UNIVERSITÉ LAVAL

GLO-4030/7030
APPRENTISSAGE PAR RÉSEAUX DE NEURONES PROFONDS

Réseaux Récurrents (RNN)

Philippe Giguère

Pourquoi ?

- Traiter des données séquentielles

  Image X vs. $\{x^{(1)}, x^{(2)}, \cdots, x^{(\tau)}\}$

  – séries temporelles

  – séquences de pixels

- Souvent de longueur *variable*
- Pas clair d'avance où l'information pertinente est située

I went to Nepal in 2009

In 2009, I went to Nepal

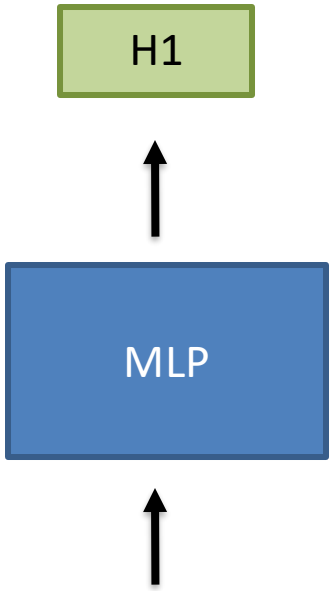
Exemple textuel

```
I like whisky
```

Exemple textuel

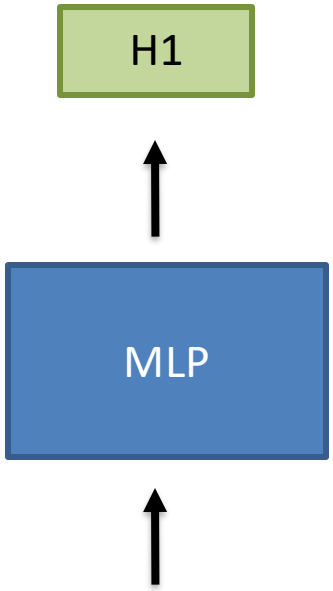
```
[I, like, whisky]
```

Exemple textuel



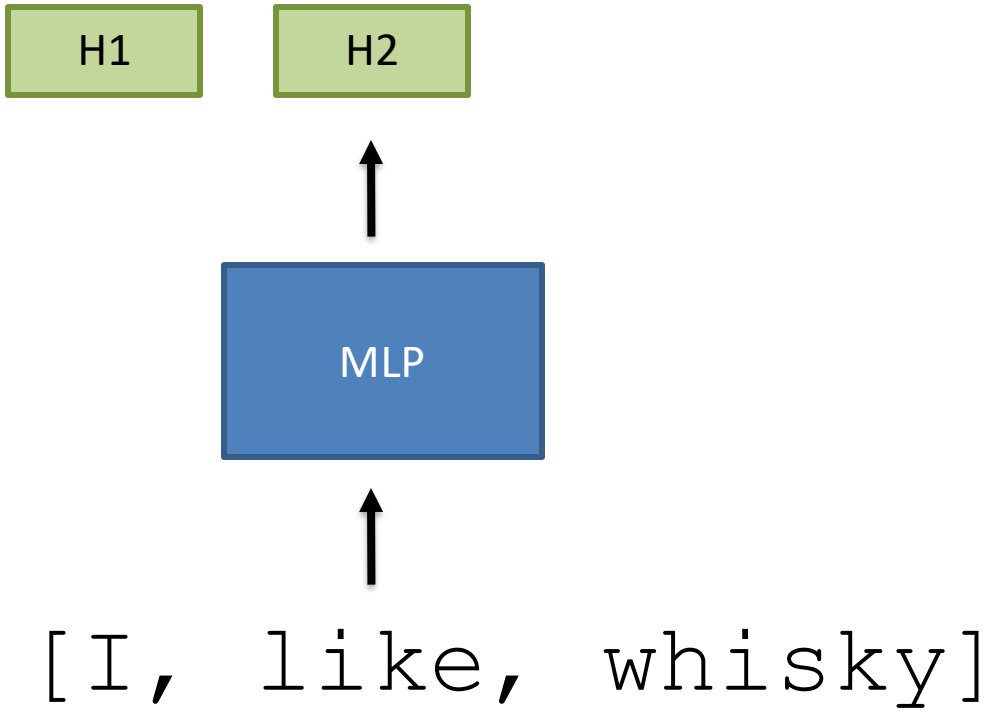
[I, like, whisky]

