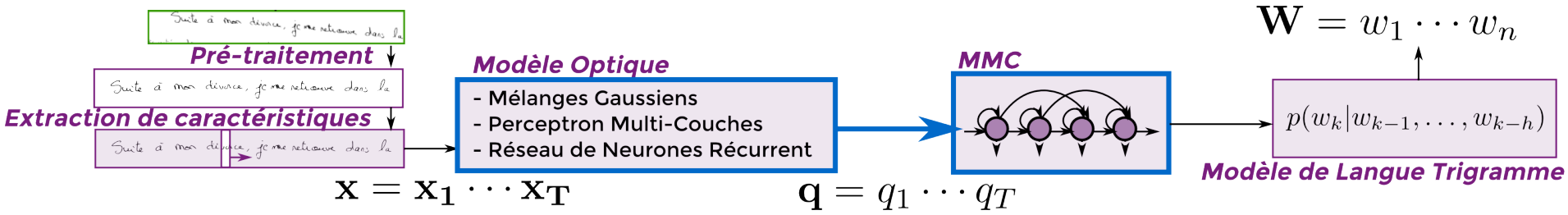
Colloque International Francophone sur l'Écrit et le Document

La CTC et son intrigant label "blank"

Théodore Bluche, Hermann Ney, Jérôme Louradour, Christopher Kermorvant

9 mars 2016

Systèmes Hybrides pour la Reconnaissance d'Écriture



Observations

- Travaux dans les années 90 sur l'entraînement intégré NN/HMM
- Aujourd'hui entraînement framewise ou CTC de RNNs

Questions

- Quelles sont les **différences entre les entraînements framewise, CTC, intégré** ?
- Peut-on appliquer la **CTC à d'autres réseaux** que des RNNs ?
- Peut-on utiliser la formulation de la **CTC avec d'autre cibles que les caractères** et le blank (e.g. états de HMM) ?
- Quel est le **rôle du symbole blank** dans la CTC? Quand et comment est-il utile?

Overview

Introduction

Entraînement de Réseaux de Neurones pour les Systèmes Hybrides NN/HMM

Comparaison des Entraînements Framewise et CTC des MLPs et RNNs

L'intrigant Label Blank de la CTC

Conclusion

Entraînement de Réseaux de Neurones pour les Systèmes Hybrides NN/HMM

Introduction

Entraînement de Réseaux de Neurones pour les Systèmes Hybrides NN/HMM

Comparaison des Entraînements Framewise et CTC des MLPs et RNNs

L'intrigant Label Blank de la CTC

Conclusion

Entraînement NN/HMM intégré

- MMC = approche à segmentation implicite $\longrightarrow$ entraînement Baum-Welch pour considérer toutes les segmentations possibles
- Forward-backward en utilisant les posteriors renormalisées comme modèle d'émission :

$$\alpha_t(s) = \frac{p(q_t=s|x_t)}{p(s)} \times \sum_r \alpha_{t-1}(r) p(q_t = s|q_{t-1} = r)$$
$$\beta_t(s) = \sum_r \frac{p(q_{t+1}=r|x_{t+1})}{p(r)} p(q_{t+1} = r|q_t = s)\beta_{t+1}(r)$$
$$p(q_t = s|\mathbf{x}, \lambda) = \frac{\alpha_t(s)\beta_t(s)}{\sum_r \alpha_t(r)\beta_t(r)}$$

- Senior & Robinson (1996); Yan et al. (1997) : 1er entraînement avec les alignements forcés, puis ajustement avec forward-backward
- Konig et al. (1996); Hennebert et al. (1997) : formulation similaire.

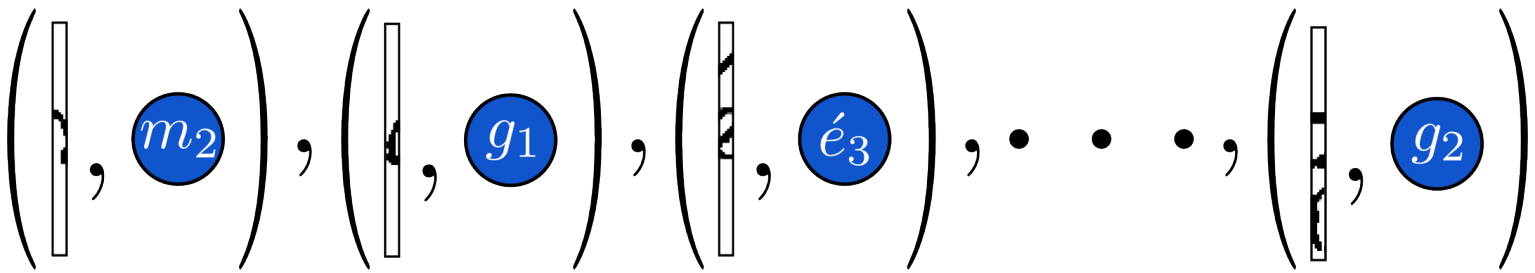
Fonction de coût

$$E_{hmm} = -\log \sum_{\mathbf{q}} \prod_t \frac{p(q_t|x_t)}{p(q_t)} p(q_t|q_{t-1})$$

- Bengio et al. (1992); Haffner (1993): coût MMI appliqué au système complet.

Entraînement Framewise Entropie Croisée

1. Calcule les alignements forcés de la séquence de trames avec le MMC de la séquence de mots cible $\longrightarrow$ jeu de données étiqueté de trames $\mathcal{S} = \{(x_t, q_t)\}$



2. Entraînement du réseau à classifier chaque trame séparemment :

Coût d'Entropie croisée :

$$E_{xent} = -\sum_{(x_t, q_t) \in \mathcal{S}} \log P(q_t | x_t)$$

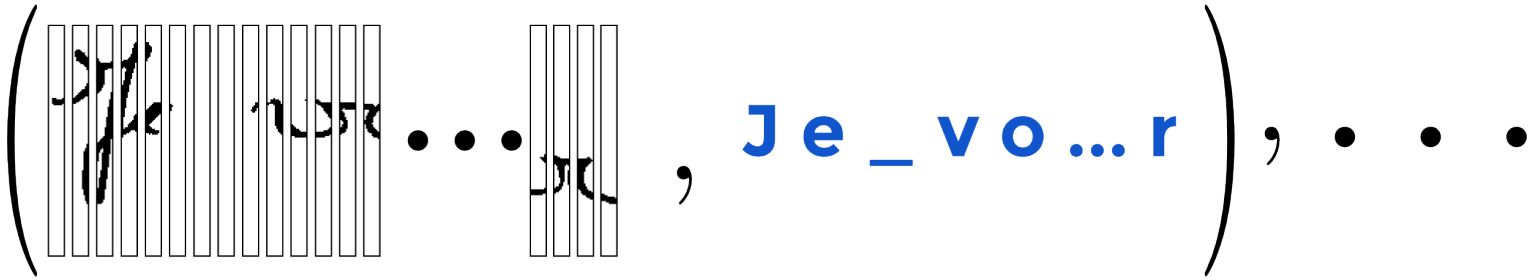
Évaluation

Frame Error Rate (FER%)

$$\frac{\text{\# trames mal classifiées}}{\text{\# trames}}$$

Entraînement "Connectionnist Temporal Classification" (CTC)

1. Jeu de données de séquences de trames, étiquetées avec les séquences de caractères $\mathcal{S} = \{(\mathbf{x}, \mathbf{c})\}$



2. Entraînement du réseau à prédire la séquence de caractères $\mathbf{c}$
   - Sorties du réseau = caractères + $\oslash$
   - Mapping $\mathcal{B} : a\ a \oslash \oslash b\ b \oslash b\ a \mapsto abba$

Coût CTC:

$$E_{ctc} = -\sum_{(\mathbf{x},\mathbf{c})\in\mathcal{S}} \log P(\mathbf{c}|\mathbf{x})$$

avec

$$P(\mathbf{c}|\mathbf{x}) = \sum_{\mathbf{q}\in\mathcal{B}^{-1}(\mathbf{c})} P(\mathbf{q}|\mathbf{x}) = \sum_{\mathbf{q}\in\mathcal{B}^{-1}(\mathbf{c})} \prod_t P(q_t|\mathbf{x})$$

Évaluation

NN - Character Error Rate (NN-CER%)

$$\frac{\text{distance d'édition}}{\text{\# caractères cibles}}$$

(Graves et al., 2006)

Résumé des Stratégies d'Entraînement

	Coût	Sorties du réseau
Framewise entropie croisée (MLPs)	$-\log \prod_t P(q_t\|x_t)$	États MMC (5-6 / caractère)
CTC (RNNs) (Graves et al., 2006)	$-\log \sum_{\mathbf{q}} \prod_t P(q_t\|\mathbf{x})$	Caractères et label blank $\varnothing$
Entr. MMC intégré (NN/HMM) (Hennebert et al., 1997)	$-\log \sum_{\mathbf{q}} \prod_t \frac{P(q_t\|x_t)}{P(q_t)} P(q_t\|q_{t-1})$	États MMC (5-6 / caractère)

Comparaison des Entraînements Framewise et CTC des MLPs et RNNs

Introduction

Entraînement de Réseaux de Neurones pour les Systèmes Hybrides NN/HMM

Comparaison des Entraînements Framewise et CTC des MLPs et RNNs

L'intrigant Label Blank de la CTC

Conclusion

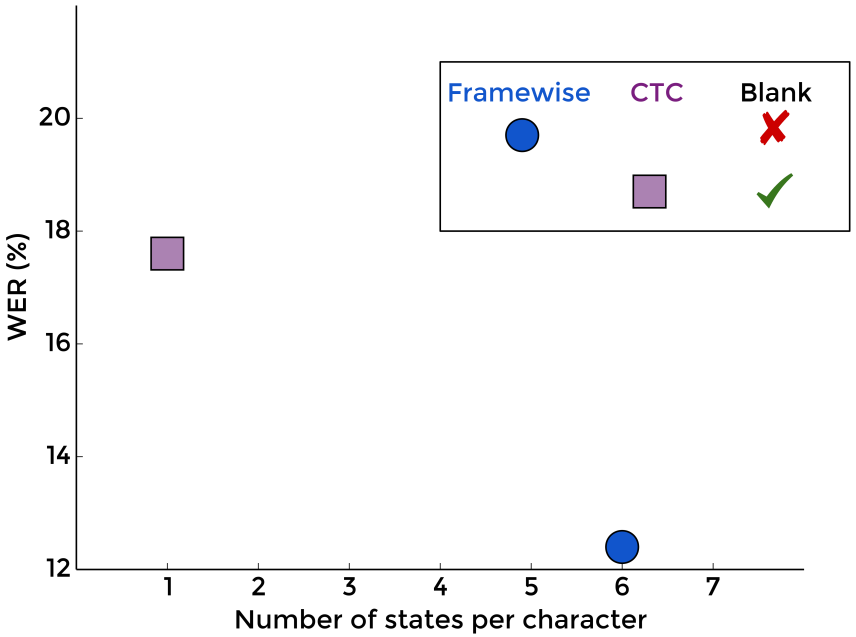
Stratégies d'Entraînement de NN/HMM

CTC = entraînement MMC intégré, sans probabilités a priori / de transition, et avec des sorties spécifiques pour la reco avec NN seul ($\approx$ 1 état de MMC / char. $+$ blank)
$\Longrightarrow$ le CTC pourrait être appliqué à d'autre topologies de MMC, et à d'autres réseaux que des RNNs (e.g. MLPs)

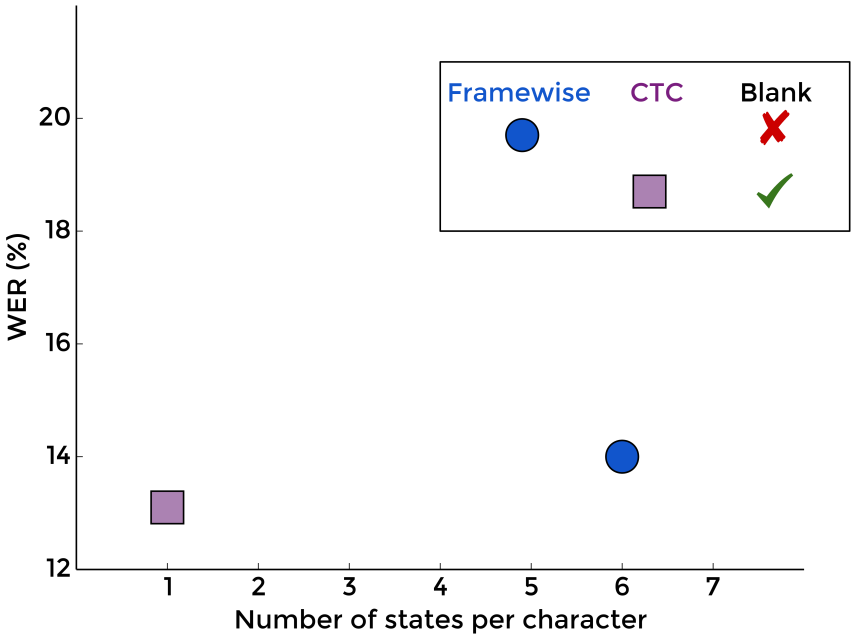
CTC = entraînement entropie croisée + forward-backward pour prendre en compte toutes les segmentations possibles
$\Longrightarrow$ on peut comparer les stratégies d'entraînement, voir l'effet du forward-backward, avec différentes topologies (nombres d'états de MMC / char.)

