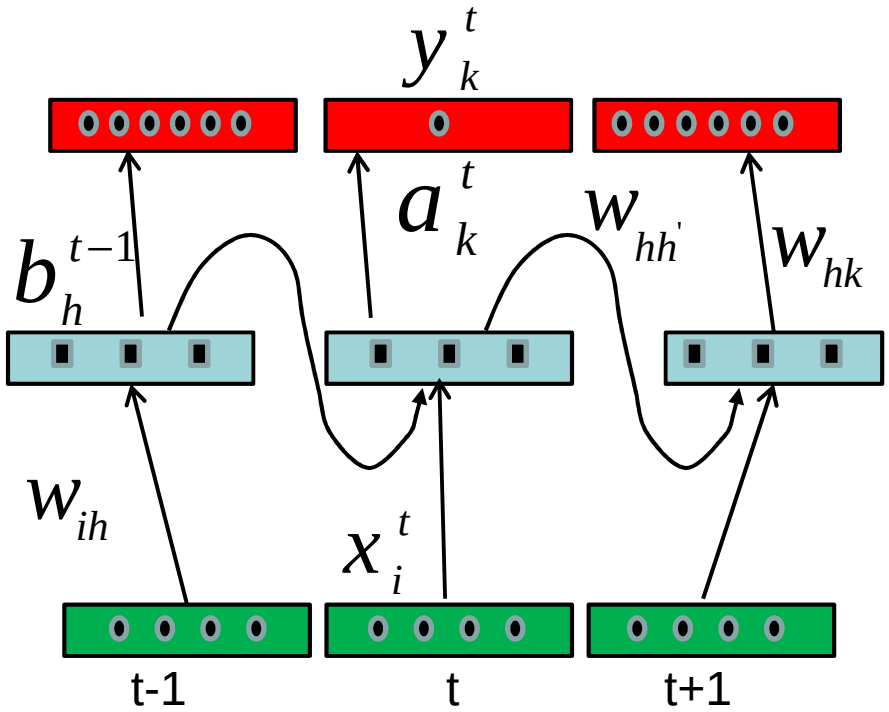
output unit k:

$$\text{input} : a_k^t = \sum_{h=1}^{H} w_{hk} b_h^t$$

$$\text{activation} : y_k^t = \theta(a_k^t)$$

θ: activation function

unrolled RNN: forward pass



$$a_k^t = \sum_{h=1}^{H} w_{hk} b_h^t$$

$$y_k^t = \theta(a_k^t)$$

$$\text{input}: a_h^t = \sum_{i=1}^{I} w_{ih} x_i^t + \sum_{h'=1}^{H} w_{h'h} b_{h'}^{t-1}$$

$$\text{activation}: b_h^t = \theta(a_h^t)$$

b_h^t activation of hidden unit h

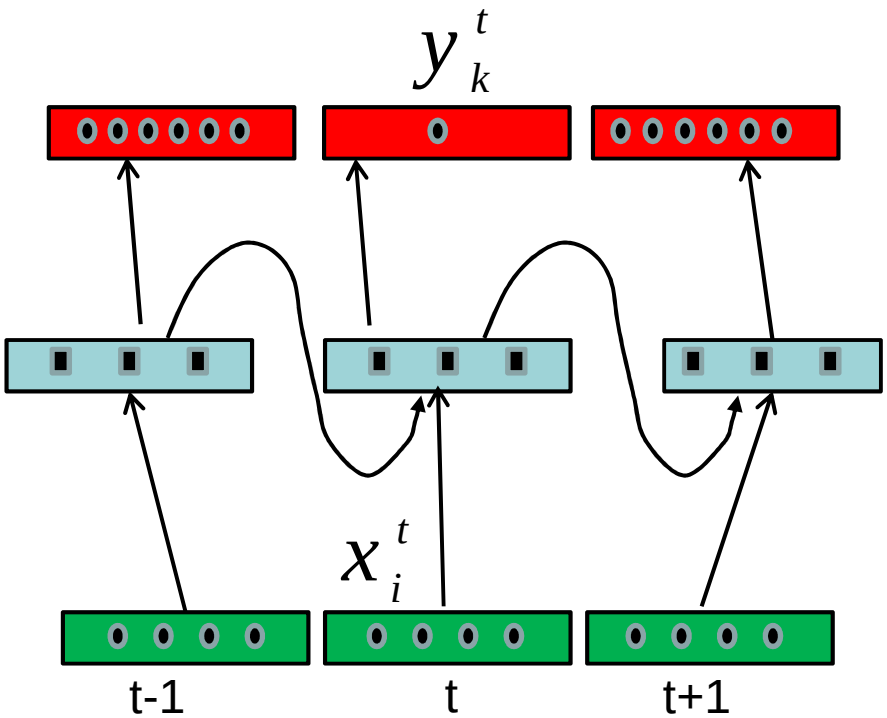
a_k^t input of output unit k

y_k^t activation of output unit k

8

RNN: unrolled

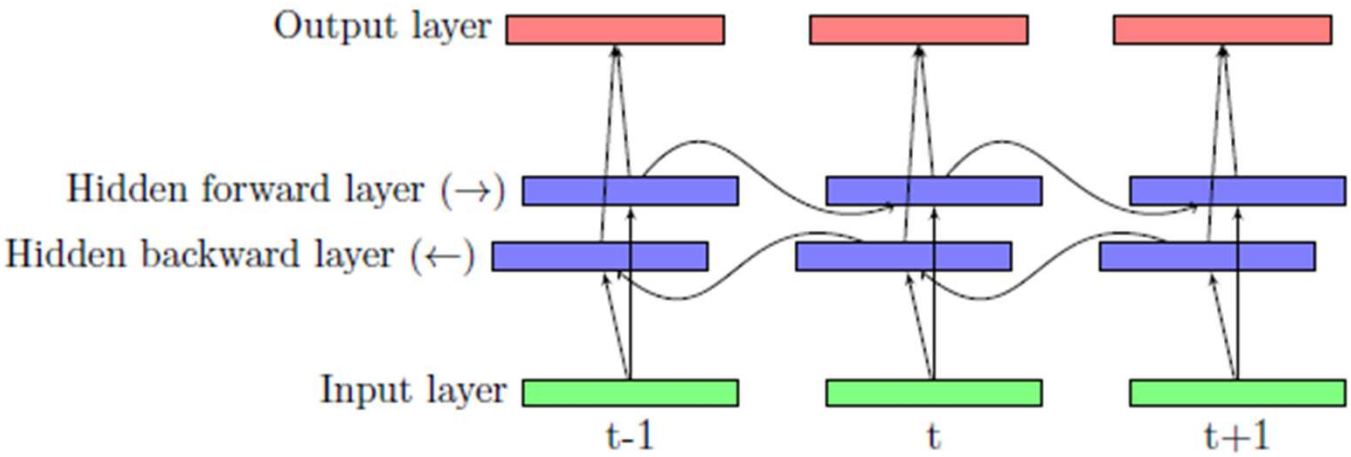
- discriminative approach: highest output wins

