Nantes ML Meetup

Question Answering
avec du Deep Learning

Yannis Flet-Berliac
Machine Learning Engineer @ iAdvize
@YannFletBerliac

6 nov. 2017

Sommaire

Introduction

“Intelligence Artificielle”: de quoi parle-t-on?

Input

- capteurs
- email
- image
- texte + question
- etc.

→

“Intelligence Artificielle”

→

Output

- mouvement
- spam: oui/non
- humain/chat/chien
- réponse
- etc.

Max Tegmark: intelligence is the “ability to accomplish complex goals”

Introduction

Machine Learning

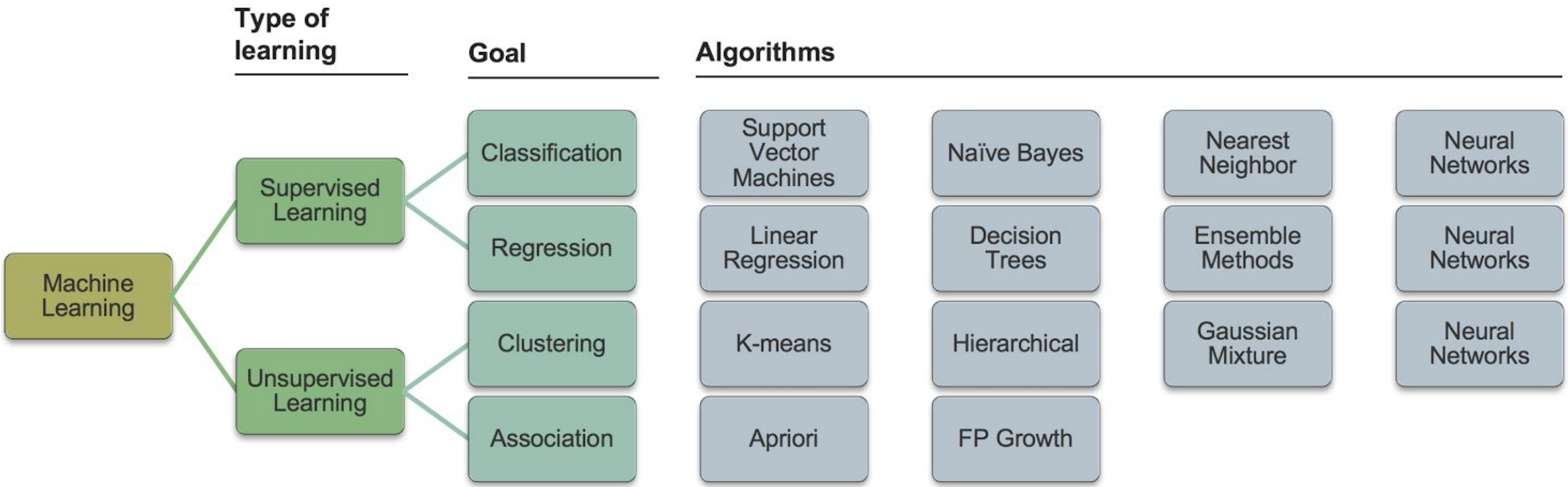
Apprentissage supervisé: entraînement de modèles à partir d'un ensemble de données étiquetées (ex: question answering)

Apprentissage non-supervisé: découverte de modèles à partir de données non étiquetées (ex: regrouper des documents similaires en fonction du contenu du texte [LDA])

+ **Apprentissage par renforcement**: apprentissage basé sur la récompense (ex: apprendre à jouer au Go en gagnant ou perdant des millions de parties)