CTC vs. niveau trame

MLPs



RNNs

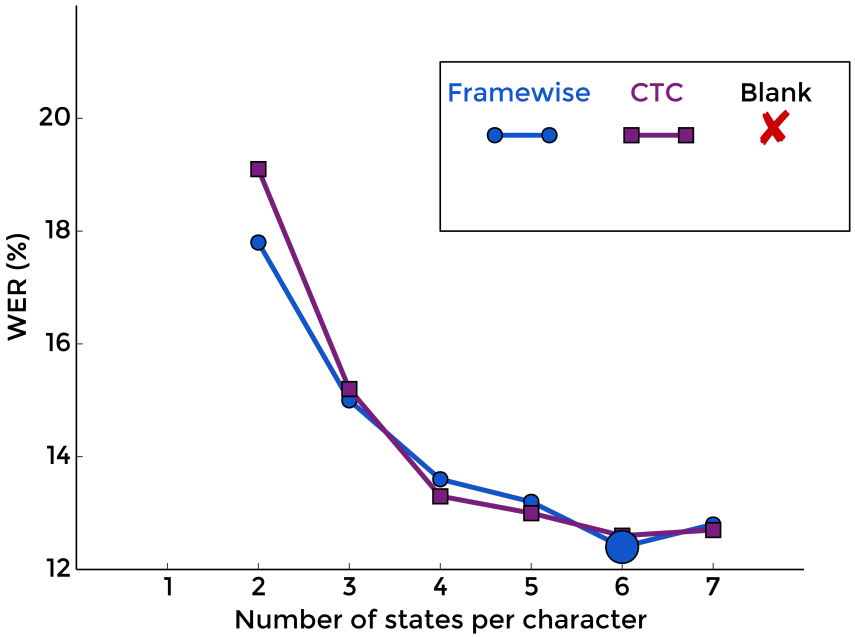


$\longrightarrow$ La CTC marche bien pour les RNNs, moins pour les MLPs

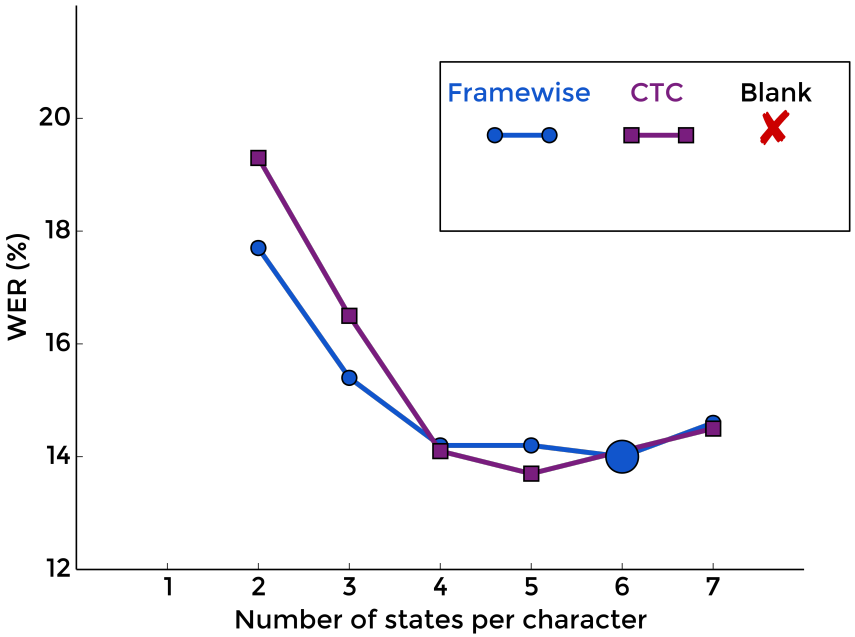
(MLP: 2x1024, ± 5 frames - RNN: 1x100)

CTC vs. niveau trame

MLPs



RNNs

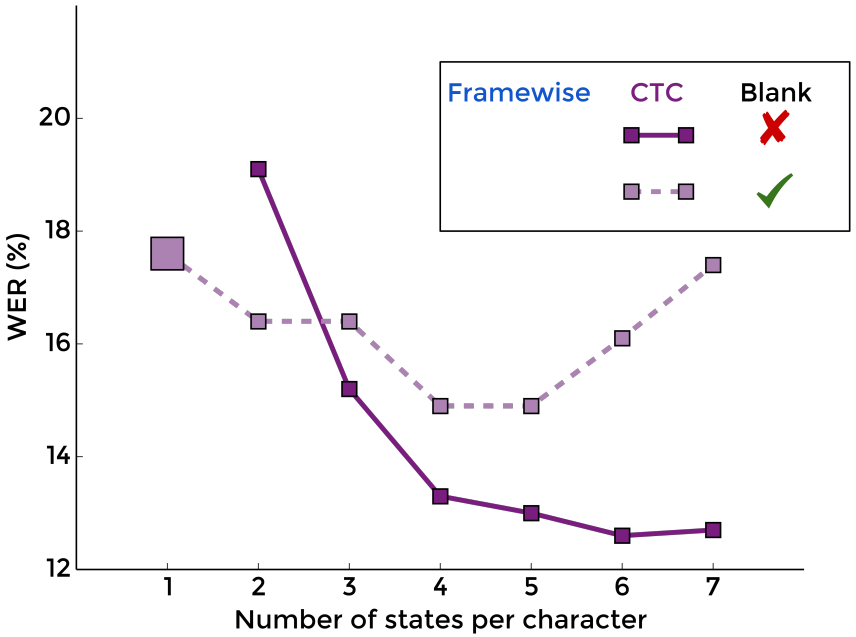


$\longrightarrow$ L'aspect Forward-backward n'améliore pas les résultats, et est pire pour peu d'états

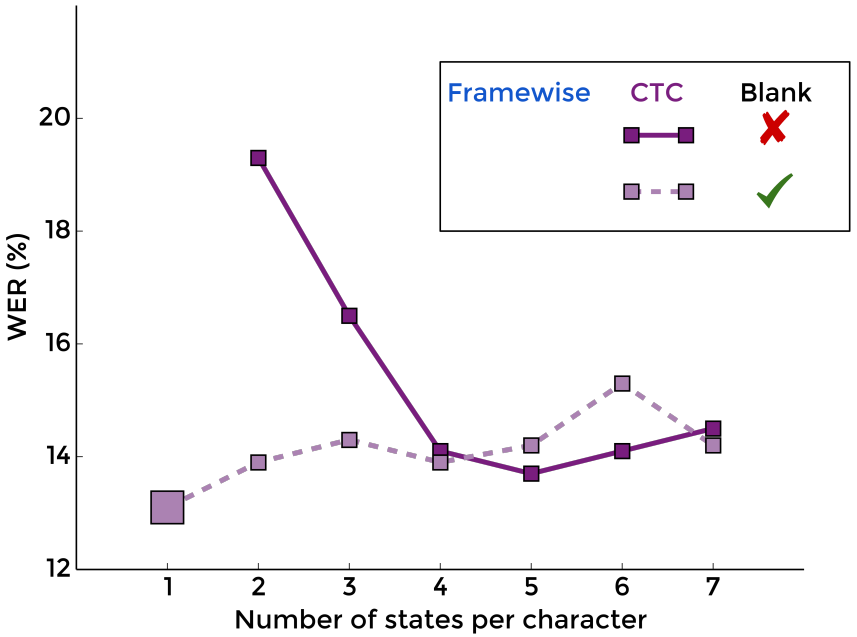
(MLP: 2x1024, ± 5 frames - RNN: 1x100)

CTC vs. niveau trame

MLPs



RNNs

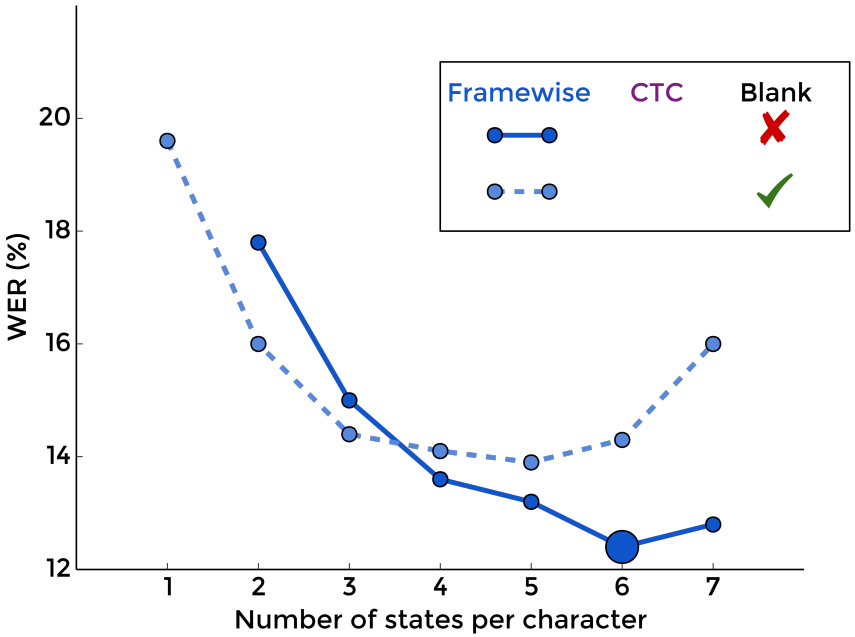


$\longrightarrow$ Le label blank aide seulement avec peu d'états pour l'entraînement CTC, ...