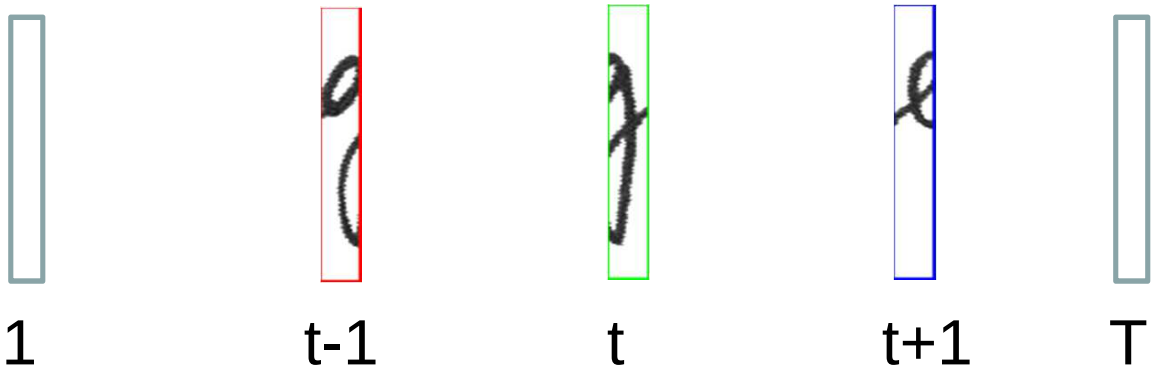


91 outputs units:
A-Z, a-z, 0-9, symbols, blank

BRNNs

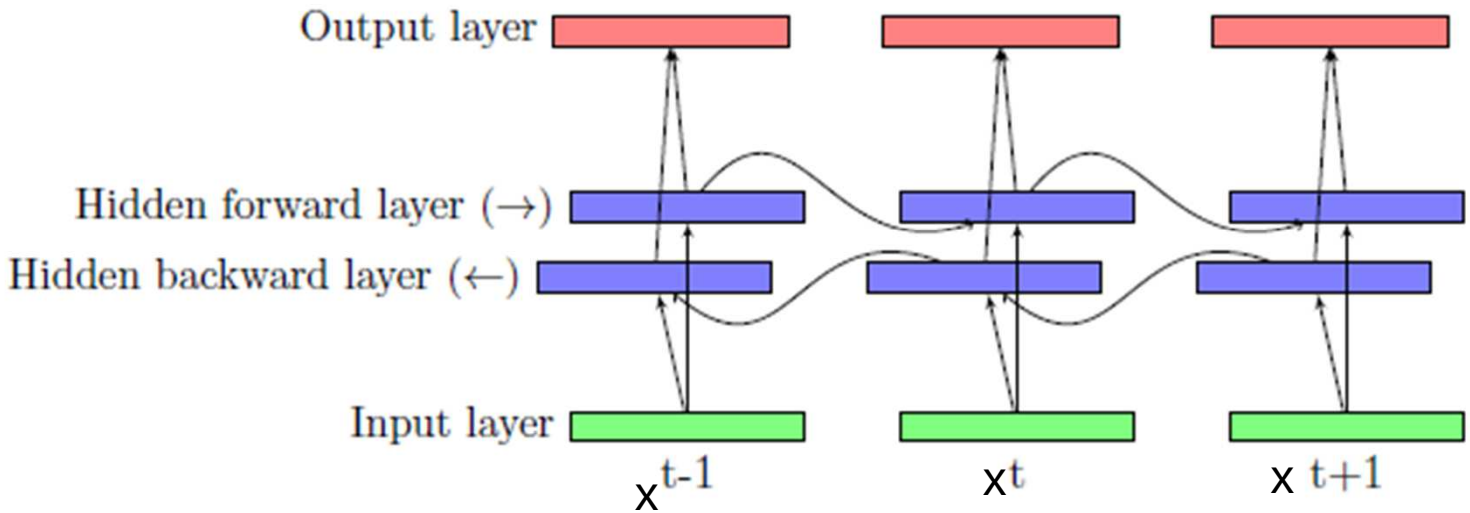
- **bi-directional recurrent neural networks [Schuster and Paliwal, 97]**



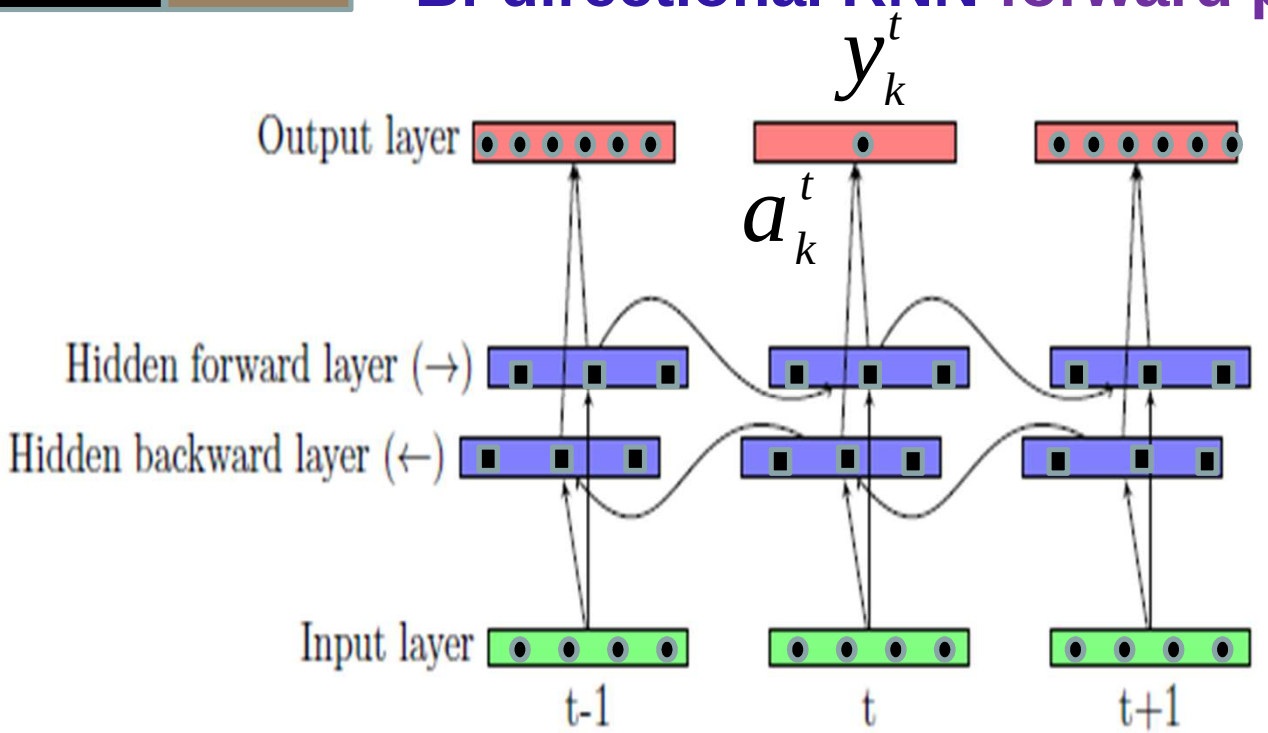


Bi-directional RNN

- coupling of two RNN networks :
  - long-term dependencies taken into account (past + future)
  - share input and output layers
  - forward network: $x^1 \ldots x^T$, backward network: $x^T \ldots x^1$



Bi-directional RNN forward pass



forward layer $\longrightarrow$ **backward layer** $\longleftarrow$

$$\text{input} \quad : a_k^t = \sum_{h=1}^{H} w_{hk} b_h^t + \sum_{l=1}^{H} w_{lk} b_l^t$$

$$\text{activation} \quad : \quad y_k^t = \theta(a_k^t)$$

b_h^t activation of hidden unit h

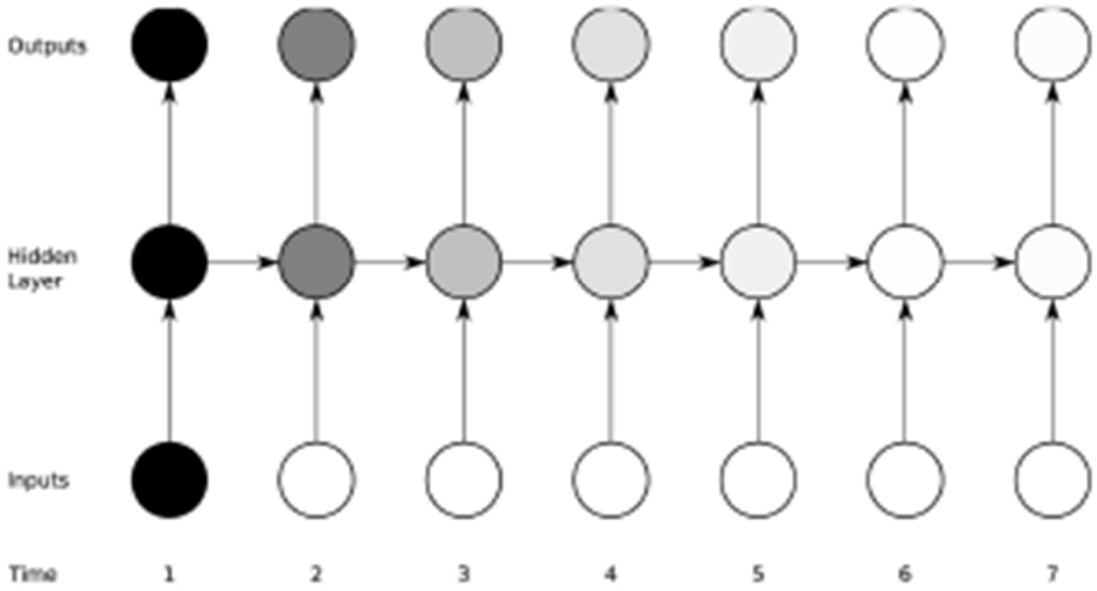
a_k^t input of output unit k

y_k^t activation of output unit k

12

Vanishing Gradient

from [Graves et al. 2009]