Introduction

Machine Learning

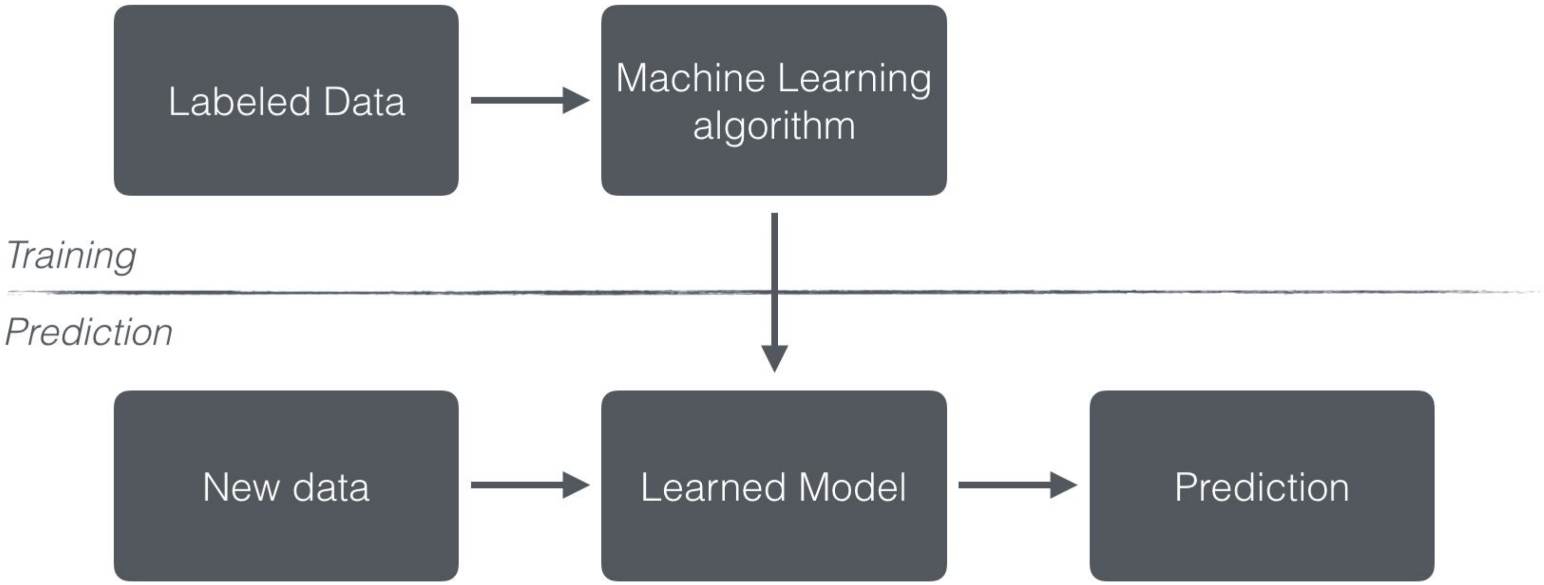


Source: Rita Waite

Introduction

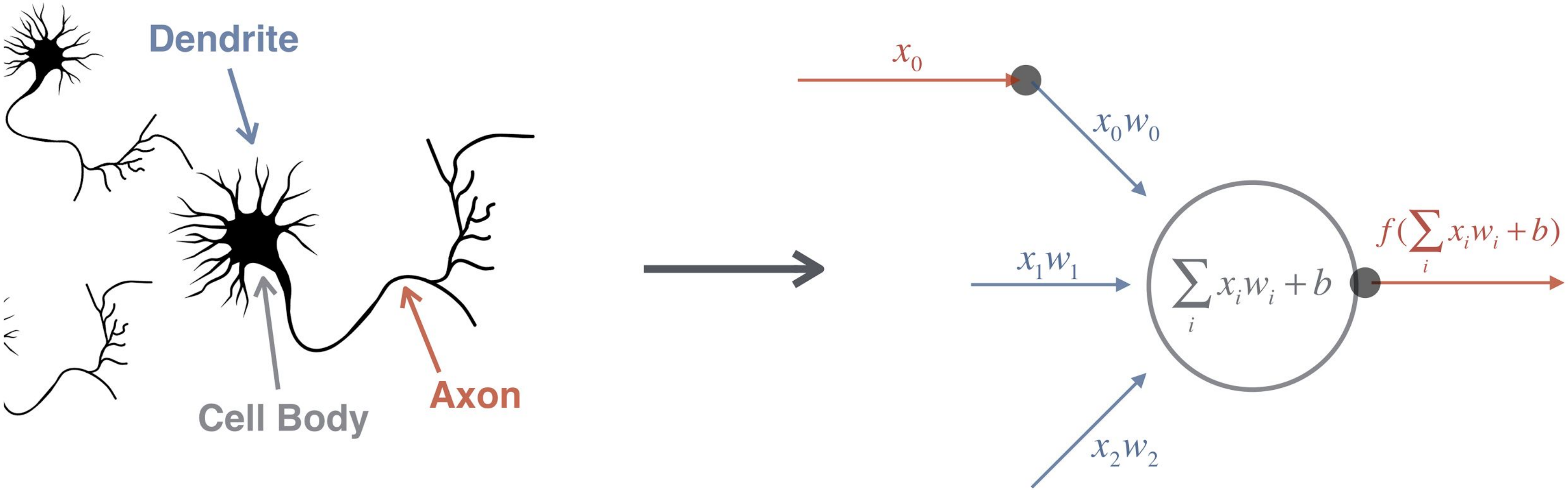
Machine Learning Framework

Apprentissage **supervisé**



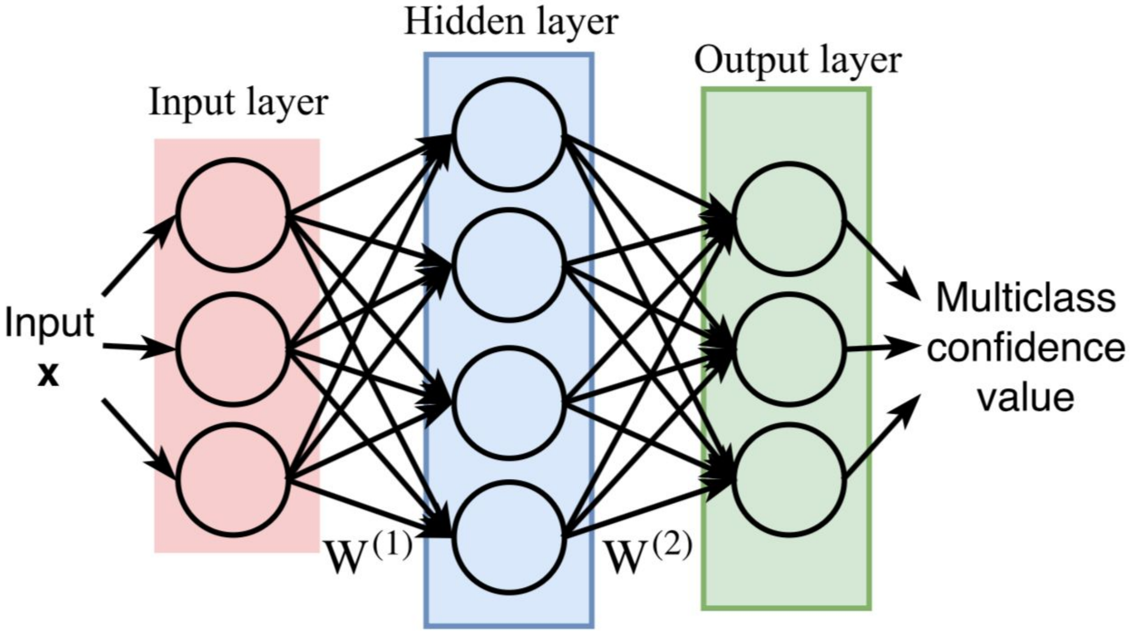
Introduction

Réseaux neuronaux, formulation



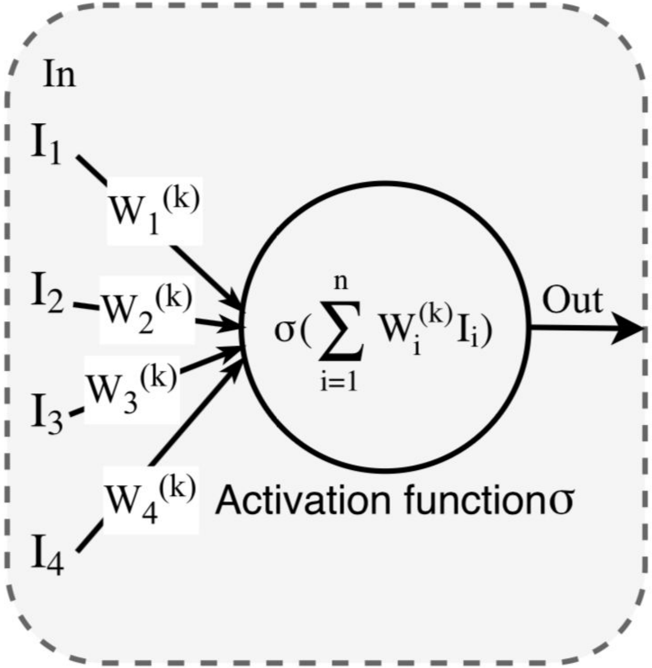
Introduction

Réseaux neuronaux



Function modeled by DNN:

$$f(x) = \sigma(W^{(2)} \cdot \sigma(W^{(1)} \cdot x))$$



Individual neurons in layer k

Introduction
Langage

Natural Language Processing (NLP): ensemble de systèmes pour des interactions de bout en bout (“end-to-end”) entre les humains et les machines

Natural Language Understanding (NLU): convertit des entrées non structurées en une représentation qui peut être interprétée par les machines

NLU with Deep Learning: a introduit d'énormes progrès dans la traduction automatique neuronale et le question answering

Introduction

Motivations

- Interfaces conversationnelles
  - Extraire une représentation d'environnement
  - Interagir en fonction des connaissances et du contexte
- Développements dans l’industrie du langage
  - Traduction
  - Droit
  - Diagnostics
  - Le savoir, la connaissance