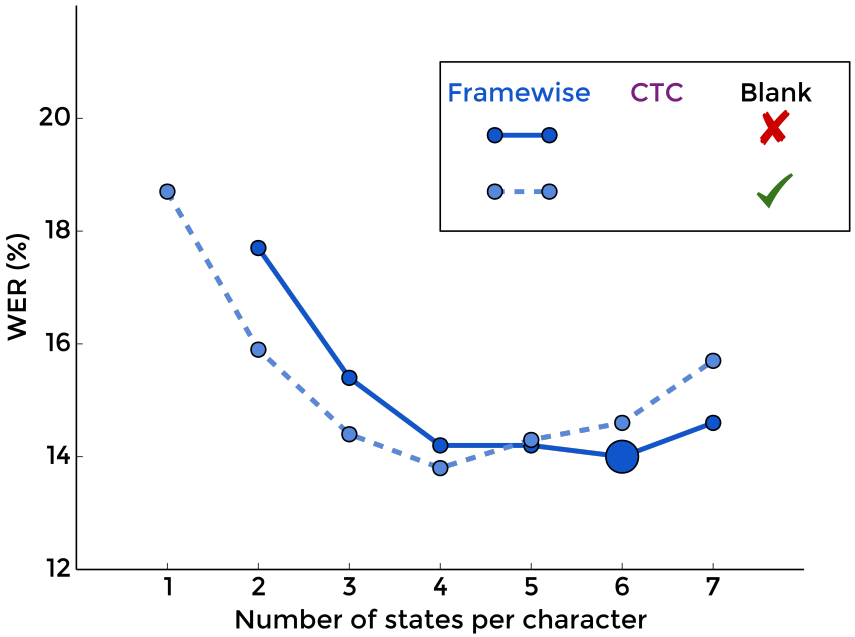
(MLP: 2x1024, ± 5 frames - RNN: 1x100)

CTC vs. niveau trame

MLPs



RNNs

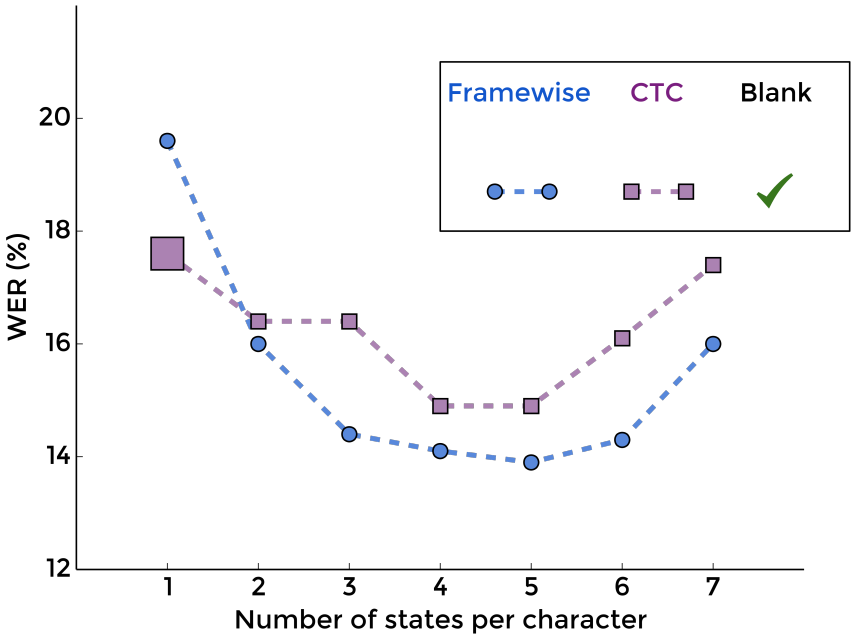


$\longrightarrow$ ... et également au niveau trame, bien que pas autant qu'ajouter un état aux modèles de caractères

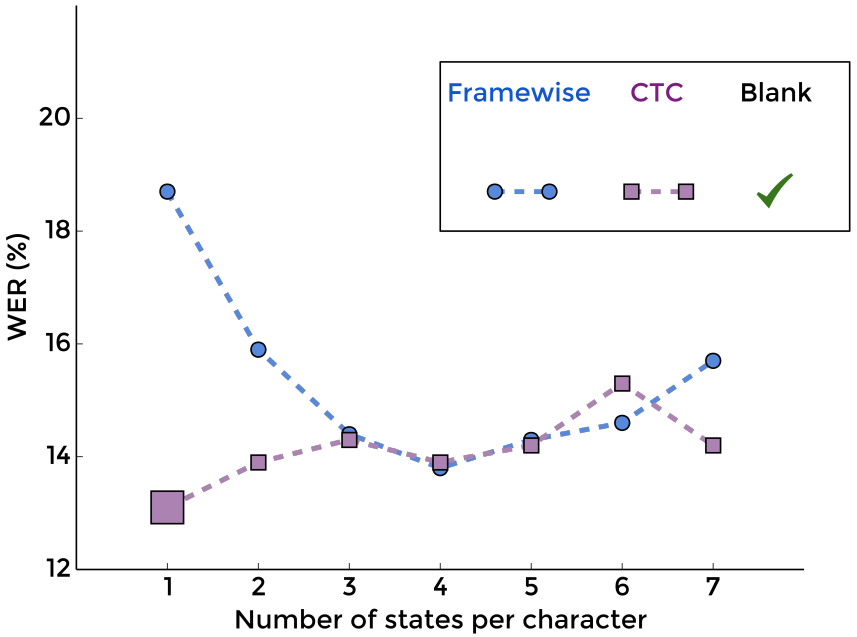
(MLP: 2x1024, ± 5 frames - RNN: 1x100)

CTC vs. niveau trame

MLPs



RNNs

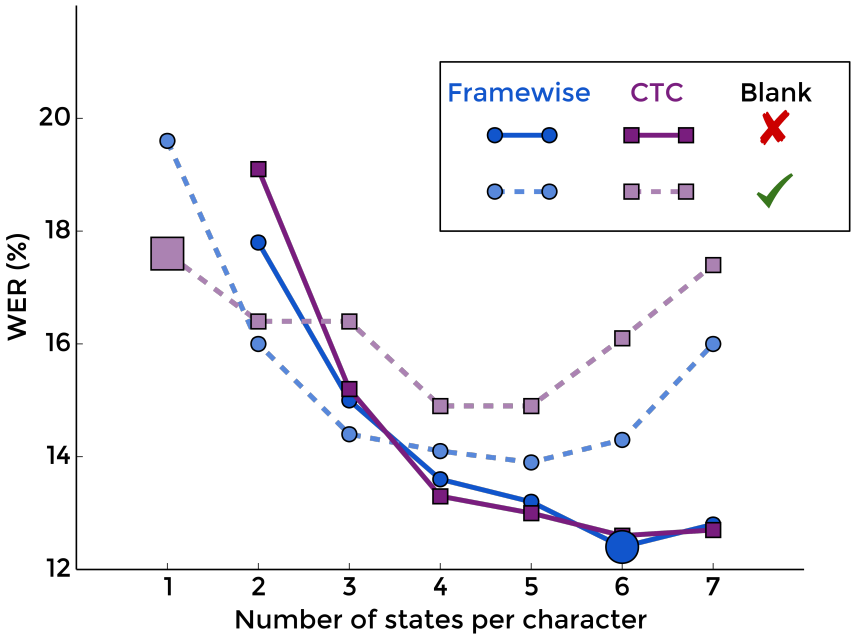


$\longrightarrow$ Forward-backward avec blank n'améliore pas vraiment les résultats sauf avec peu d'états

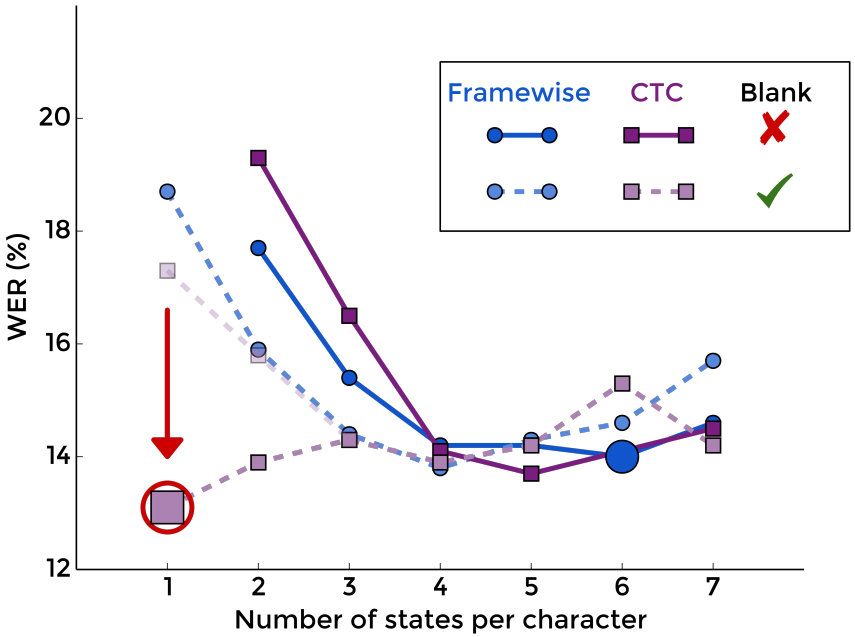
(MLP: 2x1024, ± 5 frames - RNN: 1x100)

CTC vs. niveau trame

MLPs



RNNs



$\longrightarrow$ La CTC+blank, avec des modèles à un état, est particulièrement adaptée aux RNNs

(MLP: 2x1024, ± 5 frames - RNN: 1x100)

L'intrigant Label Blank de la CTC

Introduction

Entraînement de Réseaux de Neurones pour les Systèmes Hybrides NN/HMM

Comparaison des Entraînements Framewise et CTC des MLPs et RNNs

L'intrigant Label Blank de la CTC

Conclusion

L'intrigant Label Blank de la CTC

Le label blank...

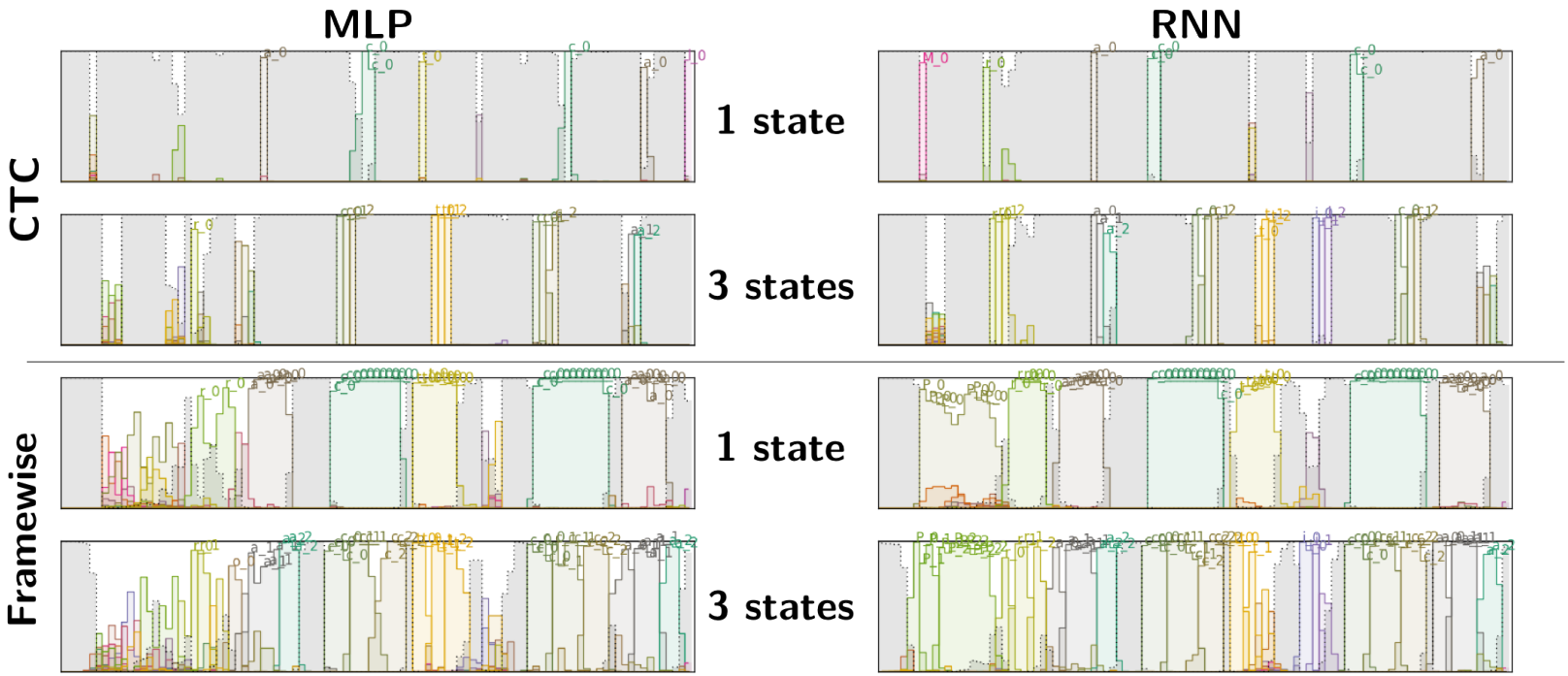
- est nécessaire pour obtenir de bons résultats avec peu d'états
- (Graves et al., 2006): requis pour le mapping et la modélisation des inter-caractères
- ... mais sinon non-informatif,
- peu d'intérêt avec plus d'états
- et a un impact sur les sorties (prochains slides)

$\longrightarrow$ Que ce passe-t-il avec ce label? Comment se comportent les systèmes quand il est présent, et en quoi peut-il être bénéfique?