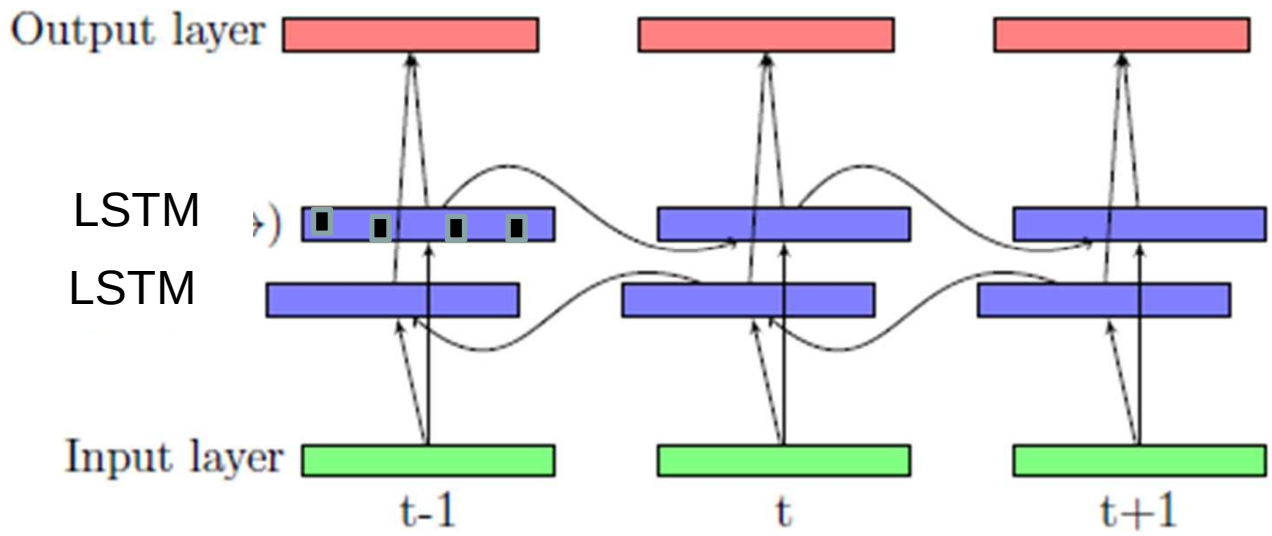
RNN: information vanishes quickly

→LSTM: Long Short -Term Memory blocks in hidden layer
[Gers, Hochreiter, Schmidhuber 2001; 1997]

13

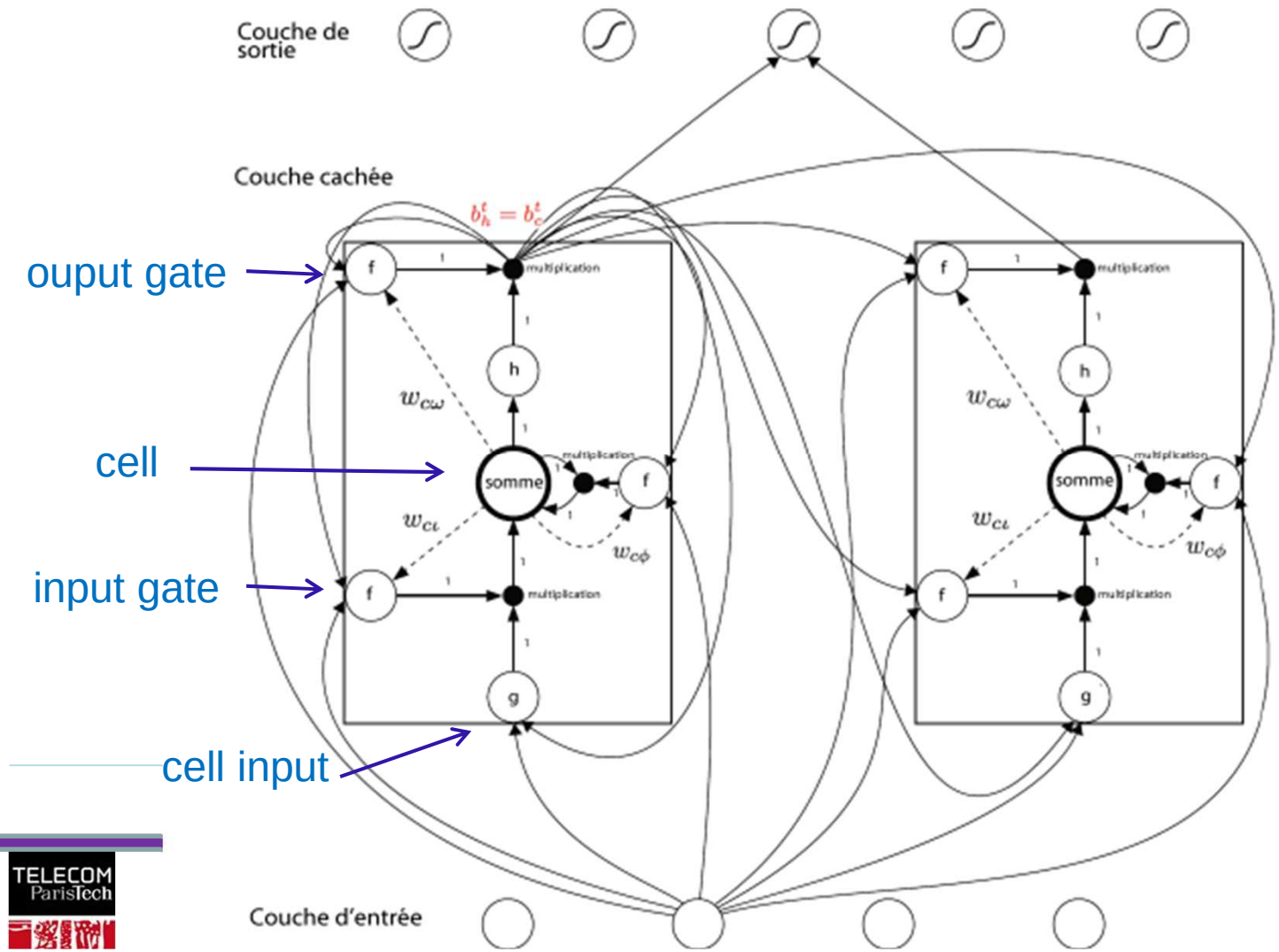
BLSTMs

- Bi-directional Long Short Term Memory
  - memory blocks LSTMs in forward and backward hidden layers



LSTM block in Hidden layer

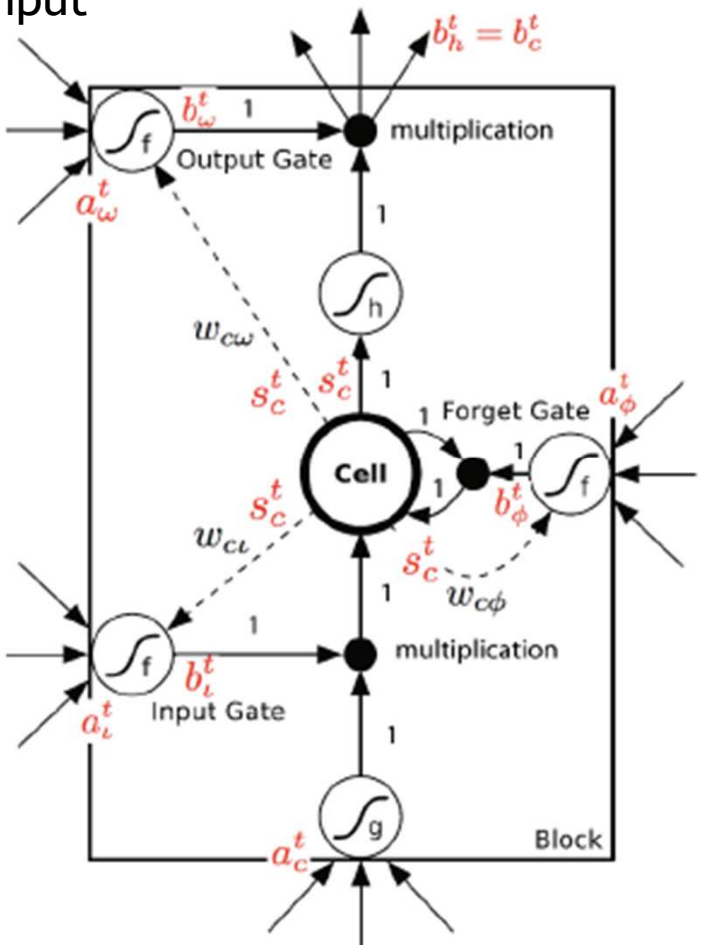
4 block entries: cell input and 3 gates



inputs at t (x_i^t) and hidden layer outputs at t-1 (b_h^{t-1})