Exemple textuel

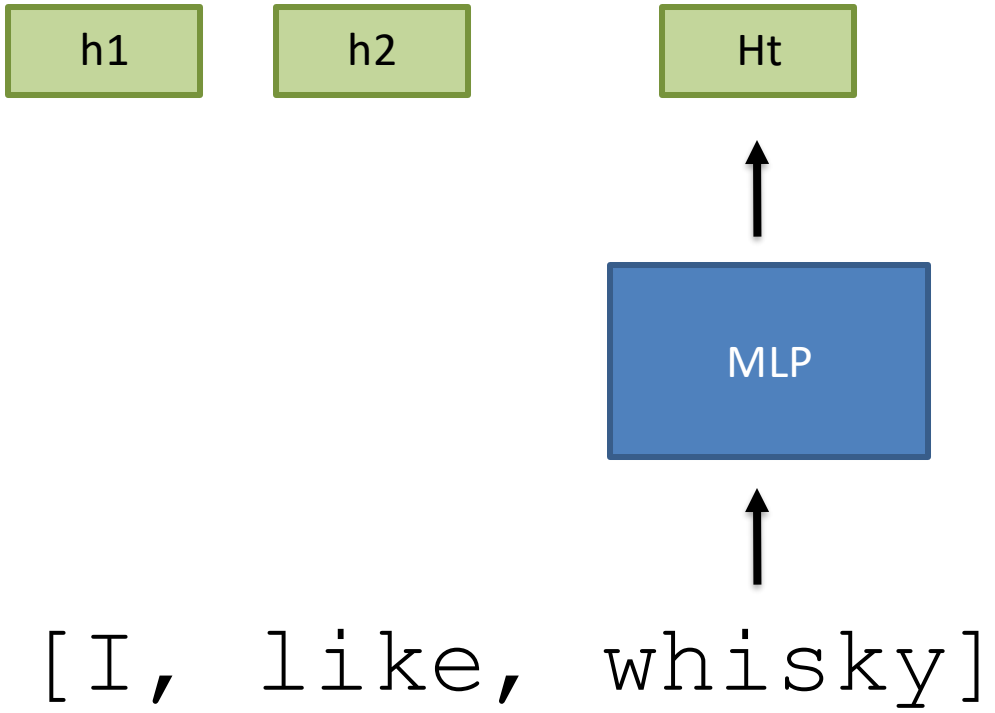


[I, like, whisky]

Exemple textuel



Exemple textuel

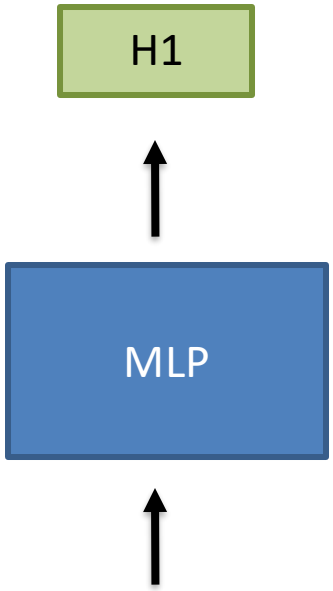


Exemple textuel

Ne correspond pas vraiment à ce qu'on désire....

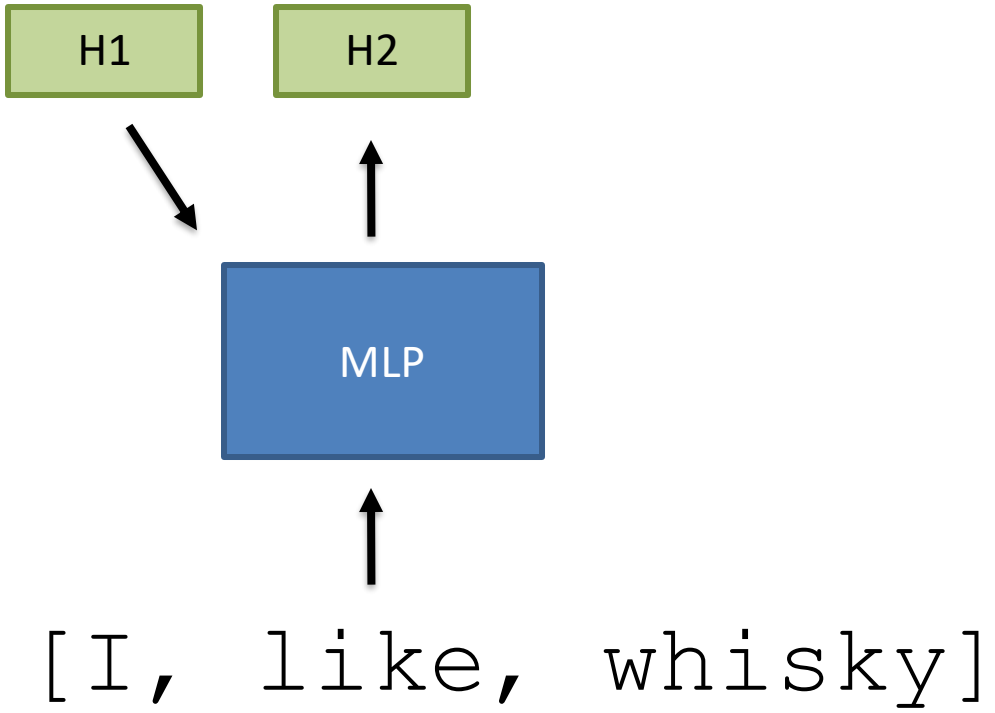
```
[I, like, whisky]
```

Vers le réseau récurrent...