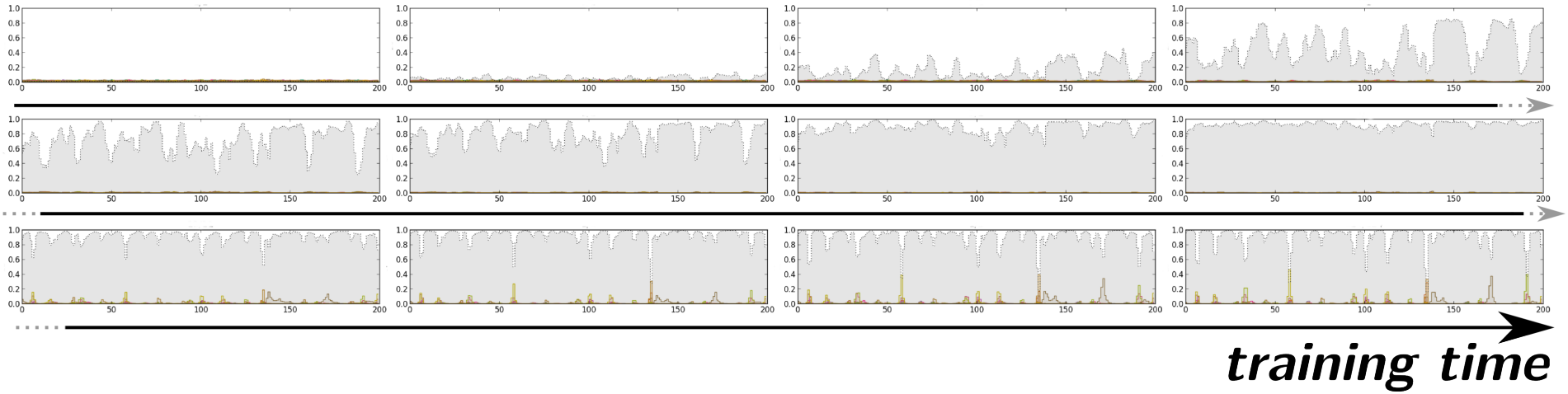
nb. ce genre de symbole non-caractère / junk apparaît dans d'autres travaux, e.g. (Tay et al., 2001; Rashid et al., 2012; Elagouni et al., 2012)

Sorties Framewise vs. CTC avec Blank

- x-axis = temps
- y-axis = probabilité prédite
- girs = blank
- couleur = caractères

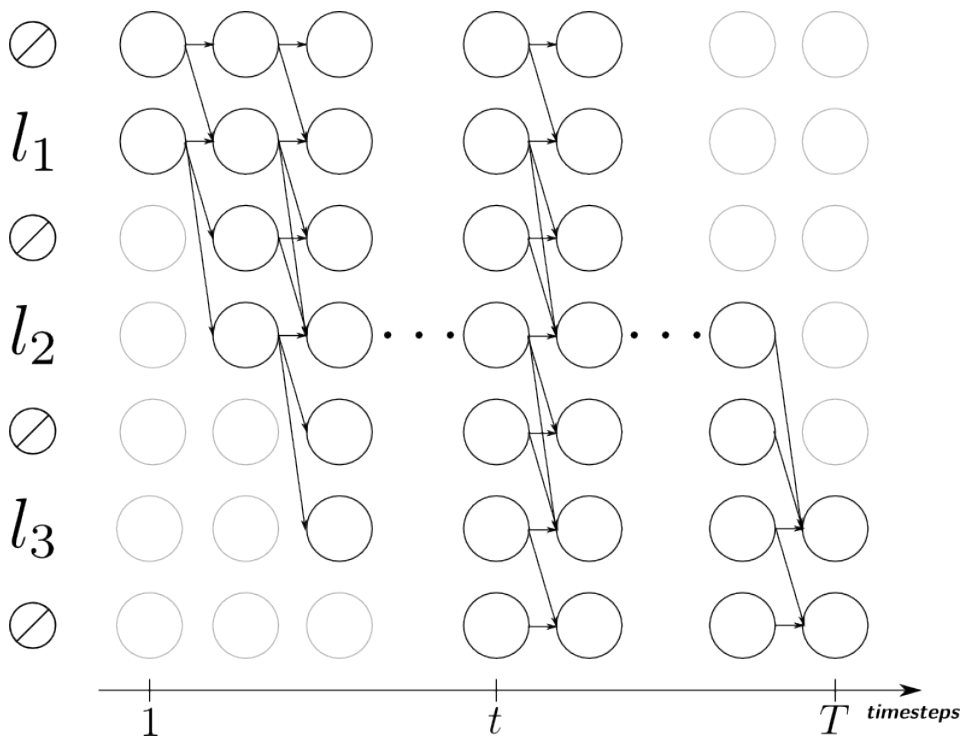


CTC - Évolution des Sorties pendant l'Entraînement



- x-axis = temps
- y-axis = probabilité prédite
- gris = blank

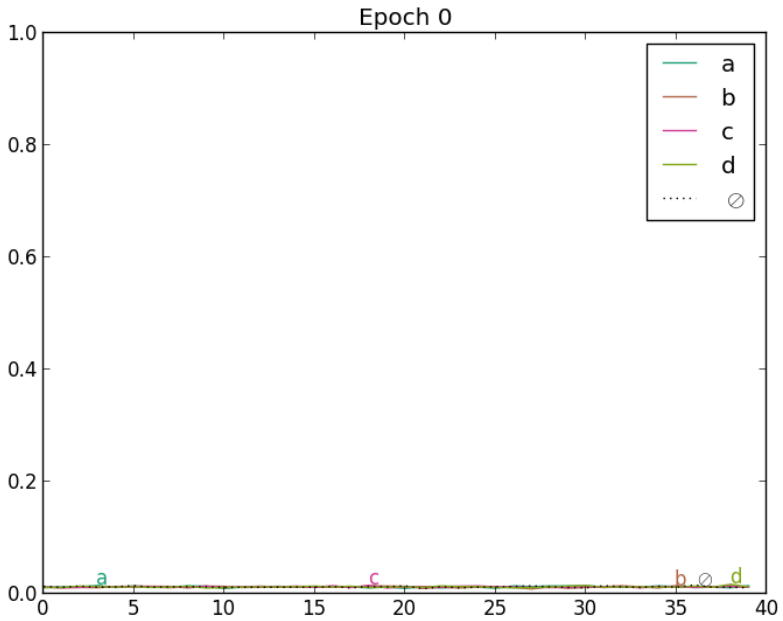
Entraînement CTC



$$\frac{\partial E}{\partial a_k^t} = p_{nn}(q_t = k|x_t) - \sum_{s:\mu(s)=k} \frac{\alpha_t(s)\beta_t(s)}{\sum_r \alpha_t(r)\beta_t(r)}$$

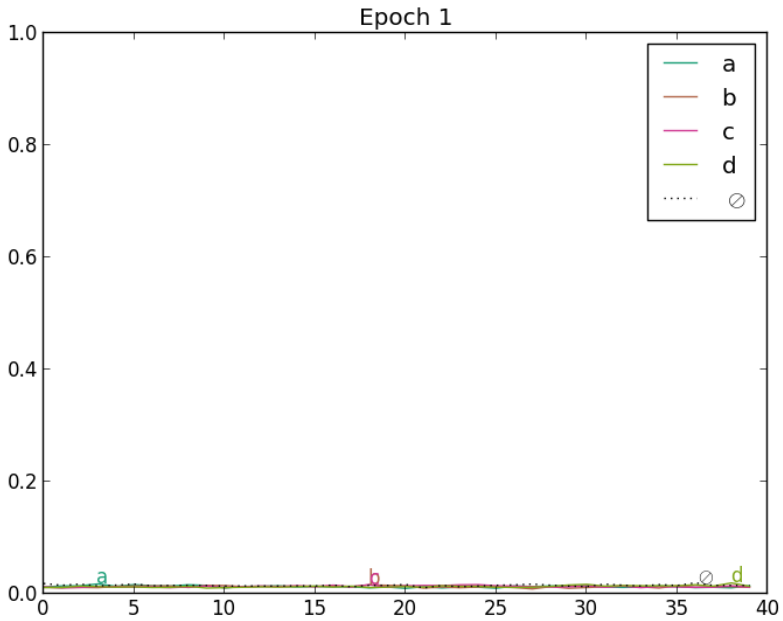
Entraînement CTC, exemple joué

Ce comportement est **dû à la CTC, pas aux RNNs**.
Exemple: descente de gradient sur les **entrées de la CTC**



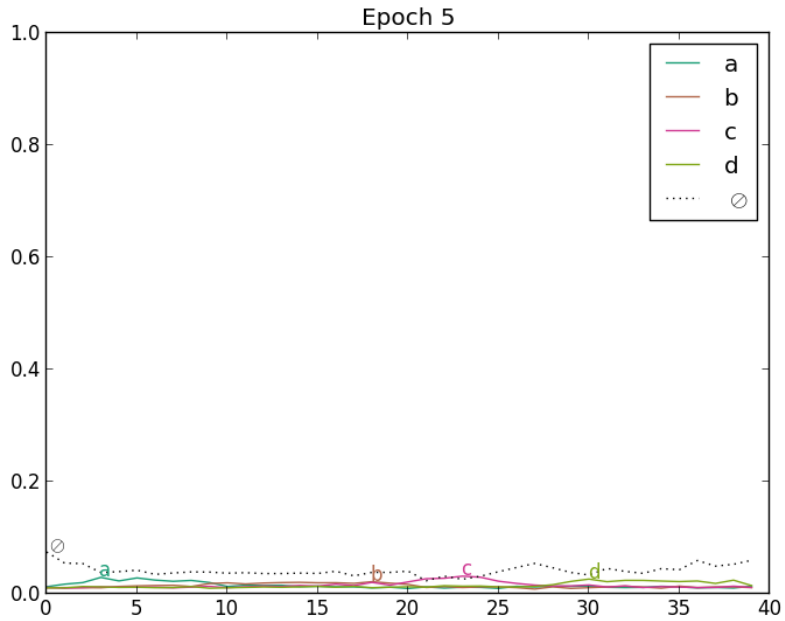
Entraînement CTC, exemple joué

Ce comportement est **dû à la CTC, pas aux RNNs**.
Exemple: descente de gradient sur les **entrées de la CTC**



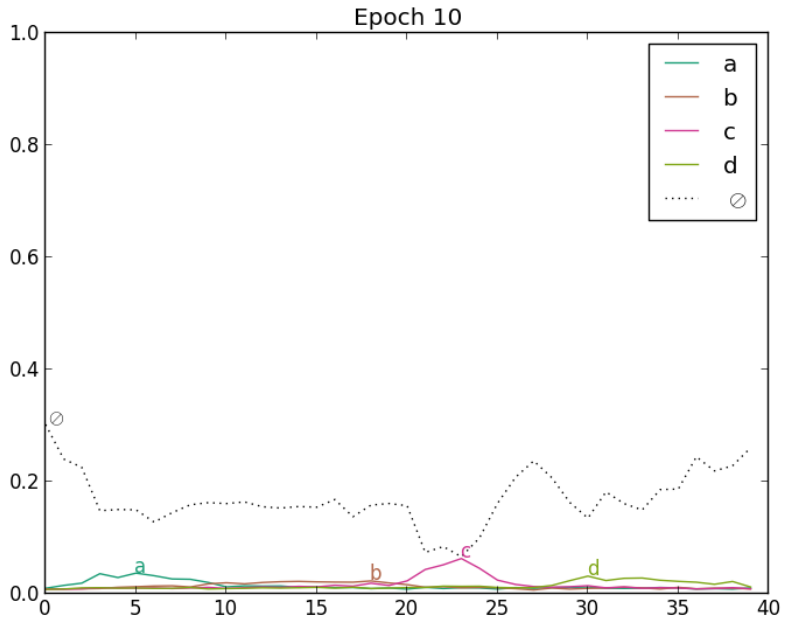
Entraînement CTC, exemple joué

Ce comportement est **dû à la CTC, pas aux RNNs**.
Exemple: descente de gradient sur les **entrées de la CTC**