Choix: ***Question Answering***

Objectifs

1. **Literature** review & choix de méthodes “**état de l’art**”
2. **Implémentation** de modèles QA:
    i. Recurrent Entity Network (**EntNet**)
    ii. Query-Reduction Network (**QRN**)
3. **Évaluer** les performances
4. Visualiser et comprendre le **processus d’apprentissage** (black-boxes)

Méthodologie

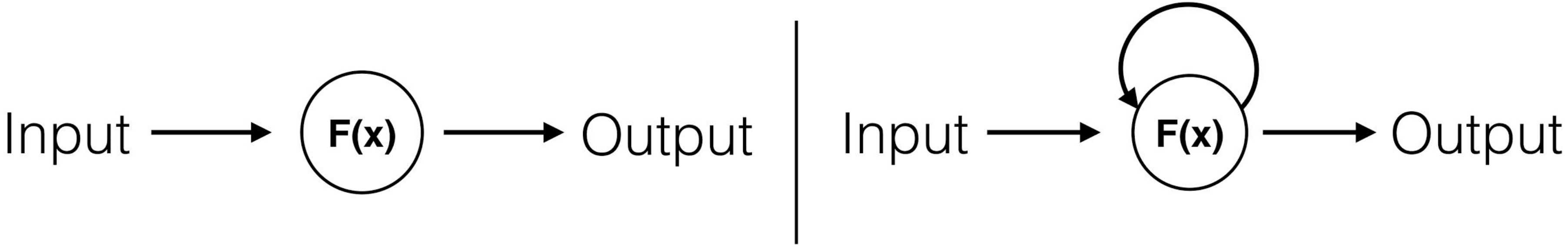
Le bAbI Dataset

- benchmark
- 20 tâches QA
- 1k ou 10k échantillons par tâche

Task 5: Three Argument Relations Mary gave the cake to Fred. Fred gave the cake to Bill. Jeff was given the milk by Bill. **Who gave the cake to Fred?** Mary **Who did Fred give the cake to?** Bill	**Task 6: Yes/No Questions** John moved to the playground. Daniel went to the bathroom. John went back to the hallway. **Is John in the playground?** no **Is Daniel in the bathroom?** yes	**Task 15: Basic Deduction** Sheep are afraid of wolves. Cats are afraid of dogs. Mice are afraid of cats. Gertrude is a sheep. **What is Gertrude afraid of?** wolves	**Task 16: Basic Induction** Lily is a swan. Lily is white. Bernhard is green. Greg is a swan. **What color is Greg?** white
Task 7: Counting Daniel picked up the football. Daniel dropped the football. Daniel got the milk. Daniel took the apple. **How many objects is Daniel holding?** two	**Task 8: Lists/Sets** Daniel picks up the football. Daniel drops the newspaper. Daniel picks up the milk. John took the apple. **What is Daniel holding?** milk, football	**Task 17: Positional Reasoning** The triangle is to the right of the blue square. The red square is on top of the blue square. The red sphere is to the right of the blue square. **Is the red sphere to the right of the blue square?** yes **Is the red square to the left of the triangle?** yes	**Task 18: Size Reasoning** The football fits in the suitcase. The suitcase fits in the cupboard. The box is smaller than the football. **Will the box fit in the suitcase?** yes **Will the cupboard fit in the box?** no

Méthodologie

Recurrent Neural Networks (RNN)