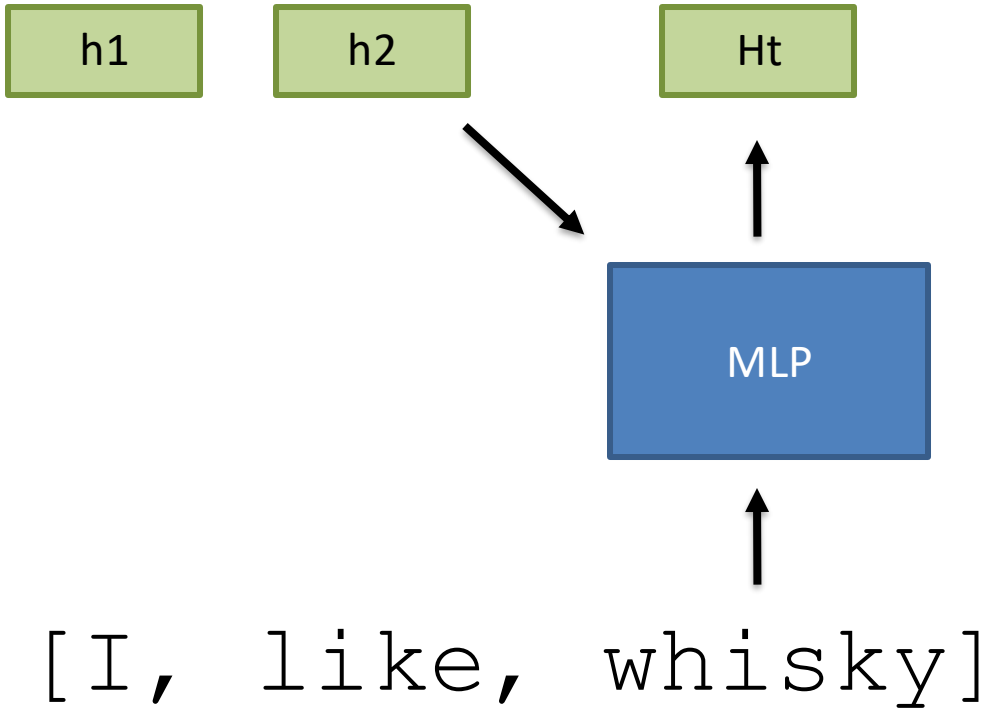


[I, like, whisky]

Vers le réseau récurrent...



Vers le réseau récurrent...



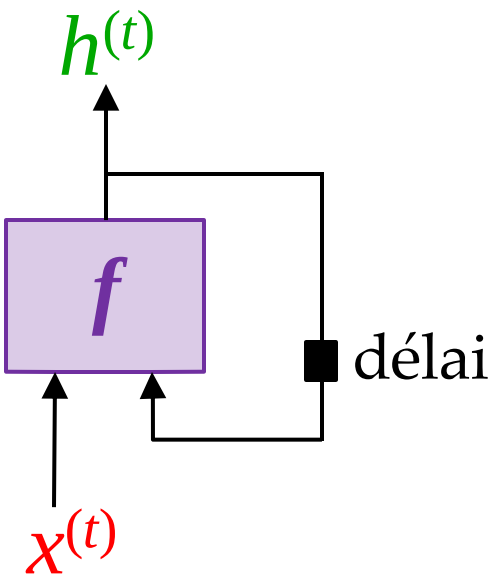
Vers le réseau récurrent...

Maintenant Ht va contenir de l'information sur toute la séquence.

`[I, like, whisky]`

Idée générale

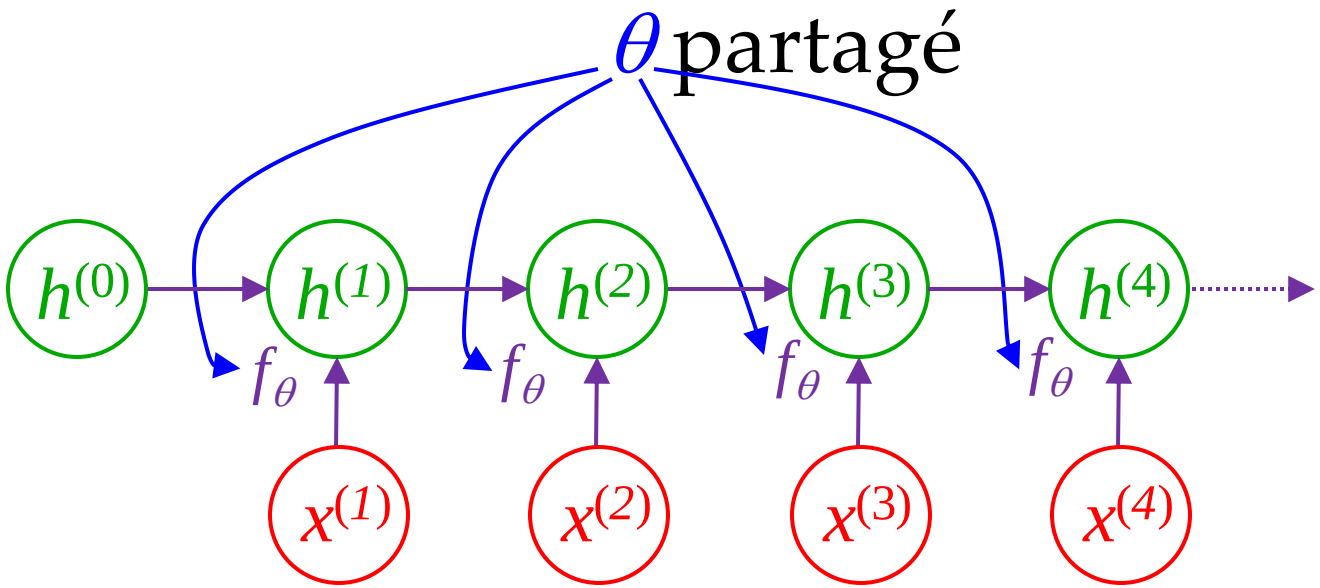
$$h^{(t)}=f(h^{(t-1)}, x^{(t)}, \theta)$$