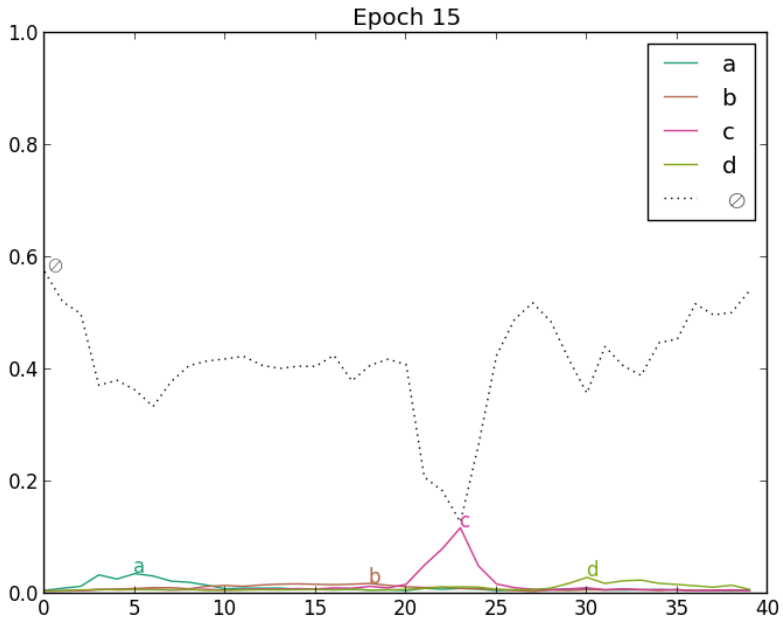
Entraînement CTC, exemple joué

Ce comportement est **dû à la CTC, pas aux RNNs**.
Exemple: descente de gradient sur les **entrées de la CTC**



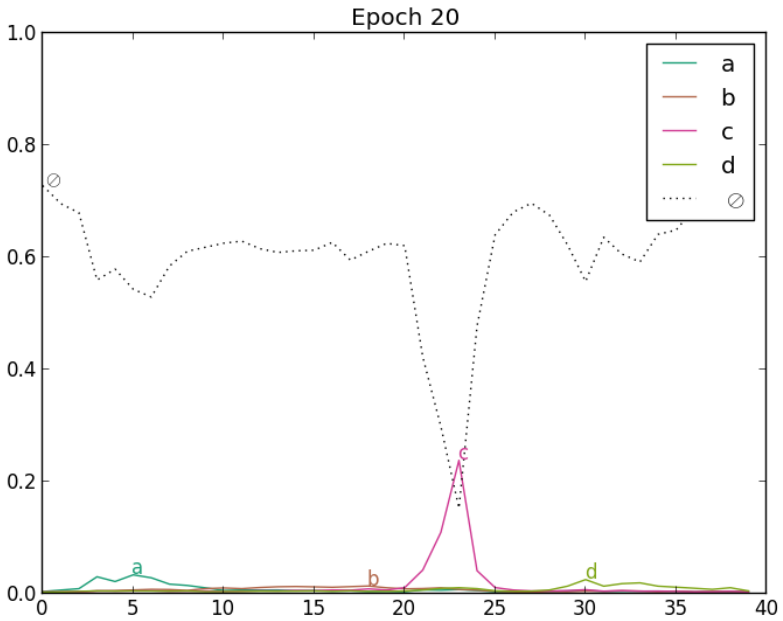
Entraînement CTC, exemple joué

Ce comportement est **dû à la CTC, pas aux RNNs**.
Exemple: descente de gradient sur les **entrées de la CTC**



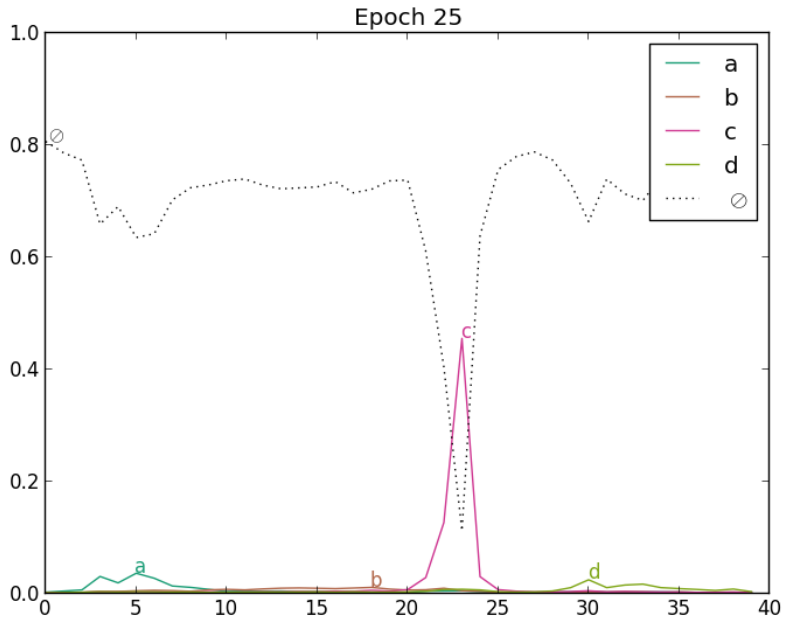
Entraînement CTC, exemple joué

Ce comportement est **dû à la CTC, pas aux RNNs**.
Exemple: descente de gradient sur les **entrées de la CTC**



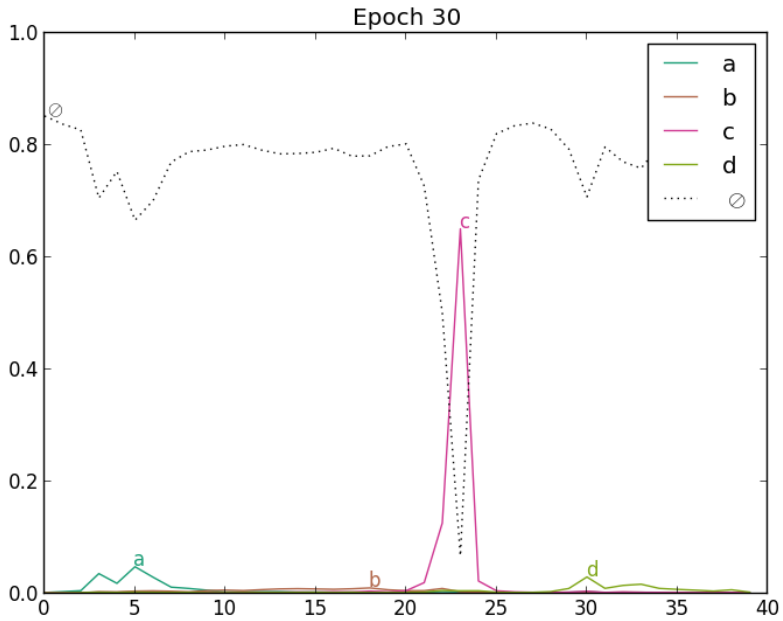
Entraînement CTC, exemple joué

Ce comportement est **dû à la CTC, pas aux RNNs**.
Exemple: descente de gradient sur les **entrées de la CTC**



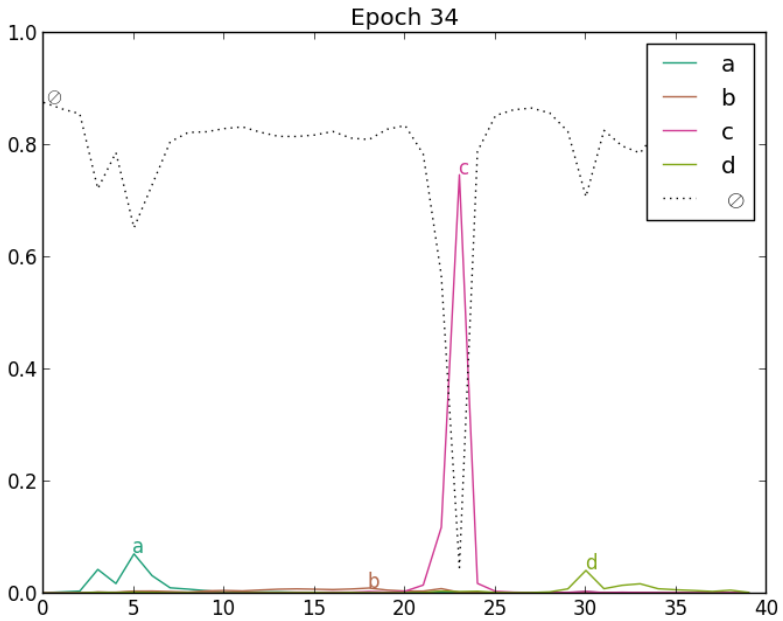
Entraînement CTC, exemple joué

Ce comportement est **dû à la CTC, pas aux RNNs**.
Exemple: descente de gradient sur les **entrées de la CTC**



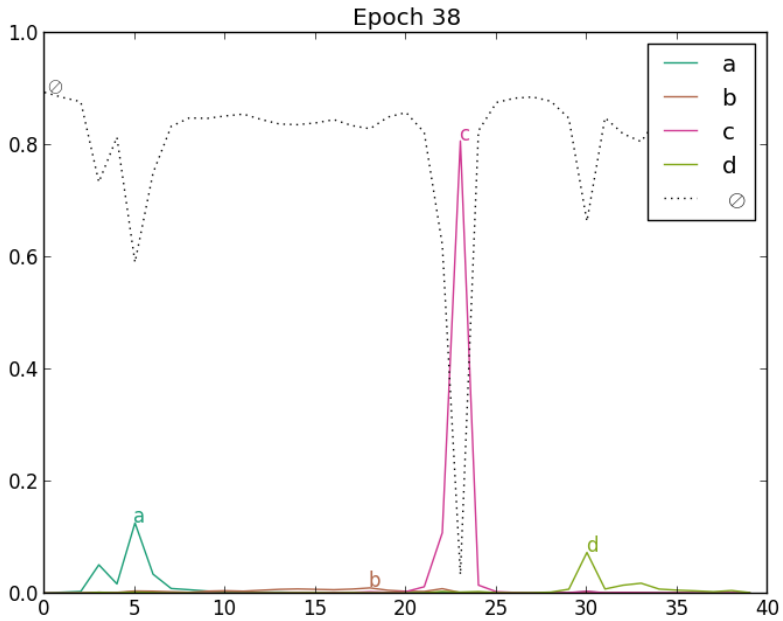
Entraînement CTC, exemple joué

Ce comportement est **dû à la CTC, pas aux RNNs**.
Exemple: descente de gradient sur les **entrées de la CTC**