LSTM blocks

- « **memory** » **cell** + one recurrent connection (carousel)
  cell state can remain constant even if no input
- **input/output gates : when closed**
  irrelevant inputs and noise do not enter
  and cell state does not influence network
- **forget gate**
  reset cell state (stop cell state increase)
- **peep-hole connections**
  observe cell state



LSTM blocks

input gate

inputs

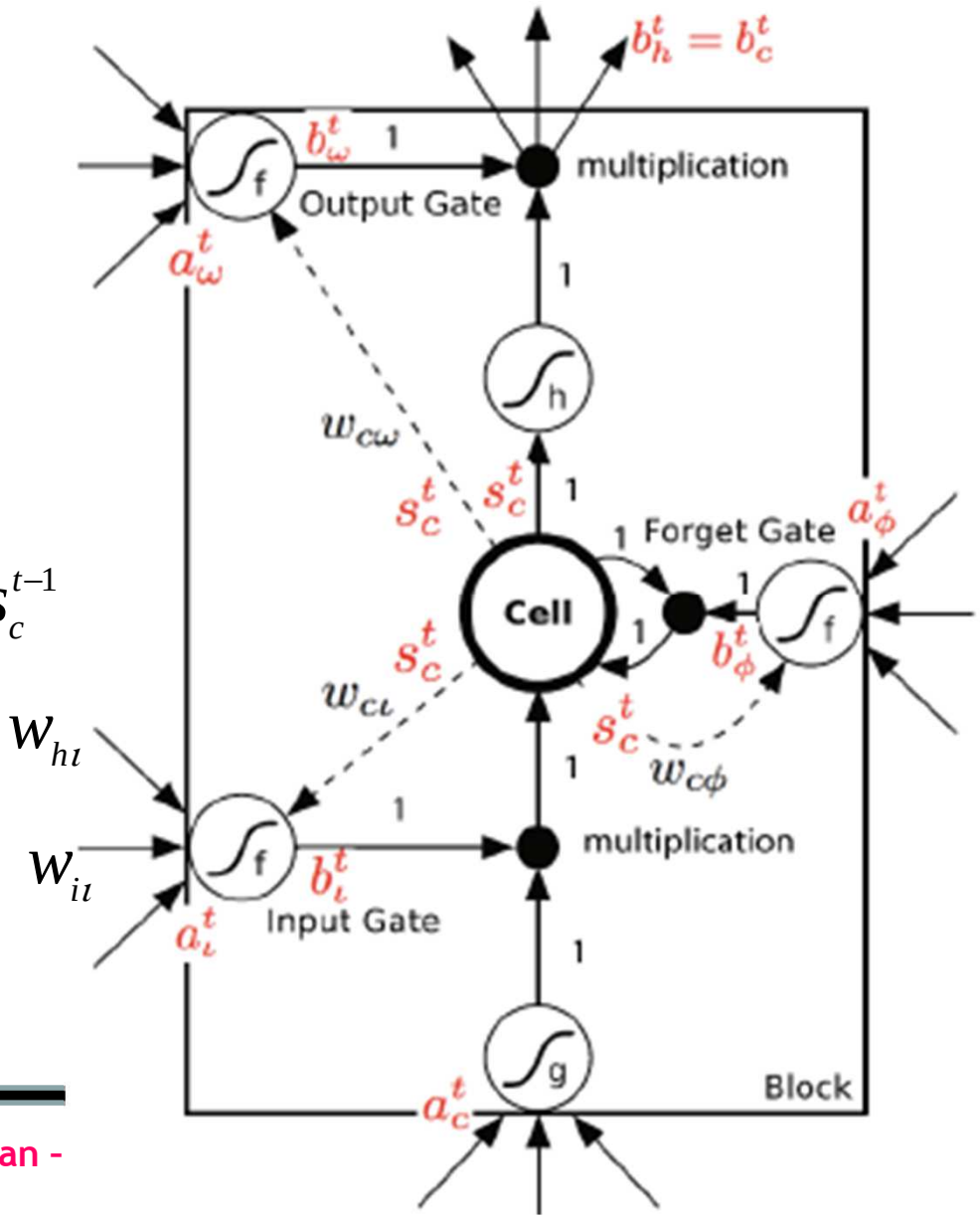
block outputs (t-1)

cell state (t-1)

$$a_\iota^t = \sum_{i=1}^{I} w_{i\iota} x_i^t + \sum_{h=1}^{H} w_{h\iota} b_h^{t-1} + w_{c\iota} s_c^{t-1}$$

$$b_\iota^t = f(a_\iota^t)$$

peephole



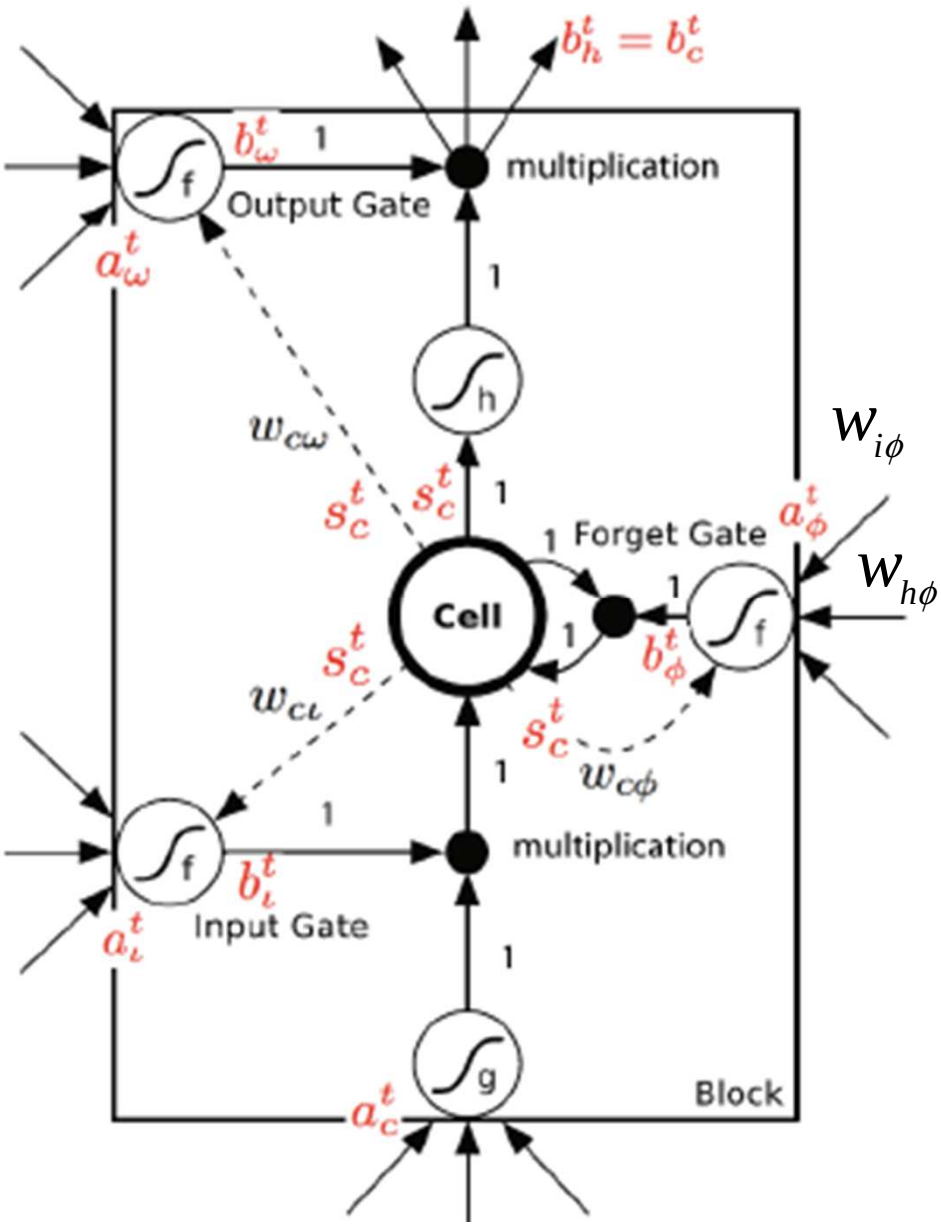
LSTM blocks

forget gate

inputs — block outputs (t-1) — cell state (t-1)