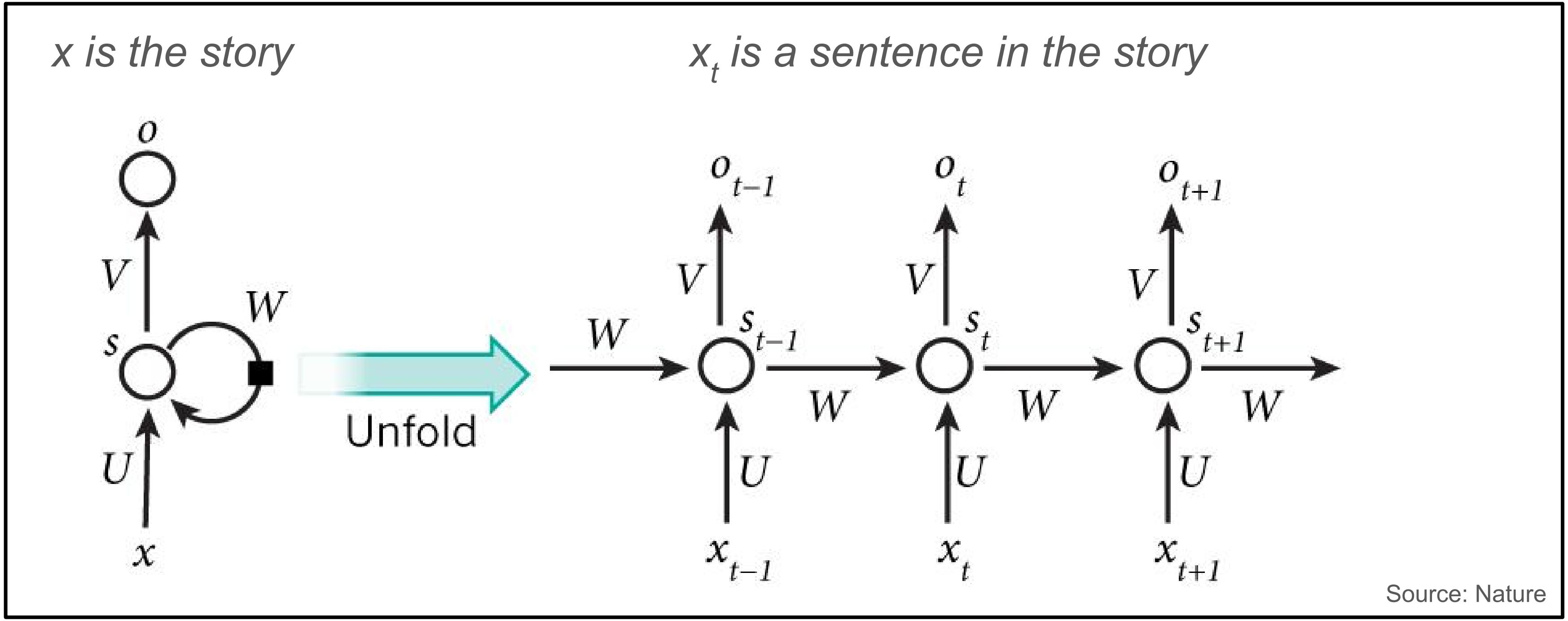


Feed-forward neuron

Recurrent neuron

Méthodologie

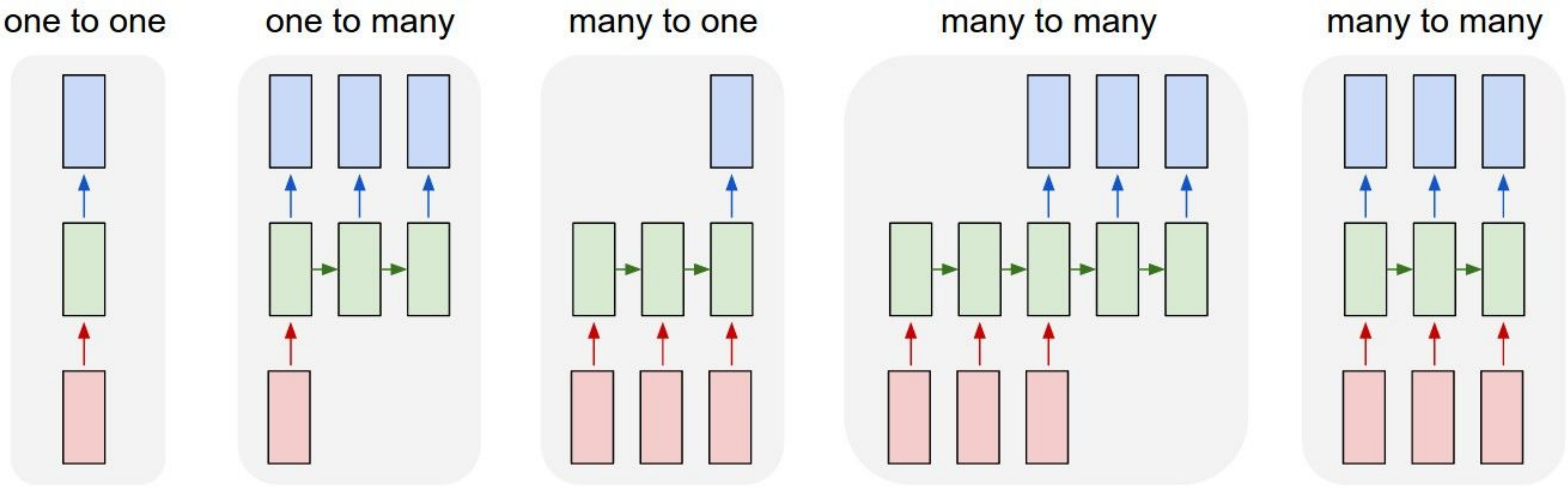
Recurrent Neural Networks (RNN)



L'architecture des **RNN** est particulièrement fondamentale aux modèles de QA.

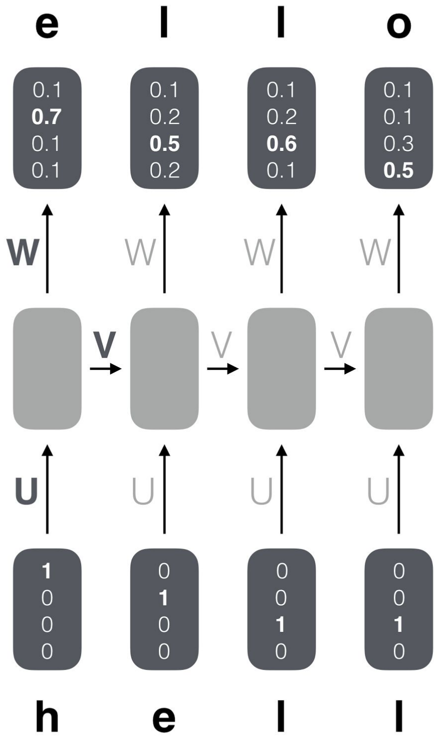
Méthodologie

Recurrent Neural Networks (RNN)



Méthodologie

RNN: exemple simple



$$y_t = \sigma(W * h_t + b_y) \qquad \sigma(x) = \frac{e^x}{\sum e^x}$$

$$h_t = tanh(U * X_t + V * h_{t-1} + b_h)$$

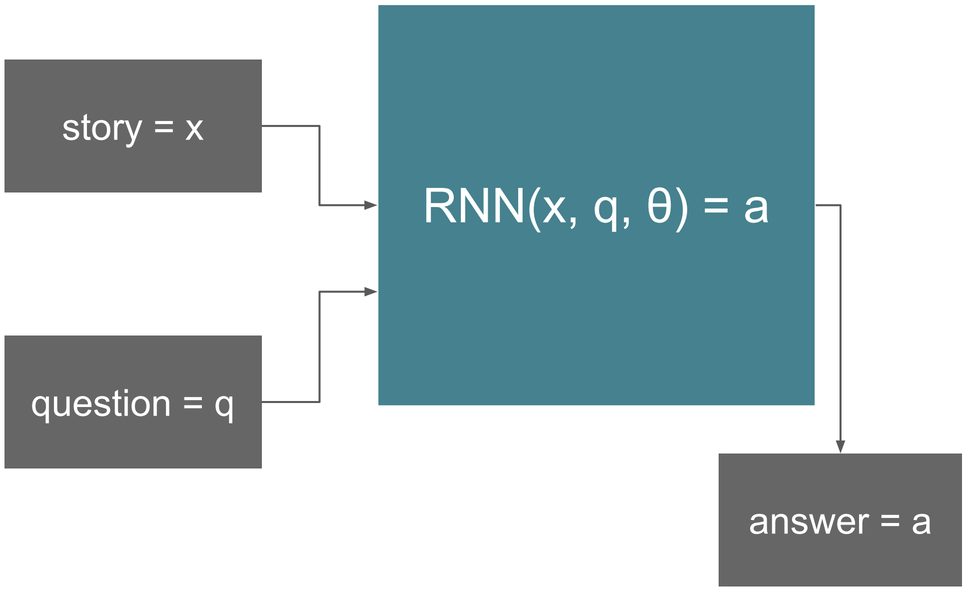
$$X_t$$

Méthodologie

Framework

Neural network-based
methods
+
QA (contexte supervisé)

- **story** (observations)
- **question** (trigger)
- **answer** (target)



Model overview

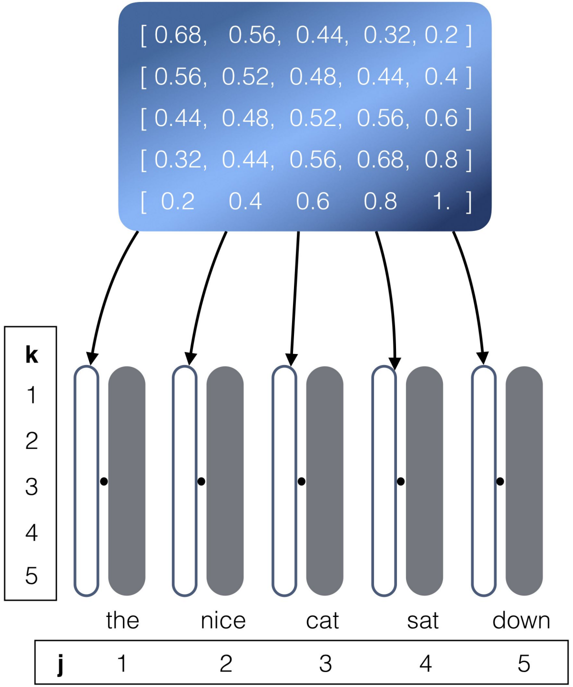
EntNet - Position Encoding

Ce module extrait:

- la représentation d'une phrase ***w***
- l'information ***l*** de l'ordre des mots dans la phrase

$$f_i = \sum_j l_j \cdot w_{ij}$$

$$l_{kj} = (1 - j/J) - (k/d)(1 - 2j/J)$$



Model overview

EntNet - Entity Memory Cells

Memory blocks:

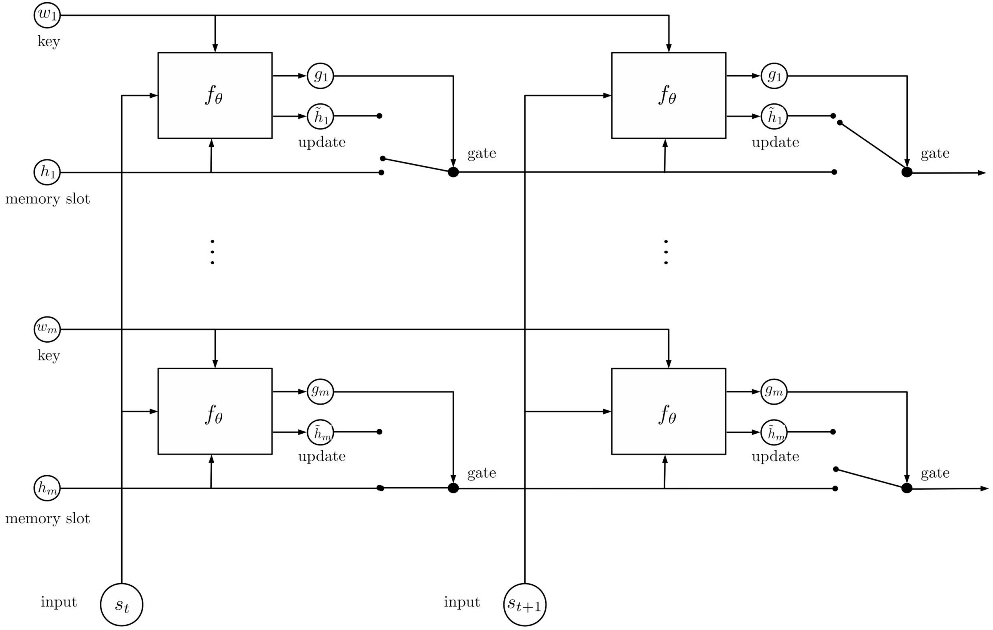
- s_t sentence *t* of a story
- h_j memory *j* from block *j*
- w_j key vector *j* from block *j*

<u>Memory update mechanism</u>

$$g_j = \sigma(s_t^T h_j + s_t^T w_j)$$

$$\widetilde{h_j} = \phi(U h_j + V w_j + W s_t)$$

Hidden state → $h_j = h_j + g_j \odot \widetilde{h_j}$



Model overview

EntNet

Memory update mechanism

$$g_j = \sigma(s_t^T h_j + s_t^T w_j)$$

$$\widetilde{h_j} = \phi(U h_j + V w_j + W s_t)$$

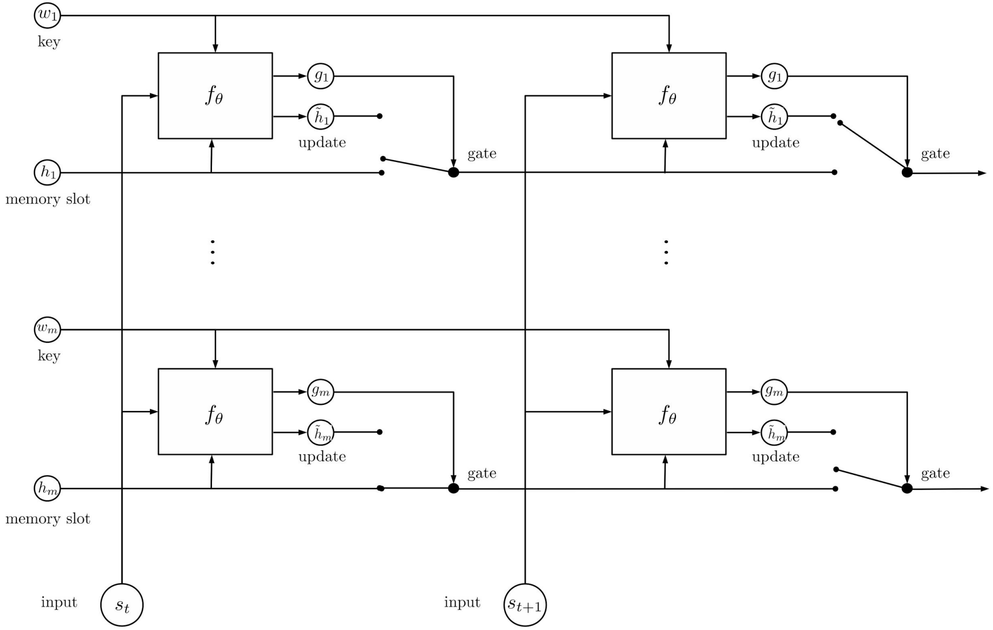
Hidden state → $h_j = h_j + g_j \odot \widetilde{h_j}$