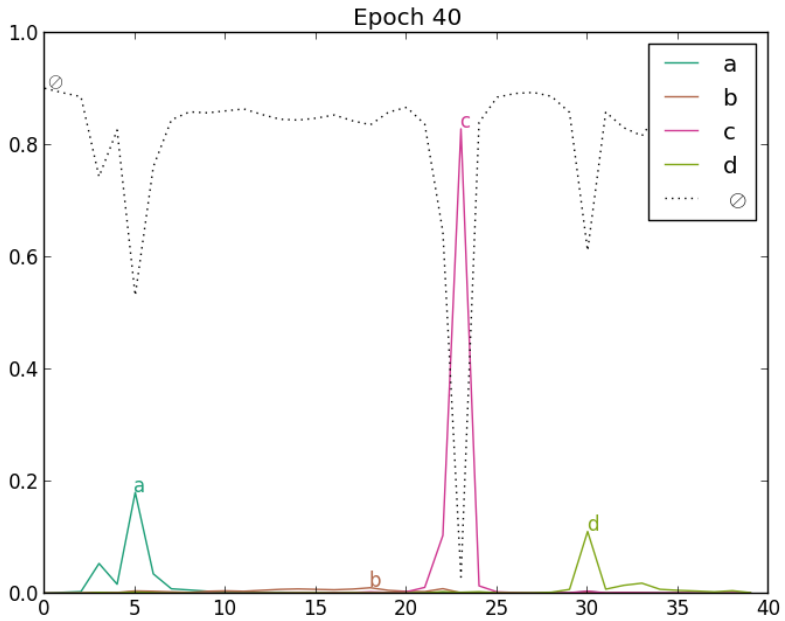
Entraînement CTC, exemple joué

Ce comportement est **dû à la CTC, pas aux RNNs**.
Exemple: descente de gradient sur les **entrées de la CTC**



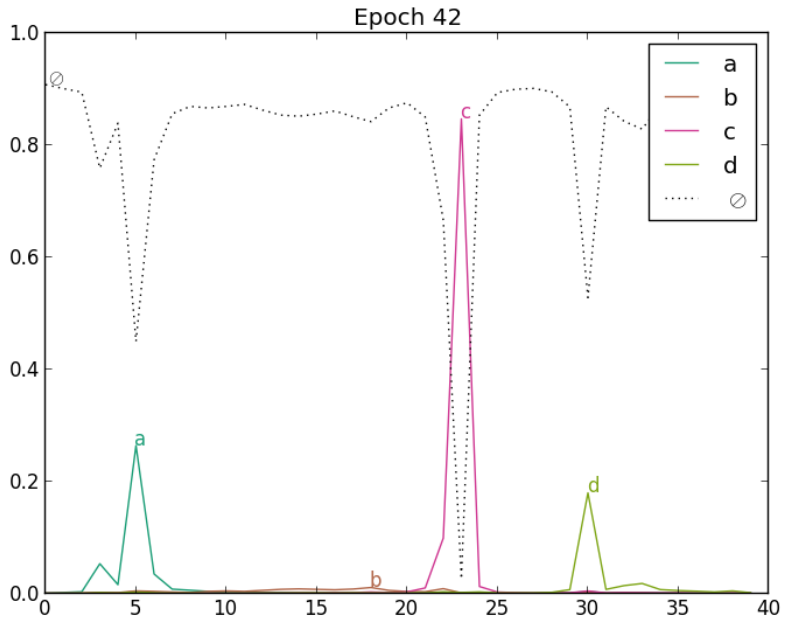
Entraînement CTC, exemple joué

Ce comportement est **dû à la CTC, pas aux RNNs**.
Exemple: descente de gradient sur les **entrées de la CTC**



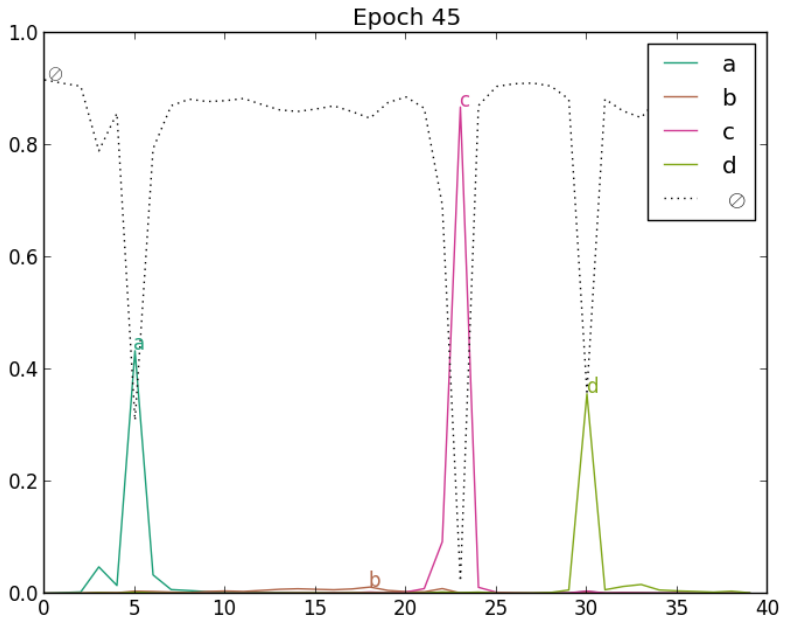
Entraînement CTC, exemple joué

Ce comportement est **dû à la CTC, pas aux RNNs**.
Exemple: descente de gradient sur les **entrées de la CTC**



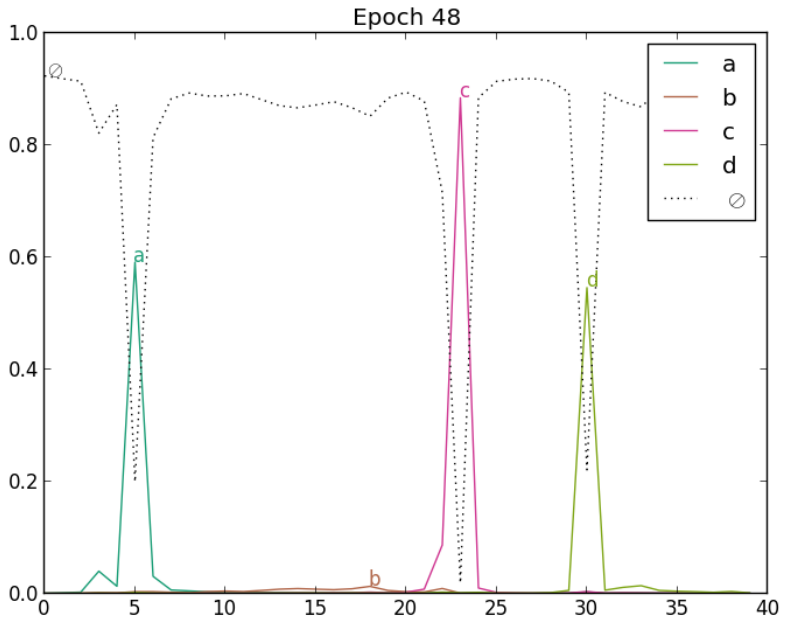
Entraînement CTC, exemple joué

Ce comportement est **dû à la CTC, pas aux RNNs**.
Exemple: descente de gradient sur les **entrées de la CTC**



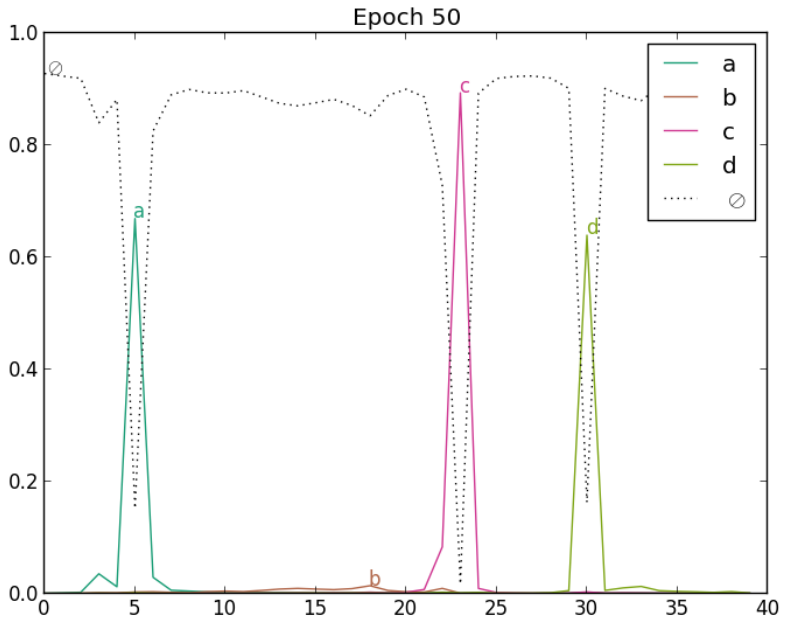
Entraînement CTC, exemple joué

Ce comportement est **dû à la CTC, pas aux RNNs**.
Exemple: descente de gradient sur les **entrées de la CTC**

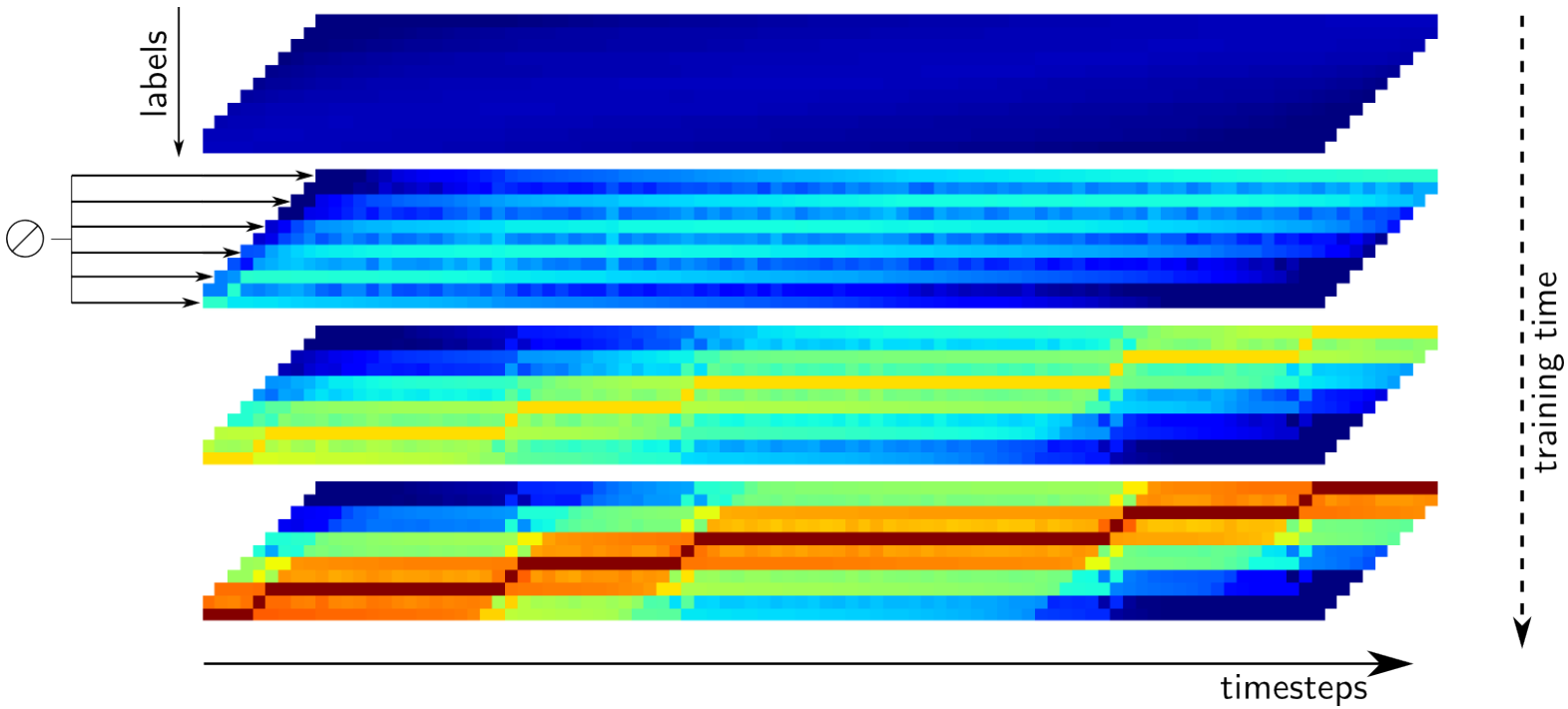


Entraînement CTC, exemple joué

Ce comportement est **dû à la CTC, pas aux RNNs**.
Exemple: descente de gradient sur les **entrées de la CTC**



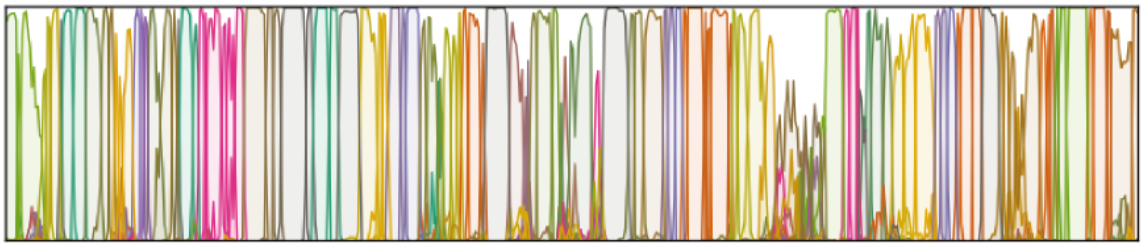
CTC - Pourquoi l'Entraînement provoque des pics



CTC - Que se passe-t-il sans Blank

(gris = espace)

Framewise



CTC
random
init

CTC
init 1ep
framewise

Les Advantages du Label Blank

Entraînement

En faisant pencher la distribution des sorties vers le blank, on évite certains problèmes d'alignements au début de l'entraînement, en particulier pour l'espace, quand les sorties du réseau ne sont pas informatives.