$$a_\phi^t = \sum_{i=1}^{I} w_{i\phi} x_i^t + \sum_{h=1}^{H} w_{h\phi} b_h^{t-1} + w_{c\phi} s_c^{t-1}$$

$$b_\phi^t = f(a_\phi^t)$$



LSTM blocks

block outputs (t-1)

inputs

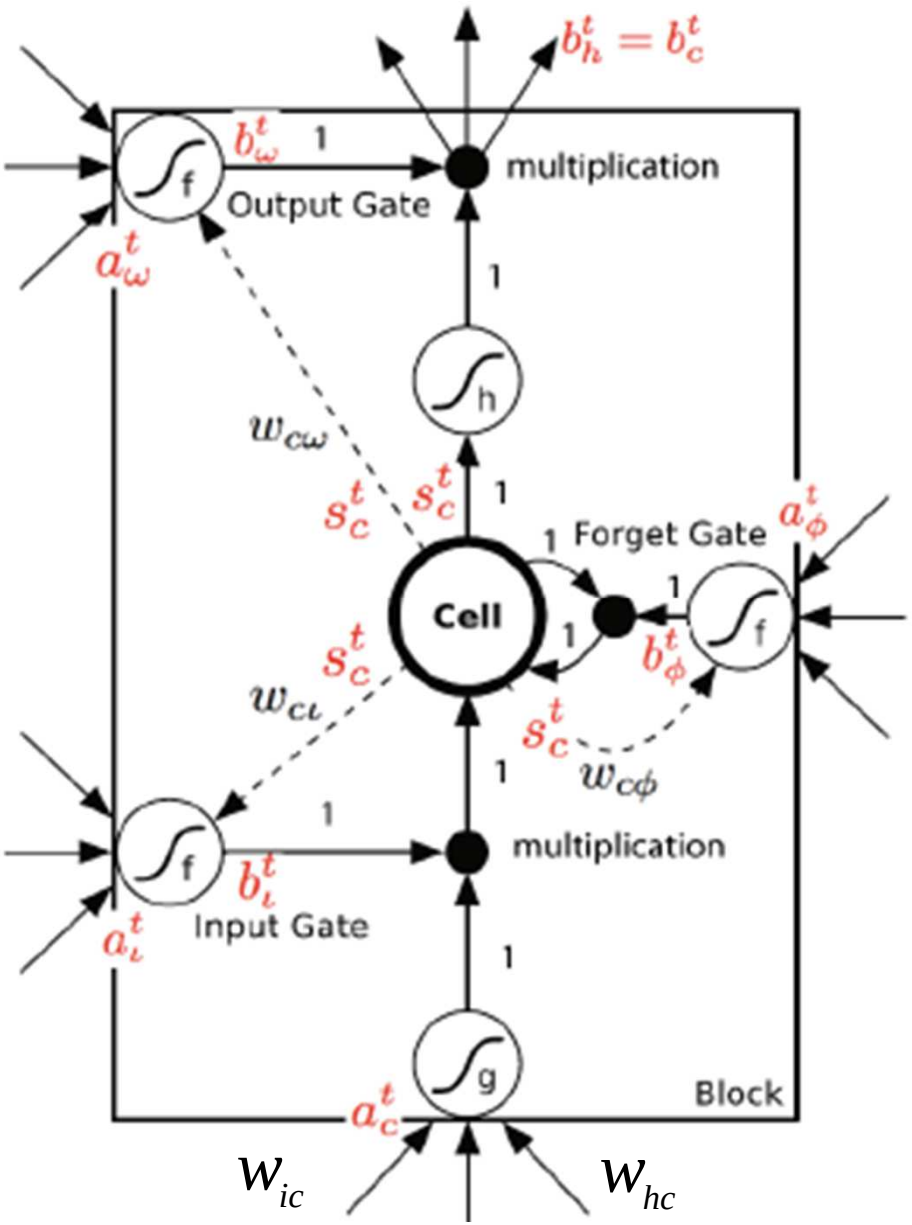
cell

$$\text{cell input}: \; a_c^t = \sum_{i=1}^{I} w_{ic} x_i^t + \sum_{h=1}^{H} w_{hc} b_h^{t-1}$$

$$\text{cell state}: s_c^t = b_\phi^t s_c^{t-1} + b_\iota^t g(a_c^t)$$

recurrent connexion

if 'input gate' closed , cell state does not depend on input data any more



LSTM blocks

output gate

inputs — block outputs (t-1) — cell state (t-1)

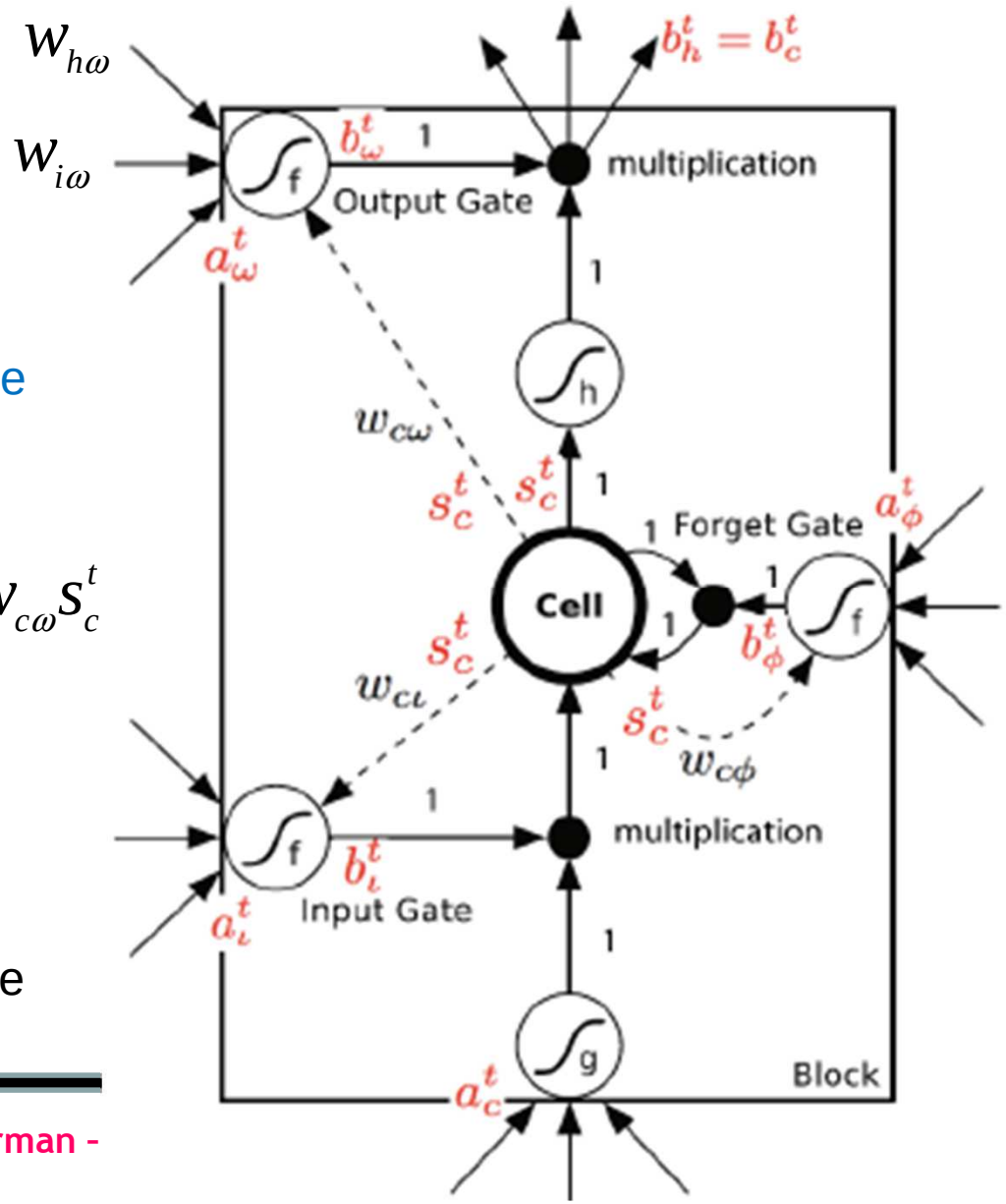
$$a_\omega^t = \sum_{i=1}^{I} w_{i\omega} x_i^t + \sum_{h=1}^{H} w_{h\omega} b_h^{t-1} + w_{c\omega} s_c^t$$