- Même θ (*weight sharing*)
- Limite le pouvoir de représentation
  - régularisation
- Relation f stationnaire : ne change pas selon t
  - p. e. règle grammaire indépendante de la position
- Lien avec systèmes dynamiques (GMC, GEL)

Graphe calcul déroulé

$$h^{(t)}=f(h^{(t-1)}, x^{(t)}, \theta)$$



Variable cachée *h*

- **Résumé sémantique** (avec perte) de la séquence (passée, si causal)
- En lien direct avec la tâche :
  - p. e. si on cherche des dates, des mots comme **`mercredi`** vont influencer *h* plus que **`Québec`**
  - backprop fera le travail de trouver la fonction *f* favorisant cette représentation
- Taille de *h* influencera la quantité d'information pouvant y être stockée
  - pourra difficilement résumer *À la recherche du temps perdu* de M. Proust (4 215 pages)
  - généralisation plus difficile si *h* est grand

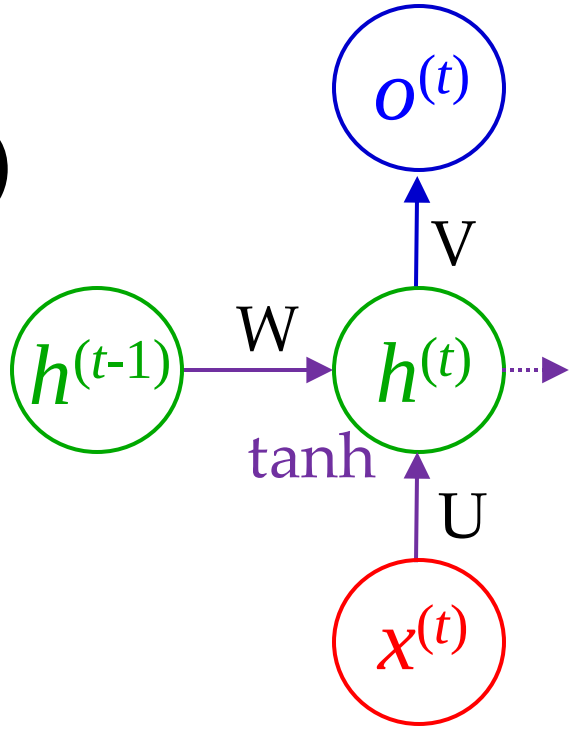
RNN universel (vanille)

- Utilise des fonctions affines
- tanh comme non-linéarité

$$h^{(t)} = \tanh(Wh^{(t-1)} + Ux^{(t)} + b)$$

$$o^{(t)} = Vh^{(t)} + c$$

- Peut accomplir autant qu'une machine de Turing
- Plus commune
- Défaut : on ne peut pas paralléliser forward/backward pass
  - doit faire la séquence au complet en sériel



RNN universel (vanille)

Combien de paramètres pour:

- Un hidden state de taille 10
- Une input de taille 5
  - W=
  - U =
  - B =
  - Total =

$$h^{(t)} = \tanh(Wh^{(t-1)} + Ux^{(t)} + b)$$

Pourquoi tanh ?

- Non-linéaire
- Toujours dérivable
- Sortie [-1,1] (enlever/ajouter)
- Symétrique
- Pas de biais systématique
  - sigmoïde va de [0,1], induit biais
- Autres ?

Topologies RNN

one to many

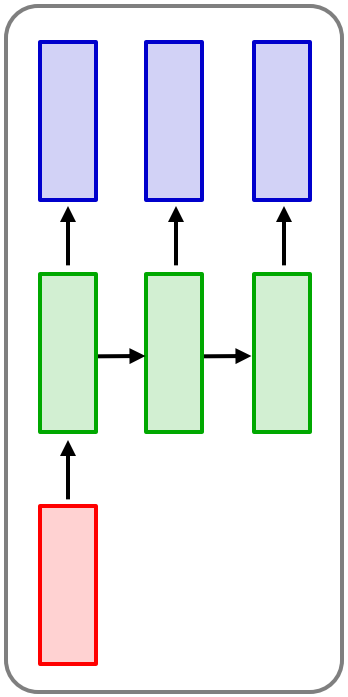
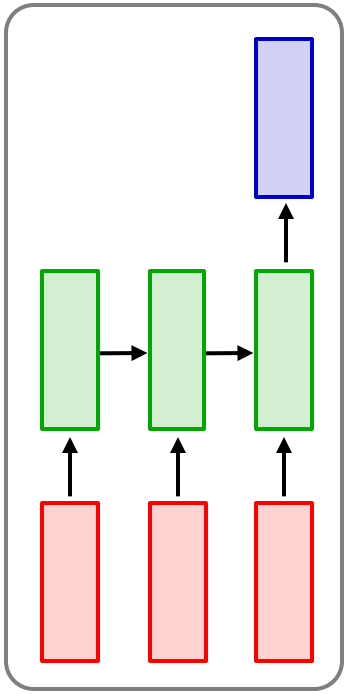