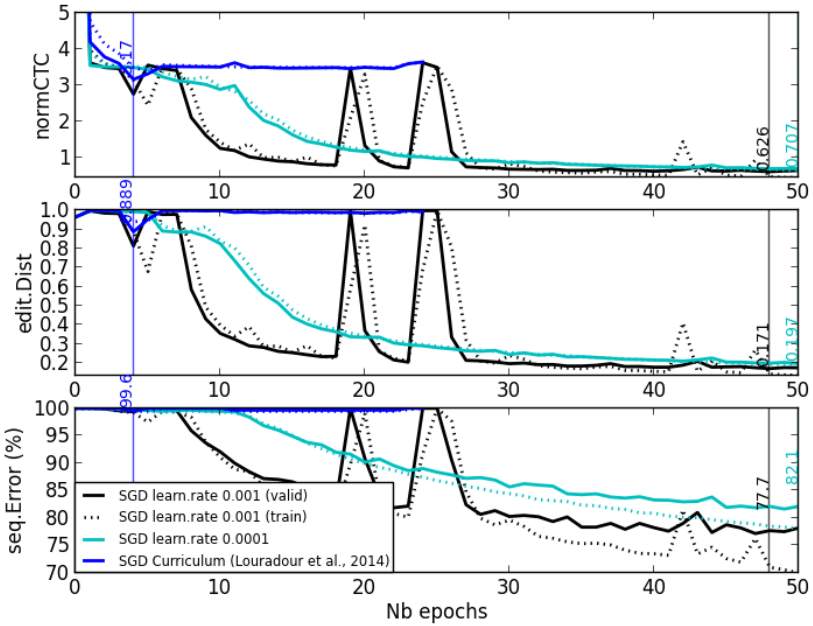
Transcription

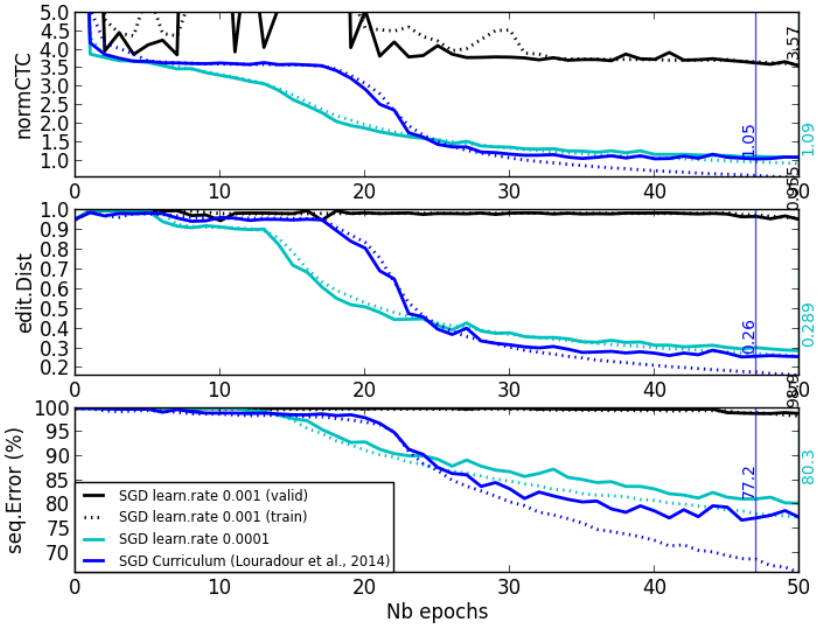
Les blanks n'apportent aucune information et sont partagés par tous les modèles de mot. Les prédictions de caractères sous forme de pics localisés permenttent des corrections moins coûteuses (une seule trame doit être changée pour corriger une substitution/deletion/insertion)

= plus rapide / plus d'hypothèses dans un même beam.

Sachant que le blank est rapidement omniprésent dans la séquence de prédictions, que peut-on faire?

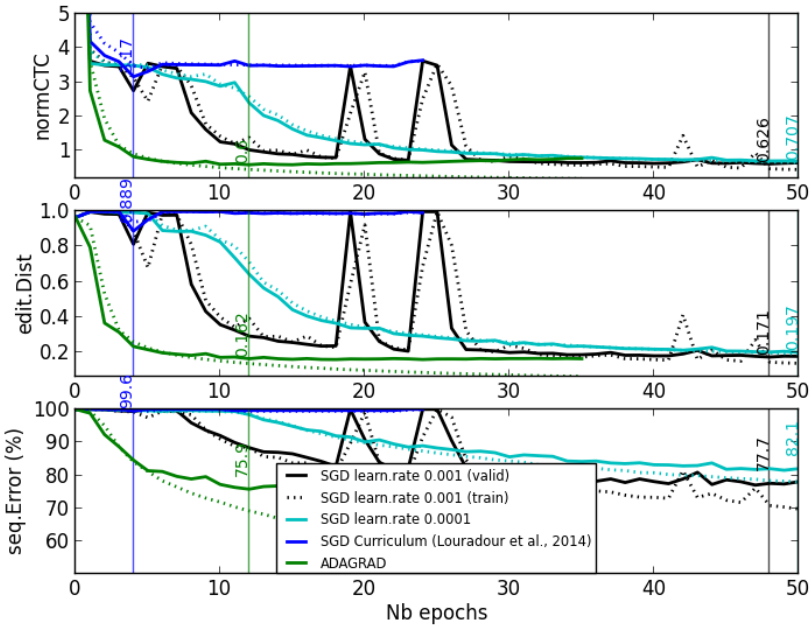
- On "perd" du temps au début en n'apprenant que des blanks non informatifs
- On observe parfois un plateau, et il faut quelque temps avant que le réseau prédise des caractères
- Un curriculum (Louradour & Kermorvant, 2014) ou un plus petit learning rate peut aider, mais le phénomène ne disparait pas

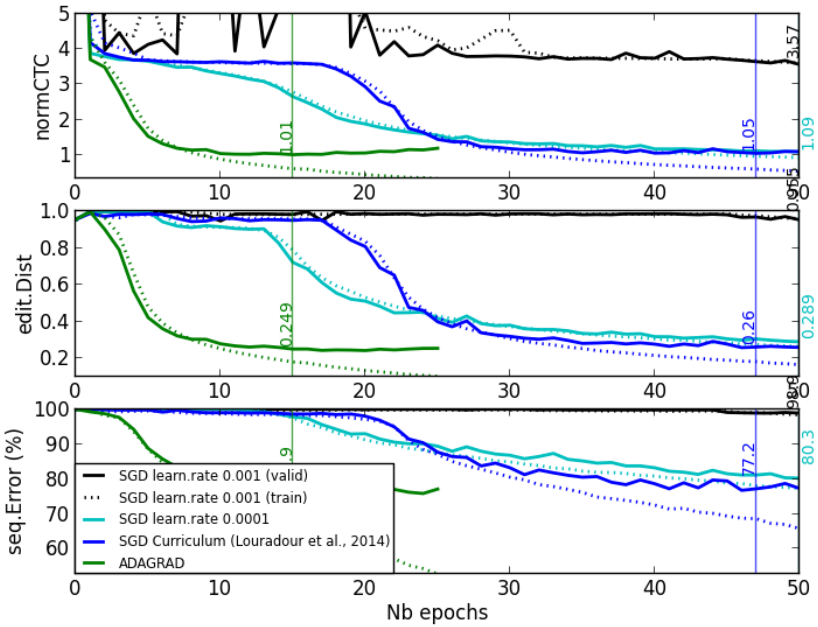




Sachant que le blank est rapidement omniprésent dans la séquence de prédictions, que peut-on faire?

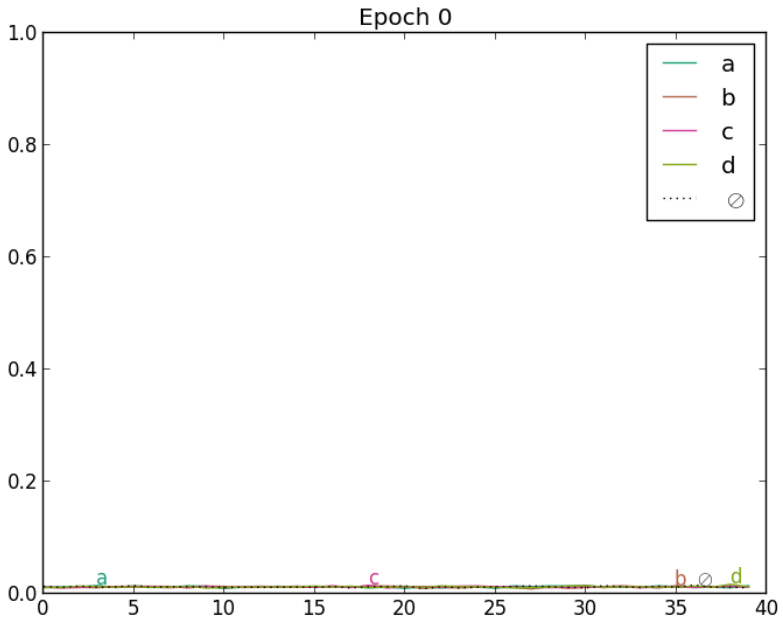
- Learning rate adaptatif par paramètre (ADAGRAD; Duchi et al. (2011)) pour donner progressivement moins d'importance au signal d'erreur venant du label blank.