$$b_\omega^t = f(a_\omega^t)$$

cell output

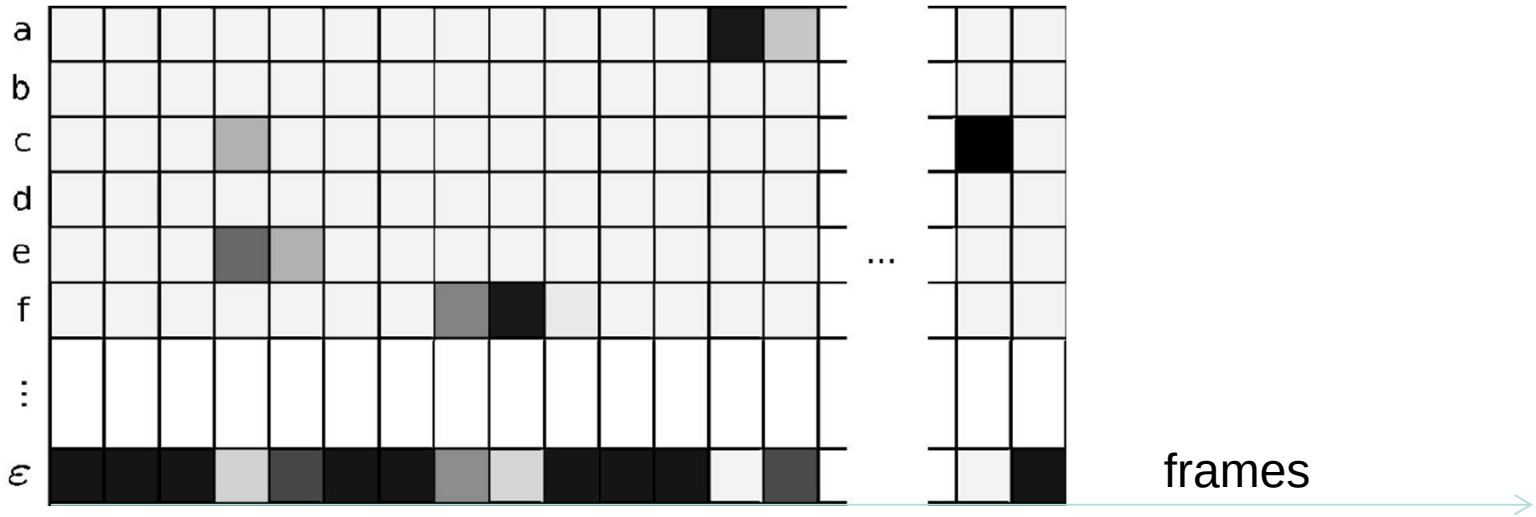
$$b_c^t = b_\omega^t h(s_c^t)$$

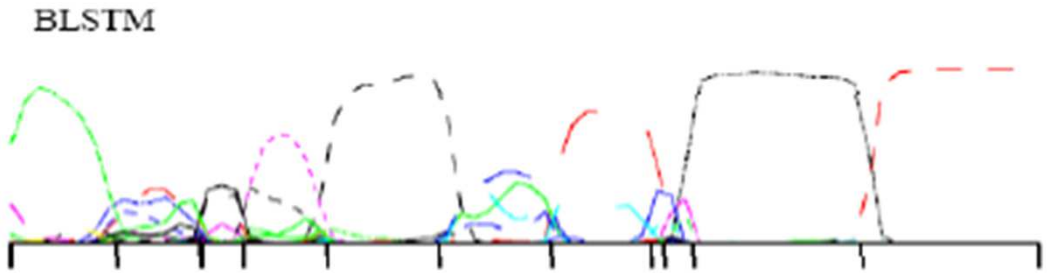
peephole



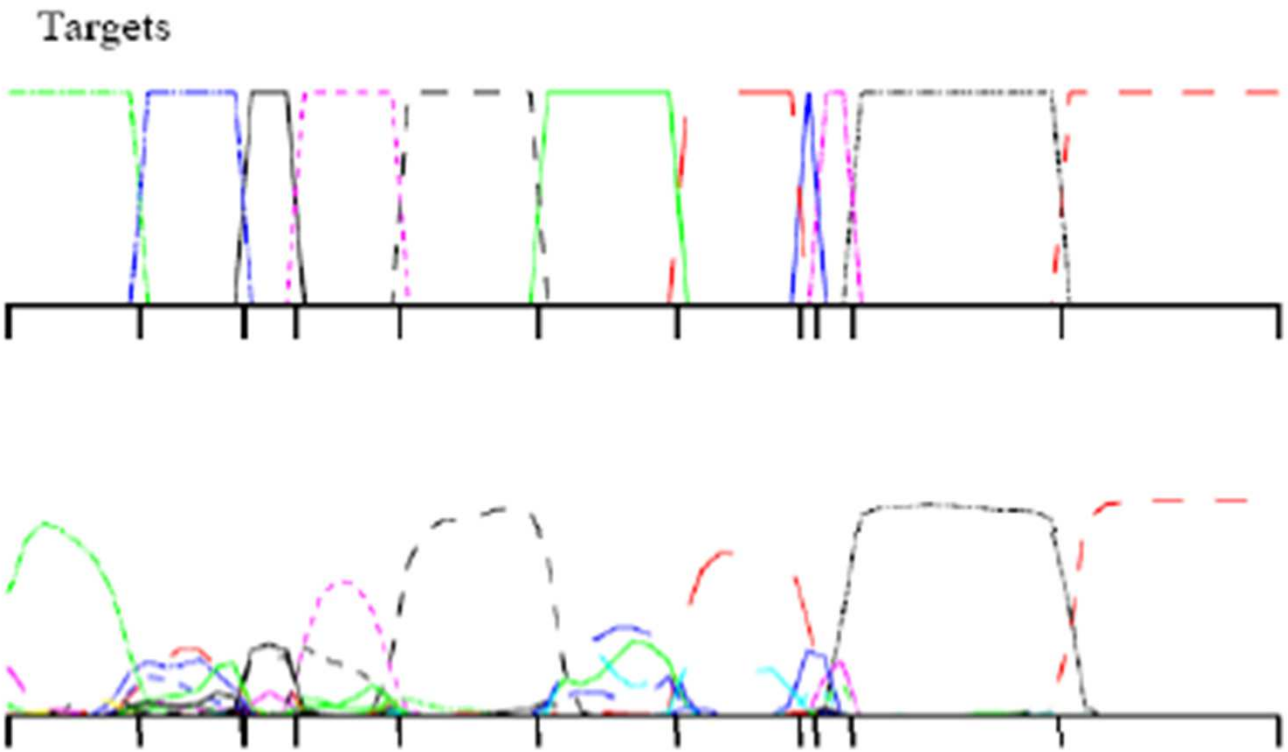
BLSTM Output layer

- **91 output units: A-Z, a-z, 0-9, symbols, ε ='sil'or no label**
- ouput activations: y^t_k probability of observing character k at time t given frame sequence x





training a RNN



training a RNN

- objective function
- sensitivity terms

$$O = -\sum_{(x,z)\in S} \log(\ p(z|x))$$

$$\delta_h^t = \theta'(a_h^t)\left(\sum_{k=1}^{K} \delta_k^t w_{hk} + \sum_{h'=1}^{H} \delta_{h'}^{t+1} w_{hk}\right)$$

when a_h^t (hidden unit)varies, it modifies input of output units at t and input of all hidden units at t+1

delta terms are computed starting at t=T till t=1

- derivative terms (for weight update)

$$\frac{\partial O}{\partial w_{ij}} = \sum_{t=1}^{T} \frac{\partial O}{\partial a_j^t}\frac{\partial O}{\partial w_{ij}} = \sum_{t=1}^{T} \delta_j^t b_i^t$$

wij not modified through t=1, T

training a BRNN

- forward pass
  - forward hidden layer : compute activation terms t=1 to T
  - backward hidden layer : compute activation terms t=T to 1
  - output layer: compute output activation terms