Output module

$$p_j = \mathrm{Softmax}(q^T h_j)$$

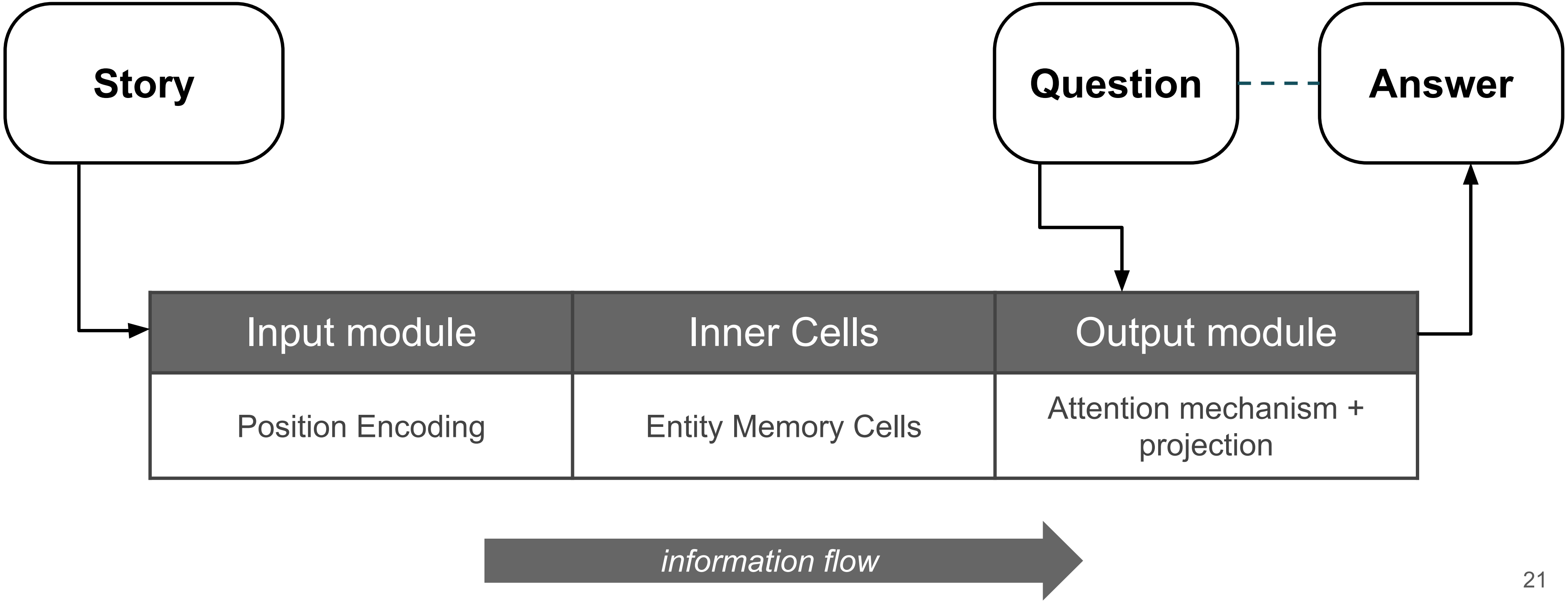
$$u = \sum_j p_j h_j$$

$$y = R\phi(q + Hu)$$



Model overview

Framework - EntNet



Implémentations
Apprentissage

Le modèle est très instable: faire plusieurs simulations (x3) avec différentes initialisations aléatoires

Tri: des textes les plus courts aux plus longs pour un apprentissage progressif

Régularisation: L2-norm, dropout

RNN bi-directionnels : permet des rétro-inférences

Implémentations

Training

- bAbI dataset divisé avec un ratio de **0.9** pour l'apprentissage et le test
- batch_size: **32** pour le dataset 1k et **128** pour le dataset 10k
- embedding size des modules d'entrée et de sortie: **50**
- fonction coût: **cross entropy**
- minimisée par **SGD** pour un maximum de **500** epochs
- learning rate contrôlé par **AdaGrad**

Un modèle réussit une tâche si son score est **supérieur à 95%**.

Experiments

EntNet

10k dataset : 13/20 tâches réussies

1k dataset :

Task ID	*Epochs*	*Train Loss*	*Train Acc.*	*Test Loss*	*Test Acc.*
Task 1	20	0.05	98.35	0.13	**95.97**
Task 2	200	0.17	97.96	3.51	32.71
Task 3	140	0.06	99.97	4.19	35.10
Task 4	20	0.00	100	0.00	**100**
Task 5	10	0.06	98.72	0.07	**98.86**
Task 6	50	0.01	99.78	0.16	**95.21**
Task 7	20	0.03	98.57	0.05	**98.23**
Task 8	20	0.02	99.19	0.07	**97.60**
Task 9	200	0.00	100	0.33	90.52
Task 10	40	0.01	99.85	0.17	**95.52**
Task 11	50	0.00	100	0.29	**95.31**
Task 12	20	0.00	100	0.01	**99.90**
Task 13	50	0.00	100	0.24	**95.52**
Task 14	200	0.00	100	0.63	79.06
Task 15	50	0.02	99.43	0.11	**95.42**
Task 16	200	0.49	80.78	0.88	47.60
Task 17	200	0.00	100	0.35	90.63
Task 18	40	0.06	96.73	0.07	**95.83**
Task 19	170	0.04	99.98	1.69	41.77
Task 20	20	0.03	98.43	0.03	**98.97**

Experiments

EntNet

10k dataset : 13/20 tâches réussies

1k dataset : 4/20 tâches réussies

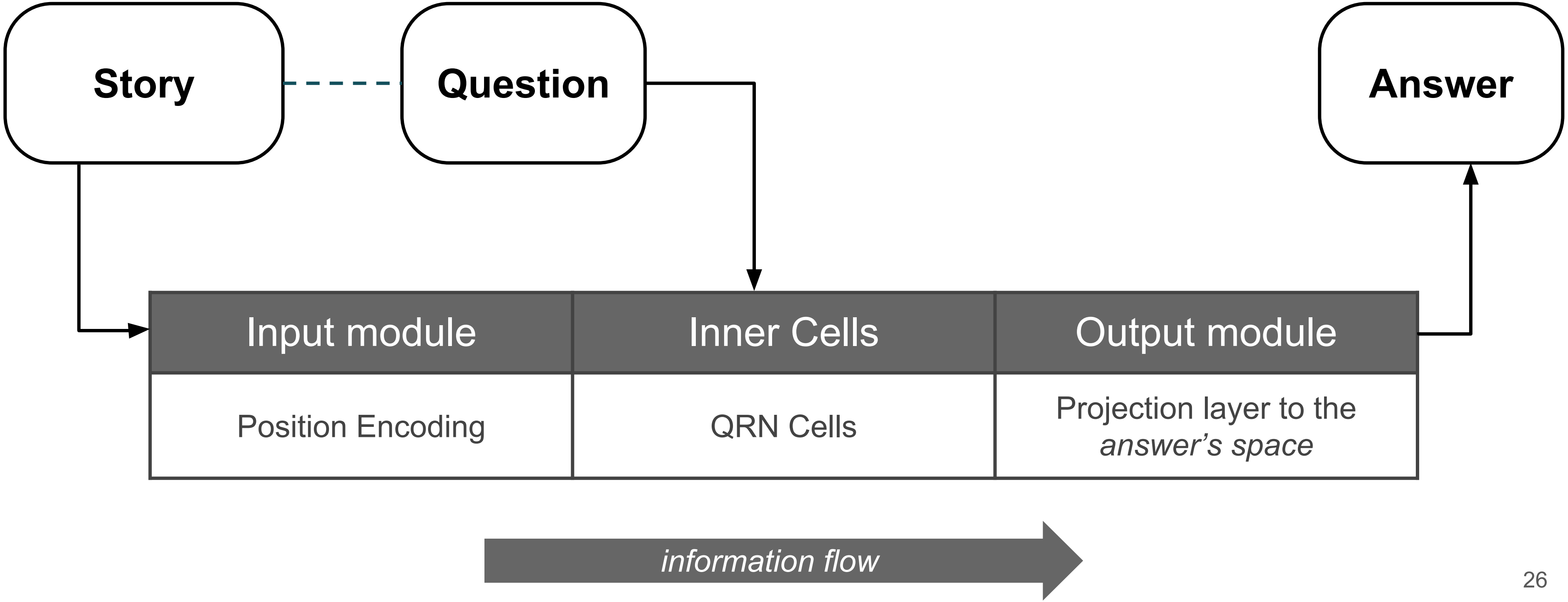
Performe mieux que, mais **est similaire** à d'autres modèles de réseaux de mémoire sur le même dataset.