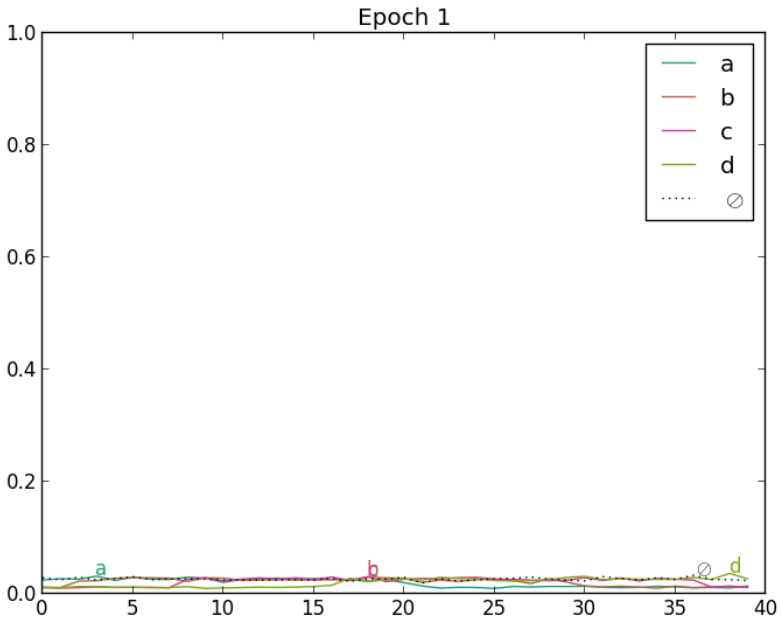
Entraînement CTC, exemple joué (2)

Exemple: descente de gradient sur les entrées de la CTC **avec ADAGRAD**



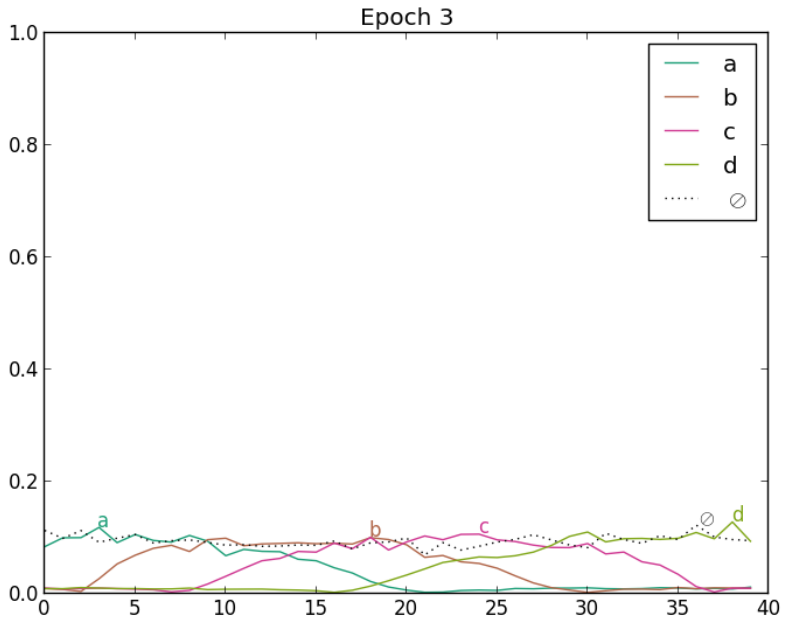
Entraînement CTC, exemple joué (2)

Exemple: descente de gradient sur les entrées de la CTC **avec ADAGRAD**



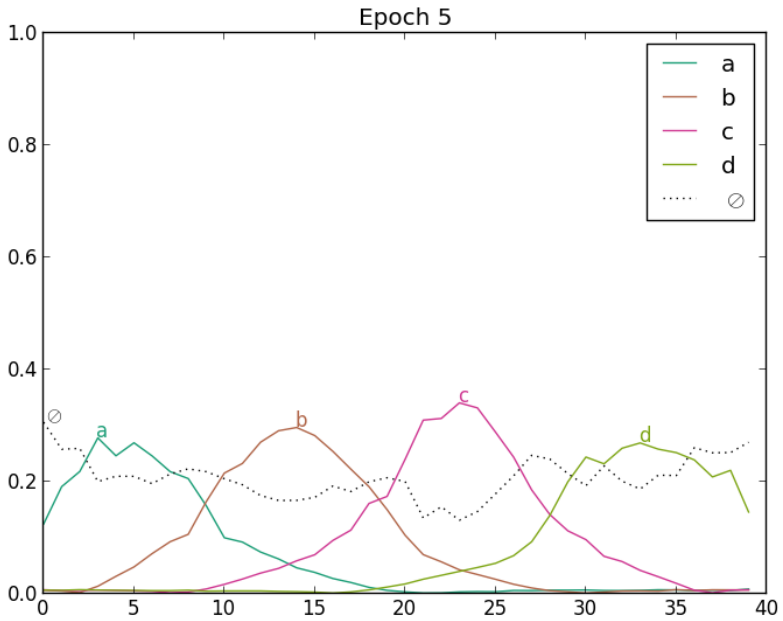
Entraînement CTC, exemple joué (2)

Exemple: descente de gradient sur les entrées de la CTC **avec ADAGRAD**



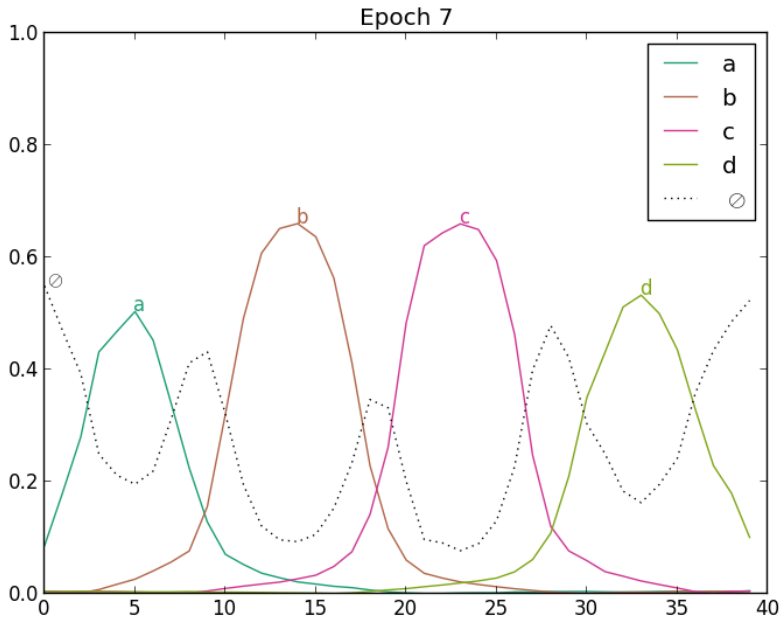
Entraînement CTC, exemple joué (2)

Exemple: descente de gradient sur les entrées de la CTC **avec ADAGRAD**



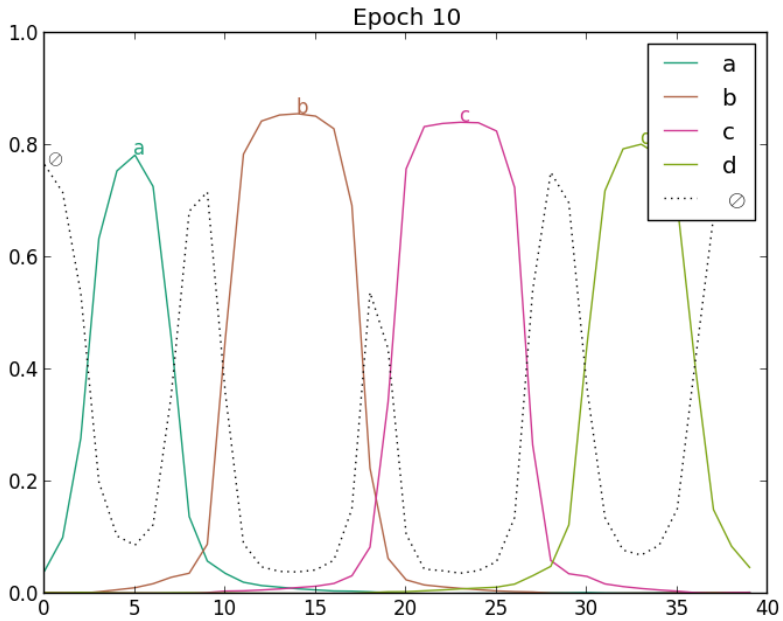
Entraînement CTC, exemple joué (2)

Exemple: descente de gradient sur les entrées de la CTC **avec ADAGRAD**



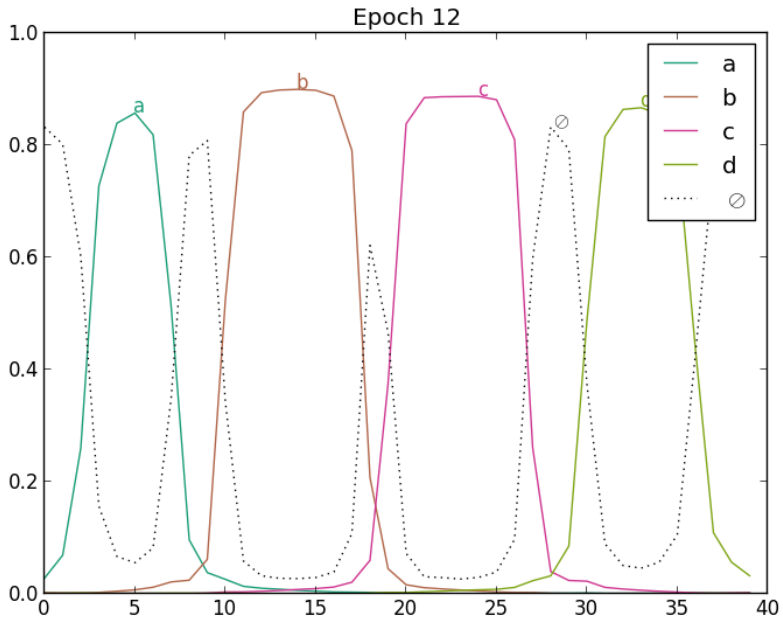
Entraînement CTC, exemple joué (2)

Exemple: descente de gradient sur les entrées de la CTC **avec ADAGRAD**



Entraînement CTC, exemple joué (2)

Exemple: descente de gradient sur les entrées de la CTC **avec ADAGRAD**



Conclusion

Introduction

Entraînement de Réseaux de Neurones pour les Systèmes Hybrides NN/HMM

Comparaison des Entraînements Framewise et CTC des MLPs et RNNs

L'intrigant Label Blank de la CTC

Conclusion

Conclusion

Nous avons étudié **l'entraînement CTC**, sa relation avec les entraînements framewise et HMM, et **le rôle du label blank** dans la CTC.

- L'entraînement CTC est similaire à:
  - l'entraînement framewise mais somme sur tous les alignments possibles
  - l'entraînement NN/HMM sans probabilités de transition et a priori
- Il ne se limite pas à un état par caractère + blank dans le cadre des systèmes hybrides NN/HMM, mais il est intéressant seulement dans ce cas
- CTC+Blank **marche particulièrement bien avec les RNNs**

Merci!

tb@a2ia.com

References

Bengio, Y., De Mori, R., Flammia, G., & Kompe, R. (1992). Global optimization of a neural network-hidden Markov model hybrid. Neural Networks, IEEE Transactions on, 3(2), 252--259.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. The Journal of Machine Learning Research, 12, 2121--2159.

Elagouni, K., Garcia, C., Mamalet, F., & Sébillot, P. (2012). Combining multi-scale character recognition and linguistic knowledge for natural scene text ocr. In Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on, (pp. 120--124). IEEE.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In International Conference on Machine learning, (pp. 369--376).

Haffner, P. (1993). Connectionist speech recognition with a global MMI algorithm. In EUROSPEECH.

Hennebert, J., Ris, C., Bourlard, H., Renals, S., & Morgan, N. (1997). Estimation of global posteriors and forward-backward training of hybrid HMM/ANN systems.

Konig, Y., Bourlard, H., & Morgan, N. (1996). Remap: Recursive estimation and maximization of a posteriori probabilities-application to transition-based connectionist speech recognition. Advances in Neural Information Processing Systems, (pp. 388--394).

Louradour, J., & Kermorvant, C. (2014). Curriculum learning for handwritten text line recognition. In Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on, (pp. 56--60). IEEE.

Rashid, S. F., Shafait, F., & Breuel, T. M. (2012). Scanning Neural Network for Text Line Recognition. In Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on, (pp. 105--109). IEEE.

Senior, A., & Robinson, T. (1996). Forward-backward retraining of recurrent neural networks. Advances in Neural Information Processing Systems, (pp. 743--749).

Tay, Y. H., Lallican, P.-M., Khalid, M., Knerr, S., & Viard-Gaudin, C. (2001). An analytical handwritten word recognition system with word-level discriminant training. In Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on, (pp. 726--730). IEEE.

Yan, Y., Fanty, M., & Cole, R. (1997). Speech recognition using neural networks with forward-backward probability generated targets. In Acoustics, Speech, and Signal Processing, IEEE International Conference on, vol. 4, (pp. 3241--3241). IEEE Computer Society.