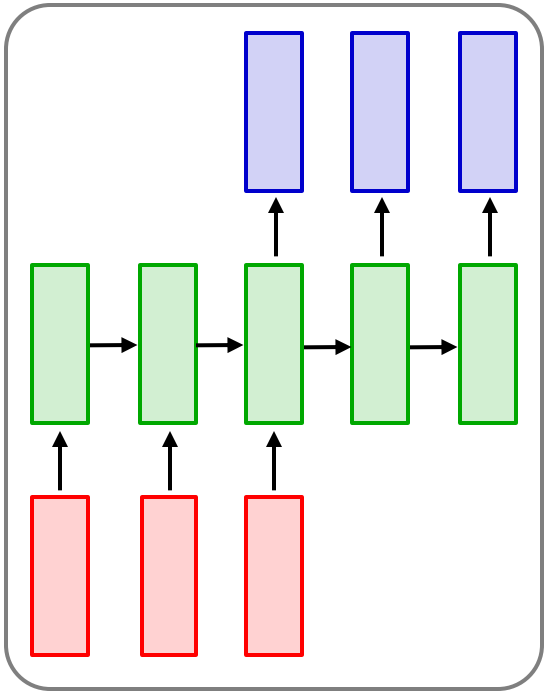
Image captioning

many to one



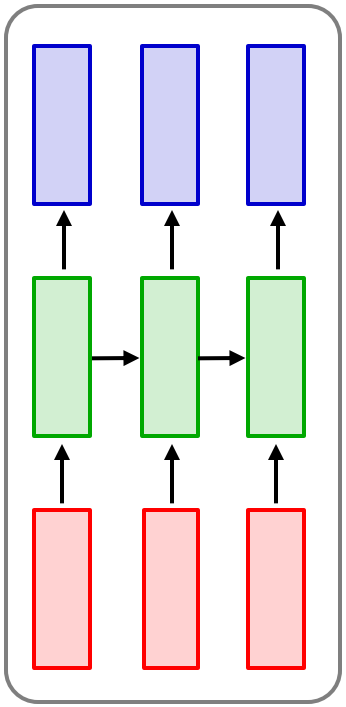
Classification de sentiment (texte)