Ces architectures sont trop complexes et souffrent de:

- sur-apprentissage (overfitting) sur les petits datasets
- vanishing gradient descent (sur les phrases et textes longs)

Models overview

Framework - QRN

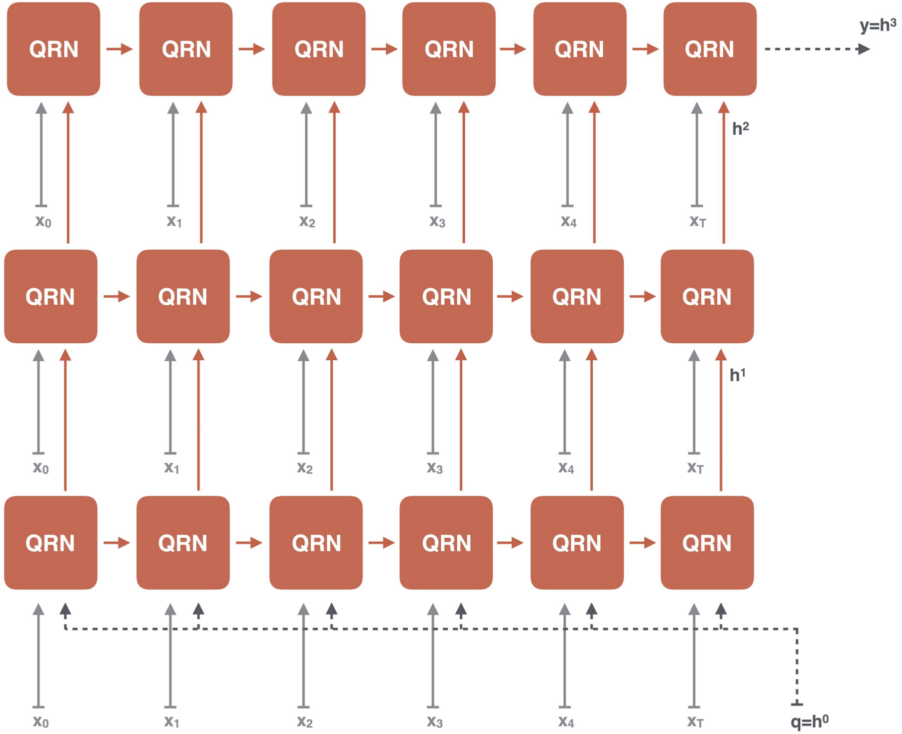


Models overview

QRN Cells

THE SLIDE WAS CORRECT!
(sorry for the blurred explanation following the question, I focused too much on the 3rd line)
h_t is indeed updated using z_t the update gate and h_{t-1} the reduced query at time-step t-1.
What is crucial here is that the gate z_t controlling the update is computed on the current memory slot **only**. Hence the gate is **locally defined** and allows the QRN cell to avoid vanishing gradient across the story when updating the hidden state in the 3rd equation.
If this raises more questions, don't hesitate to contact me: yannisfbe@gmail.com :-)

Réduction de la *question* avec chaque représentation de *phrase*.



QRN cell mechanism

$$z_t = \alpha(x_t, q_t) = \sigma(W^{(z)}(x_t \circ q_t) + b^{(z)})$$

$$\tilde{h}_t = \rho(x_t, q_t) = \tanh(W^{(h)}[x_t; q_t] + b^{(h)})$$

Hidden state → $h_t = z_t\tilde{h}_t + (1 - z_t)h_{t-1}$

Models overview

QRN

QRN cell mechanism

$$z_t = \alpha(x_t, q_t) = \sigma(W^{(z)}(x_t \circ q_t) + b^{(z)})$$

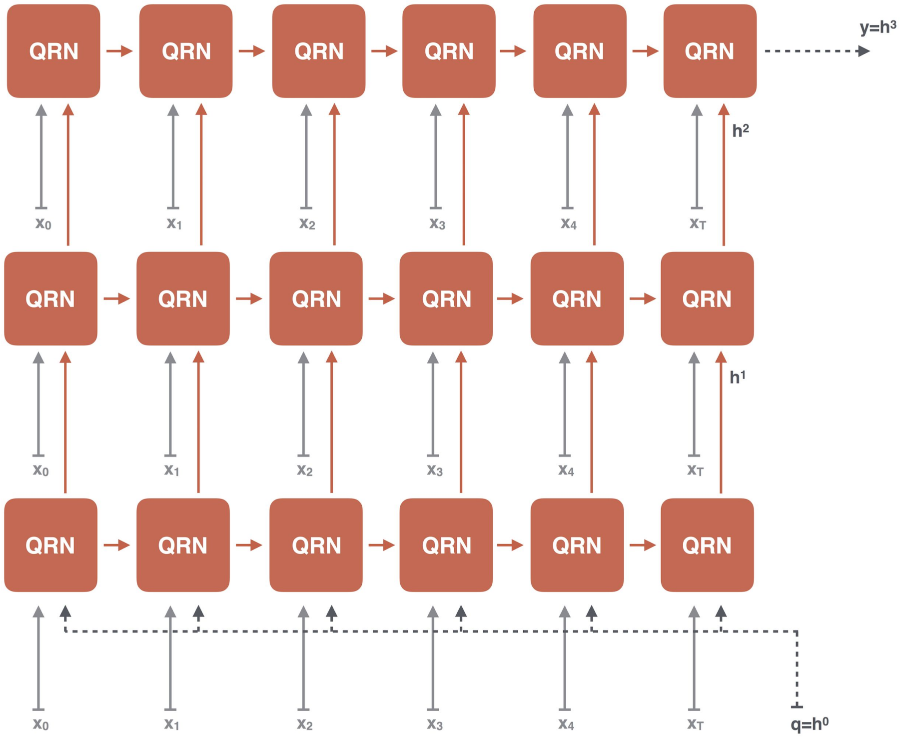
$$\tilde{h}_t = \rho(x_t, q_t) = \tanh(W^{(h)}[x_t; q_t] + b^{(h)})$$