- backward pass
  - output layer: compute δ terms t=T to 1 and update weights
  - forward hidden layer : compute δ terms t=T to 1 and update weights
  - backward hidden layer : compute δ terms t=1 to T and update weights

LSTM backward pass

$$\varepsilon_c^t = \frac{\partial O}{\partial b_c^t} \qquad \varepsilon_s^t = \frac{\partial O}{\partial s_c^t}$$

cell output $\varepsilon_c^t = \left(\sum_{k=1}^{K} \delta_k^t w_{ck} + \sum_{h=1}^{H} \delta_h^{t+1} w_{hc}\right)$

ouput gate $\delta_\omega^t = f'(a_\omega^t)\, h(s_c^t)\, \varepsilon_c^t$

States

$$\epsilon_s^t = b_\omega^t h'(s_c^t)\epsilon_c^t + b_\phi^{t+1}\epsilon_s^{t+1} + w_{c\iota}\delta_\iota^{t+1} + w_{c\phi}\delta_\phi^{t+1} + w_{c\omega}\delta_\omega^t$$

cell $\delta_c^t = b_\iota^t g'(a_c^t)\, \varepsilon_s^t$

error truncation

forget gate $\delta_\phi^t = f'(a_\phi^t)\, s_c^{t-1} \varepsilon_s^t$

input gate $\delta_\iota^t = f'(a_\iota^t)\, g(a_c^t)\, \varepsilon_s^t$

CTC training/decoding

- RNN target: should be frame-based
  - but sequence of labels and frame sequence have different lengths
- CTC: Connexionist Temporal Classification
  - maps frame sequence x (length T) with character sequence z (length L << T) [Graves 2009]
  - with CTC, no frame alignment necessary
- CTC training
  - backward-forward algorithm
- CTC decoding
  - token passing algorithm
  - with bi-grams and dictionary

CTC forward-backward algorithm

- CTC objective function

$$O = -\sum_{(x,z)\in S} \log(\ p(z|x))$$

transcription

frame sequence

- extended label sequence
  - z'=z with blanks inserted

z= **e l l e**

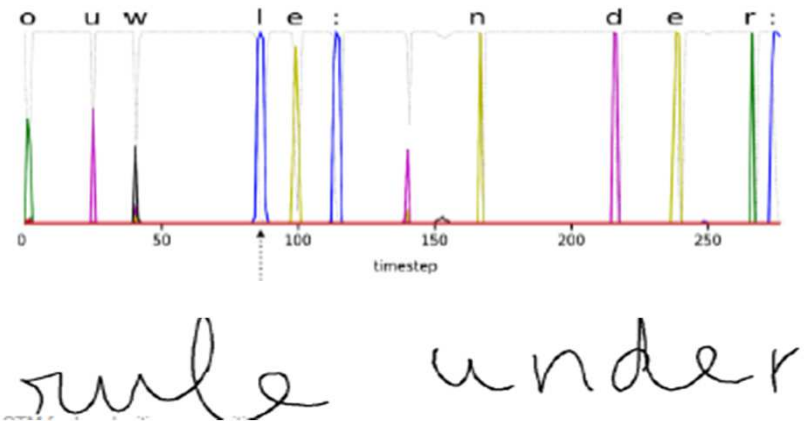
z'= **ε e ε l ε l ε e ε**

- alpha $\alpha(t,s)$
  - sum of probabilities of all paths (length t) arriving on sth symbol of z' at time t
- beta $\beta(t,s)$
  - sum of probabilities of all paths (length T-t) starting on sth symbol of z' at time t

$$p(z|x) = \sum_{s=1}^{|z'|} \alpha(t,s)\beta(t,s)$$

output example

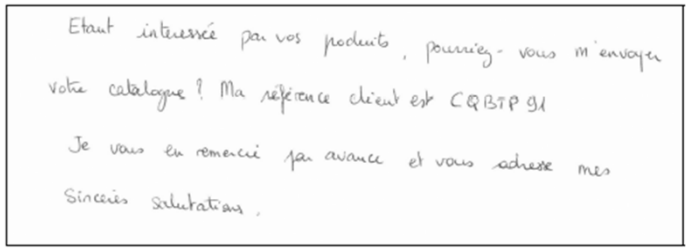
: with CTC, BLSTMs output spikes around characters



recognition example

- **BLSTM recognizer output (dic+LM)**

Etant intéressée par vos produits , pourriez vous m' envoyer
votre catalogue ? Ma référence client est CCP 21
Je vous en remercie par avance et vous adresse mes
Sincères salutations .


Etant intéressée par vos produits, pourriez-vous m'envoyer
votre catalogue ? Ma référence client est CQBTP 91
Je vous en remercie par avance et vous adresse mes
Sincères salutations.

- **BLSTM recognizer output (no dictionary, no langage model !)**

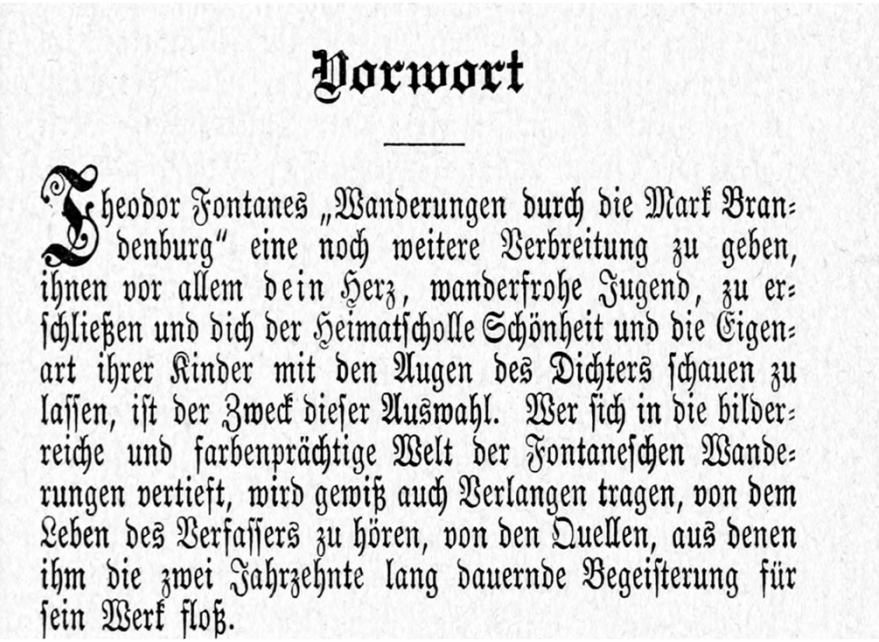
Etant interessée par vos p[r]oduits, pourriez vous m'env[o]yer
votre catal[o]gue l lMMa référence dclient est CRQHRP 91
e vous en remercié par avauce et vous adresse mes
Sinceres ssalutations.

- **HMM recognizer output (dictionary, language model)**

Etant intéressée par vos produits. pourriez-vous m'envoyer
Votre catalogue? Ma référence client est CE site si
`a vous en remercie par avance et vous adresse mes
sincères salutations

Printed character recognition

original image



Vorwort

Theodor Fontanes „Wanderungen durch die Mark Brandenburg“ eine noch weitere Verbreitung zu geben, ihnen vor allem dein Herz, wanderfrohe Jugend, zu erschließen und dich der Heimatscholle Schönheit und die Eigenart ihrer Kinder mit den Augen des Dichters schauen zu lassen, ist der Zweck dieser Auswahl. Wer sich in die bilderreiche und farbenprächtige Welt der Fontaneschen Wanderungen vertieft, wird gewiß auch Verlangen tragen, von dem Leben des Verfassers zu hören, von den Quellen, aus denen ihm die zwei Jahrzehnte lang dauernde Begeisterung für sein Werk floß.