many to many



Traduction, Réponse aux questions

many to many



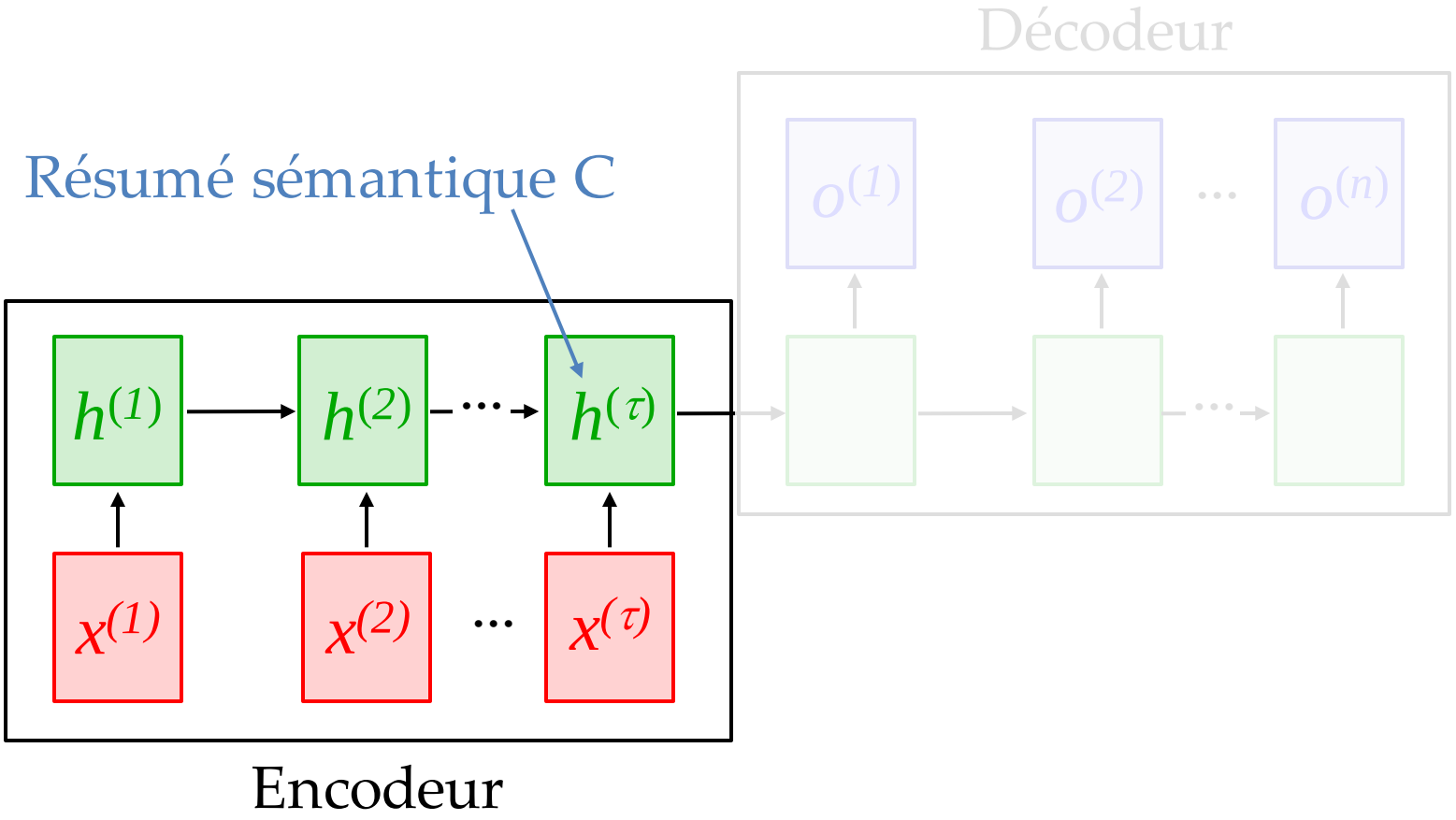
Classification de trames vidéos

adapté de cs231n

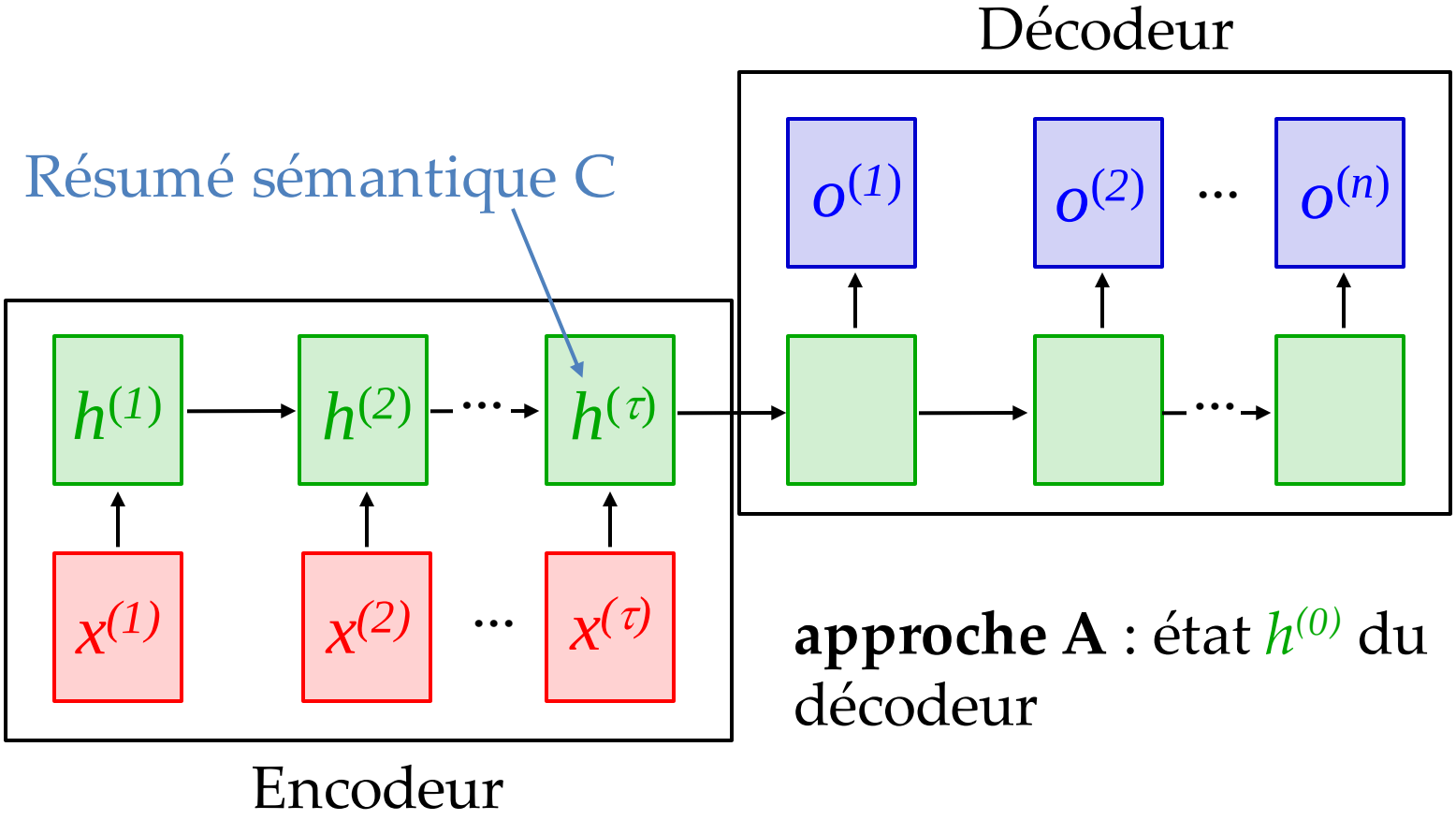
Sequence-to-sequence