Hidden state $\rightarrow$ $h_t = z_t \tilde{h}_t + (1 - z_t) h_{t-1}$

Output module

$$\hat{y} = softmax(W^y \hat{h})$$

+ argmax



Experiments

QRN

1k dataset : 15/20 tâches réussies

Pour rappel, **EntNet:**

10k dataset : 13/20 tâches réussies

Task ID	*Epochs*	*Train Loss*	*Train Acc.*	*Test Acc.*
Task 1	500	0.07	100	**100**
Task 2	500	0.29	100	**100**
Task 3	560	0.82	99.21	**96.43**
Task 4	500	0.14	100	**100**
Task 5	500	0.09	100	**98.79**
Task 6	500	0.07	100	**100**
Task 7	700	0.12	99.98	**95.87**
Task 8	500	0.15	99.88	**99.20**
Task 9	700	0.06	100	**97.68**
Task 10	500	0.08	100	**97.78**
Task 11	500	0.07	100	**100**
Task 12	500	0.10	100	**100**
Task 13	500	0.12	99.88	**96.78**
Task 14	500	0.16	100	82.76
Task 15	500	0.07	100	**100**
Task 16	500	0.43	100	46.17
Task 17	500	0.31	96.65	51.62
Task 18	500	0.16	95.09	92.74
Task 19	500	0.67	100	15.63
Task 20	500	0.11	98.44	**97.98**

Experiments

QRN

1. Nos résultats sont proches des résultats du papier: **11,43** au lieu de **11,3** (average loss score).
2. Surpasse les autres modèles basés sur la mémoire pour la petite et grande version du bAbI dataset.
3. $\tilde{h}_t$ **ne dépend pas** de l'état caché précédent.
4. La cellule QRN est à la fois un **mécanisme d'attention** et un **RNN**.

<u>QRN cell</u>

$$z_t = \alpha(x_t, q_t) = \sigma(W^z(x_t \circ q_t) + b^z)$$

$$r_t = \beta(x_t, q_t) = \sigma(W^r(x_t \circ q_t) + b^r)$$

$$\tilde{h}_t = \rho(x_t, q_t) = \tanh(W^h[x_t; q_t] + b^h)$$

$$h_t = z_t r_t \tilde{h}_t + (1 - z_t) h_{t-1}$$

Visualize the learning

QRN

Cibler une entité (John)

70	Where is the apple	02	1. Sandra moved to the garden 2. Daniel took the apple there 3. Daniel dropped the apple 4. Sandra travelled to the office 5. John moved to the kitchen 6. Sandra moved to the garden 7. Daniel went back to the bathroom 8. Mary travelled to the hallway 9. Sandra went to the bathroom 10. John went back to the bedroom 11. Mary moved to the garden 12. John picked up the apple there 13. John went back to the kitchen 14. John went back to the garden	yp = garden y = garden

facts	1			2			3		
	z	rf	rb	z	rf	rb	z	rf	rb
Sandra moved to the garden	0.01	0.98	0.27	0.00	0.54	0.35	0.03	0.00	0.00
Daniel took the apple there	0.98	0.81	0.94	0.92	0.90	0.96	0.00	0.00	0.00
Daniel dropped the apple	0.97	0.65	0.94	0.99	0.09	0.89	0.04	0.00	0.00
Sandra travelled to the office	0.06	0.92	0.31	0.00	0.31	0.29	0.02	0.00	0.00
John moved to the kitchen	0.04	0.84	0.26	0.00	0.38	0.52	0.94	0.00	0.00
Sandra moved to the garden	0.01	0.98	0.27	0.00	0.33	0.34	0.05	0.00	0.00
Daniel went back to the bathroom	0.01	0.68	0.19	0.00	0.21	0.33	0.03	0.00	0.00
Mary travelled to the hallway	0.04	0.84	0.19	0.00	0.38	0.30	0.07	0.00	0.00
Sandra went to the bathroom	0.00	0.94	0.18	0.00	0.55	0.21	0.02	0.00	0.00
John went back to the bedroom	0.03	0.85	0.27	0.00	0.15	0.40	0.91	0.00	0.00
Mary moved to the garden	0.01	0.99	0.29	0.00	0.19	0.43	0.09	0.00	0.00
John picked up the apple there	0.98	0.84	0.95	0.99	0.95	0.99	0.00	0.00	0.00
John went back to the kitchen	0.09	0.61	0.27	0.02	0.62	0.43	0.94	0.00	0.00
John went back to the garden	0.03	0.86	0.31	0.06	0.44	0.47	0.93	0.00	0.00

Visualize the learning

QRN

Motivations

885	Why did yann go to the garden	20	1. Yann is bored 2. Antoine is tired 3. Yann moved to the garden	(see facts table below)	yp = bored y = bored

facts	**1**			**2**		
	z	rf	rb	z	rf	rb
Yann is bored	0.62	0.85	0.77	0.96	0.00	0.00
Antoine is tired	0.12	0.54	0.48	0.78	0.00	0.00
Yann moved to the garden	0.76	0.98	0.95	1.00	0.00	0.00

Connaissances

883	Where will yann go	20	1. Yann is bored	(see facts table below)	yp = garden y = garden

facts	**1**			**2**		
	z	rf	rb	z	rf	rb
Yann is bored	1.00	0.96	0.91	1.00	0.00	0.00

884	Where will antoine go	20	1. Yann is bored 2. Antoine is tired	(see facts table below)	yp = bedroom y = bedroom

facts	**1**			**2**		
	z	rf	rb	z	rf	rb
Yann is bored	0.87	0.77	0.75	0.99	0.00	0.00
Antoine is tired	0.98	0.90	0.86	0.97	0.00	0.00

Conclusion

- **Memory-based networks (EntNet)** extraient des “vecteurs” de mémoire pertinents avant de les confronter à la question. Mais ils nécessitent énormément de paramètres.
- **QRN** utilise une cellule récurrente à porte pour:
  - extraire à chaque fenêtre de temps une représentation des faits pertinents d’une histoire
  - tirer parti de la capacité récurrente des RNN pour modéliser des données séquentielles
  - évite le problème de dépendance à long terme sur les longs textes
- **RNN** manifeste de fortes performances pour modéliser le langage:
  - question answering
  - machine translation
  - speech recognition

Future works

- **Real-world** datasets:
  - SQuAD
  - NewsQA

- Module d’entrée **hybride:**
  - character-level and word-level
  - le reste du modèle est partagé entre les modules d’entrée

References

M. Seo et al., “Query-Reduction Networks for Question Answering”, June 2016. (https://arxiv.org/pdf/1606.04582.pdf)

M. Henaff, Y. LeCun et al., “Tracking the world state with recurrent entity networks”, 2016. (https://arxiv.org/pdf/1612.03969.pdf)

A. Kumar et al., “Ask me anything: Dynamic memory networks for natural language processing”, International Conference on Machine Learning, pp. 1378–1387, 2016. (https://arxiv.org/pdf/1506.07285.pdf)

S. Sukhbaatar, J. Weston et al., “End-to-end memory networks”, Advances in neural information processing systems, pp. 2440–2448, 2015. (https://arxiv.org/pdf/1503.08895.pdf)

X. Zhang, Y. LeCun, “Text Understanding from Scratch”, 2015. (https://arxiv.org/pdf/1502.01710.pdf)

Merci.

Avez-vous des questions?