BLSTM output (no language modeling, no dictionary)
training: 20,000 text lines

Porwort
rMeheodor Fontanes,,Wanderungen durch die Mark Bran-
F denburg" eine noch weitere Verbreitung zu geben,
ihnen vor allem dein Herz, wanderfrohe Jugend, zu er-
schließen und dich der Heimatscholle Schönheit und die
Eigen-art ihrer Kinder mit den Augen des Dichters schauen zu
lassen, ist der Zweck dieser Auswahl. Wer sich in die bilder-
reiche und farbenprächtige Welt der Fontaneschen Wande-
rungen vertieft, wird gewiß auch Verlangen tragen, von dem
Leben des Verfassers zu hören, von den Quellen, aus denen
sein Werk floß.

T. Breuel et al, High-Performance OCR for Printed English and Fraktur using LSTM Networks, **ICDAR 2013**

Performance comparison

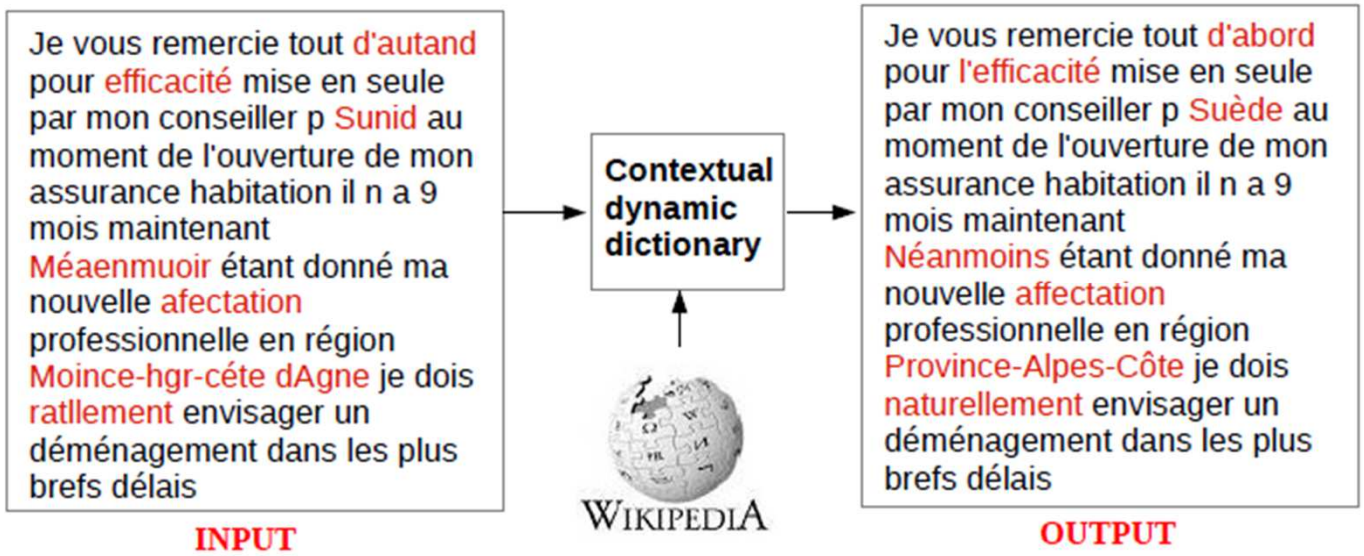
on Rimes 2011 database

System	WER
HMM without LM	45%
HMM with LM	**31,2%**
BLSTM without LM , with dictionary	26,5%
BLSTM with LM	**16,1 %**
BLSTM without dictionary, without LM	60,2%

interest of RNNs over HMMs : RNNs perform better in terms of accuracy
(generative versus discriminative approach)

Reconnaissance d'images de mots hors vocabulaire

- utilisation du BLSTM sans dictionnaire statique (pour les OOVs)
  - chaîne de caracteres
- puis comparaison des chaînes avec lexique Wikipedia
  - dictionnaire dynamique (mots wikipedia)
- décodage de l'image du mot avec dictionnaire dynamique



Conclusions et perspectives

- BLSTMs meilleurs que HMMs
  - taux de reconnaissance
  - discriminatifs/génératifs
  - mais plus longs à entraîner

- Autres architectures :
  - MDLSTM: multi-dimensional recurrent neural networks
  - 4 parcours au lieu de 2

- BLSTMs peuvent être associés à des Deep Nets
  - Bluche et al., 2014

References

- Graves et al. 2006, Connectionist Temporal Classification: labelling unsegmented sequence data with RNNs, Int. Conf. on Machine Learning, Pittsburgh, 2006.
- Alex Graves, Supervised labeling with recurrent neural networks, 2012, Springer.
- Alex Graves, Supervised sequence labelling with recurrent neural networks, PhD thesis, Technishen Universitat, Munich, 2008.
- S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Computation, 1997.
- M Schuster, KK Paliwal , Bidirectional recurrent neural networks, Signal Processing, IEEE Transactions on, 1997
- F. A. Gers and J. Schmidhuber. LSTM Recurrent Networks Learn Simple Context Free and Context Sensitive Languages. IEEE Trans. on Neural Networks, 2001
- T. Bluche et al., A Comparison of Sequence-Trained Deep Neural Networks and Recurrent Neural Networks Optical Modeling for Handwriting Recognition, SLSP 2014.

Acknowledgements

- **Olivier Morillot:** *Reconnaissance d'écriture manuscrite par modèles de Markov cachés et réseaux récurrents, These TPTech, janvier 2014.*
- → Comparaison HMM et BLSTM

- **Cristina Oprean:** *Reconnaissance de textes manuscrits grand vocabulaire, These TPTech en cours*
- → Traitement des mots hors vocabulaire avec ressources Web

- **Marcus Eichenberger-Liwicki** (DFKI, Germany), Seminaire GRCE-TPTech 2011
  - Handwriting Recognition with Long Short-Term Memory Networks

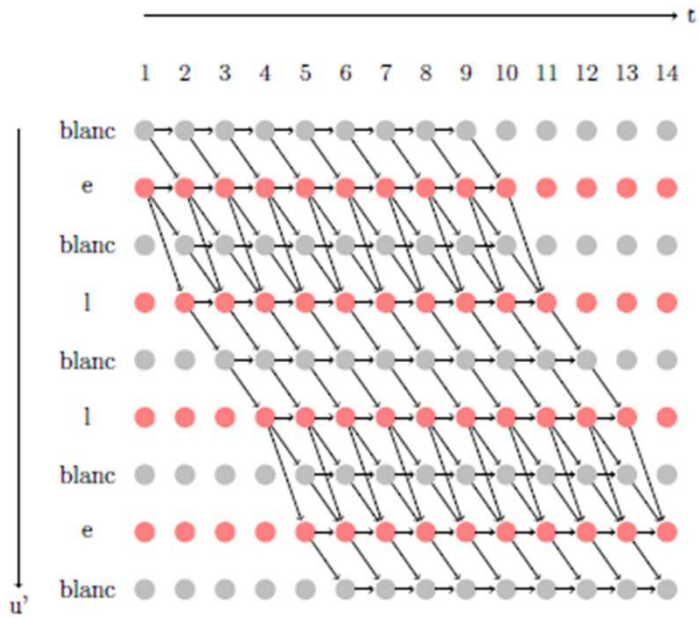
Connectionnist Temporal classification CTC

- adds the « blank » model
  - beginning, end and between characters
- provides score of label sequence l
  - l= word, text line

$$p(l|x) = \sum_{\text{admissible paths } \pi \text{ for l}} p(\pi|x)$$

$$p(\pi|x) = \prod_{t=1}^{T} y^t_{\pi_t}$$

$$p(l|x) = \sum_{s=1}^{|l'|} \frac{\alpha_s(t)\beta_t(s)}{y^t_{l'_s}}$$



l=« **e l l e** »

l'=«ε **e** ε **l** ε **l** ε **e** ε »

admissible path=«ε e ε l l ε l ε e **e**εε