- Architecture *many to many*
- Généré une séquence à partir d'un résumé C

contexte

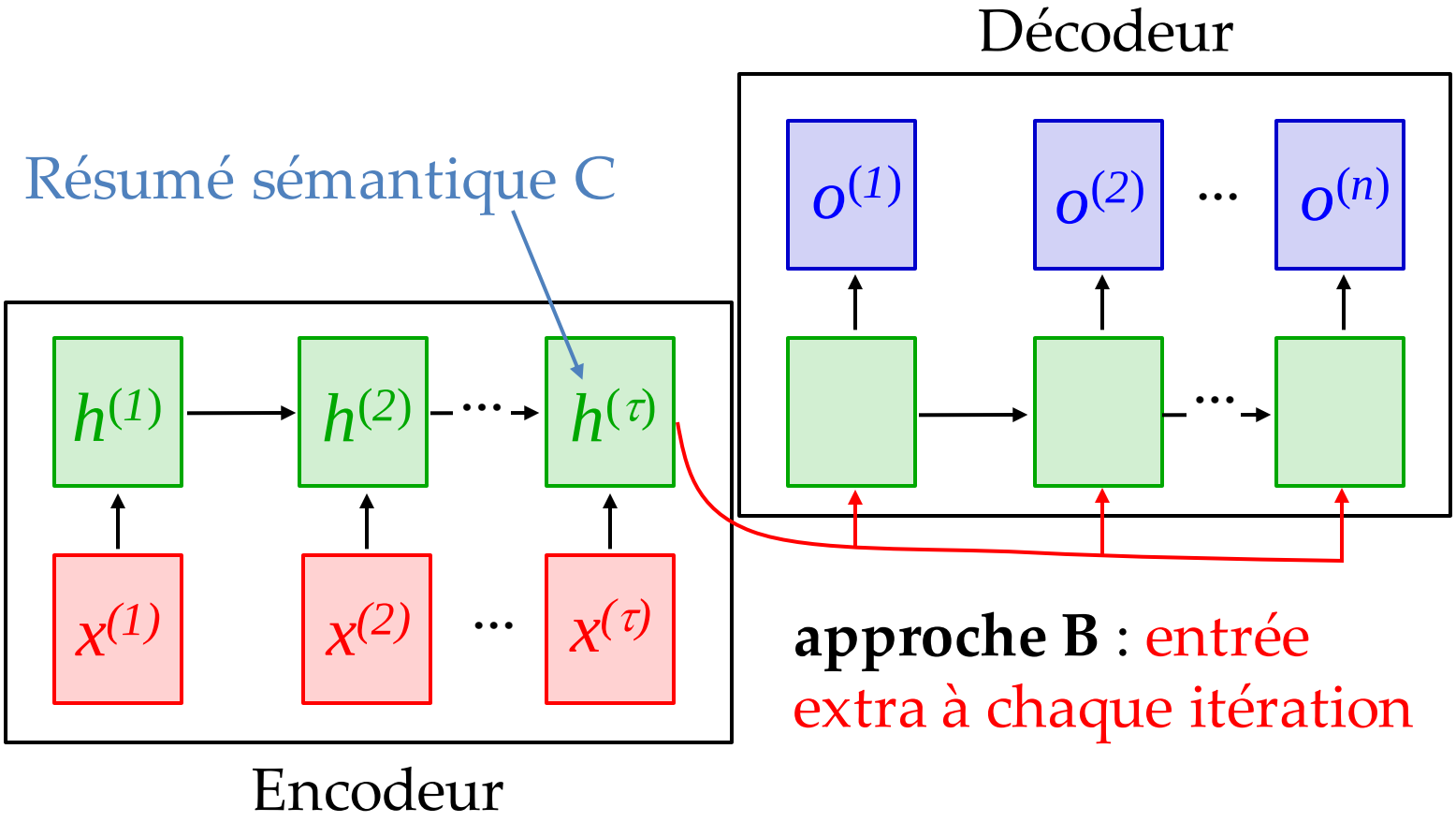


Sequence-to-sequence



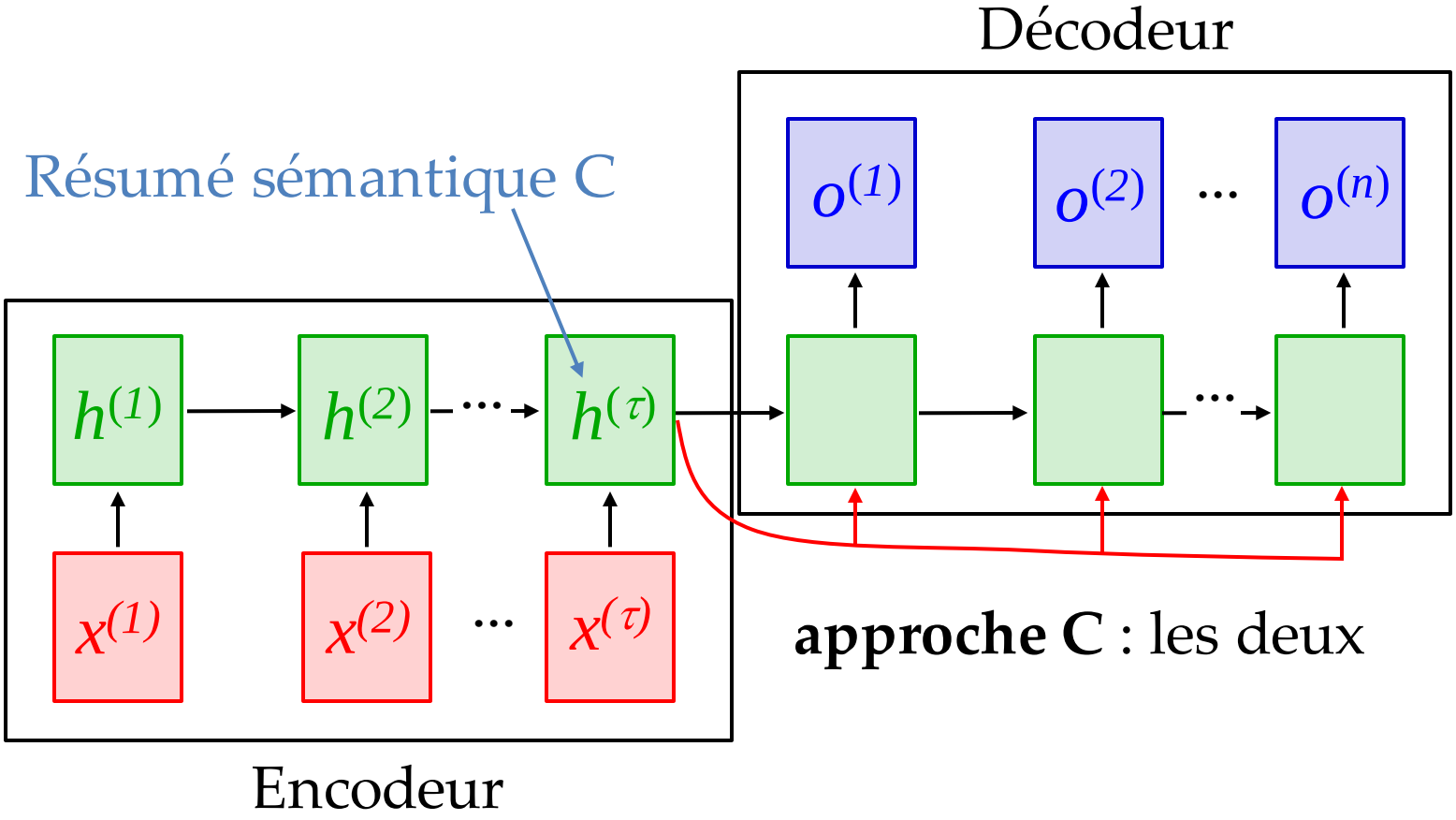
approche A : état $h^{(0)}$ du décodeur

Sequence-to-sequence



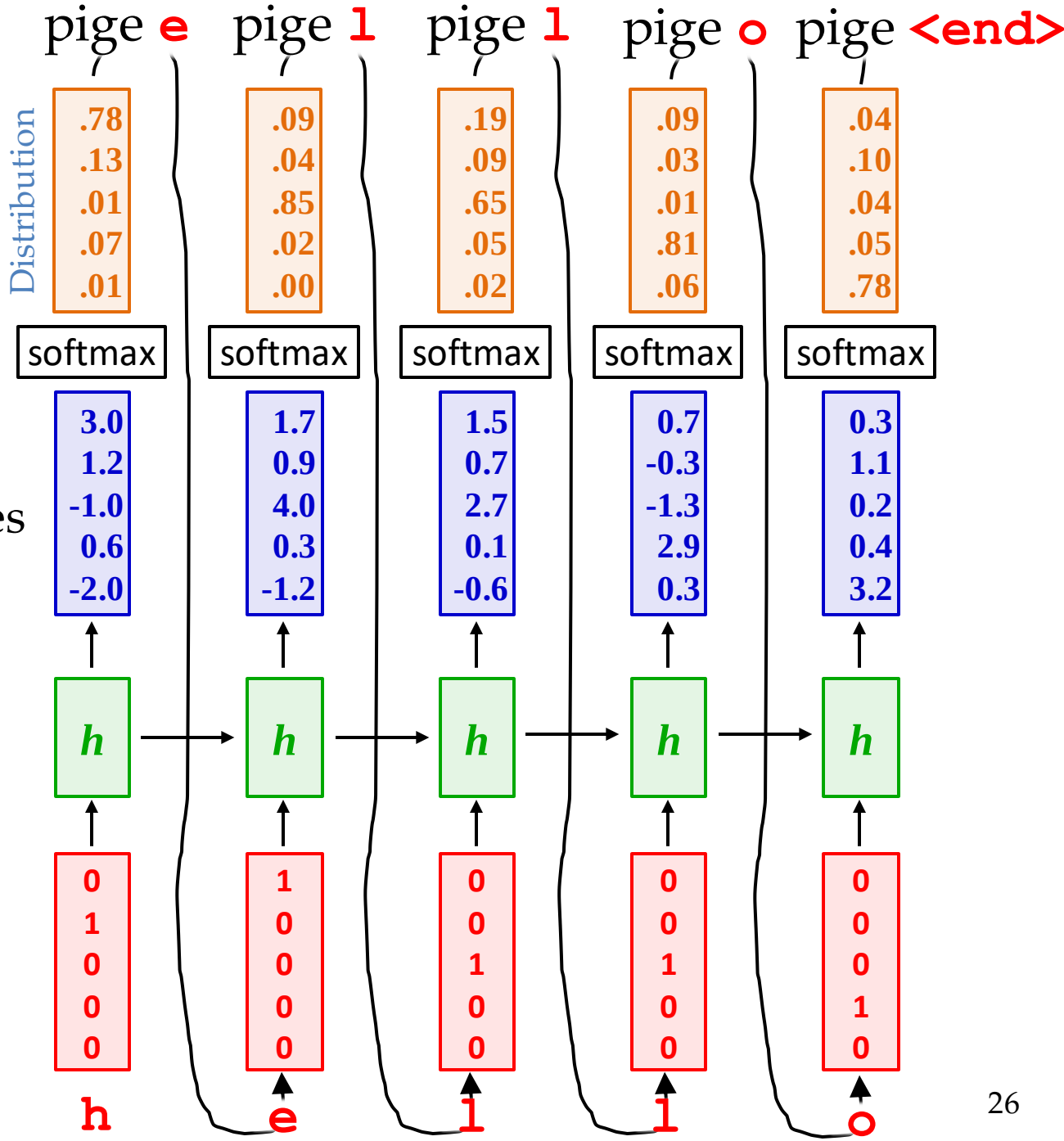
approche B : entrée extra à chaque itération

Sequence-to-sequence



Exemple de génération avec RNN

- Réseau entraîné à prédire des caractères `{e,h,l,o,<end>}`
- Entraîné sur `hello`



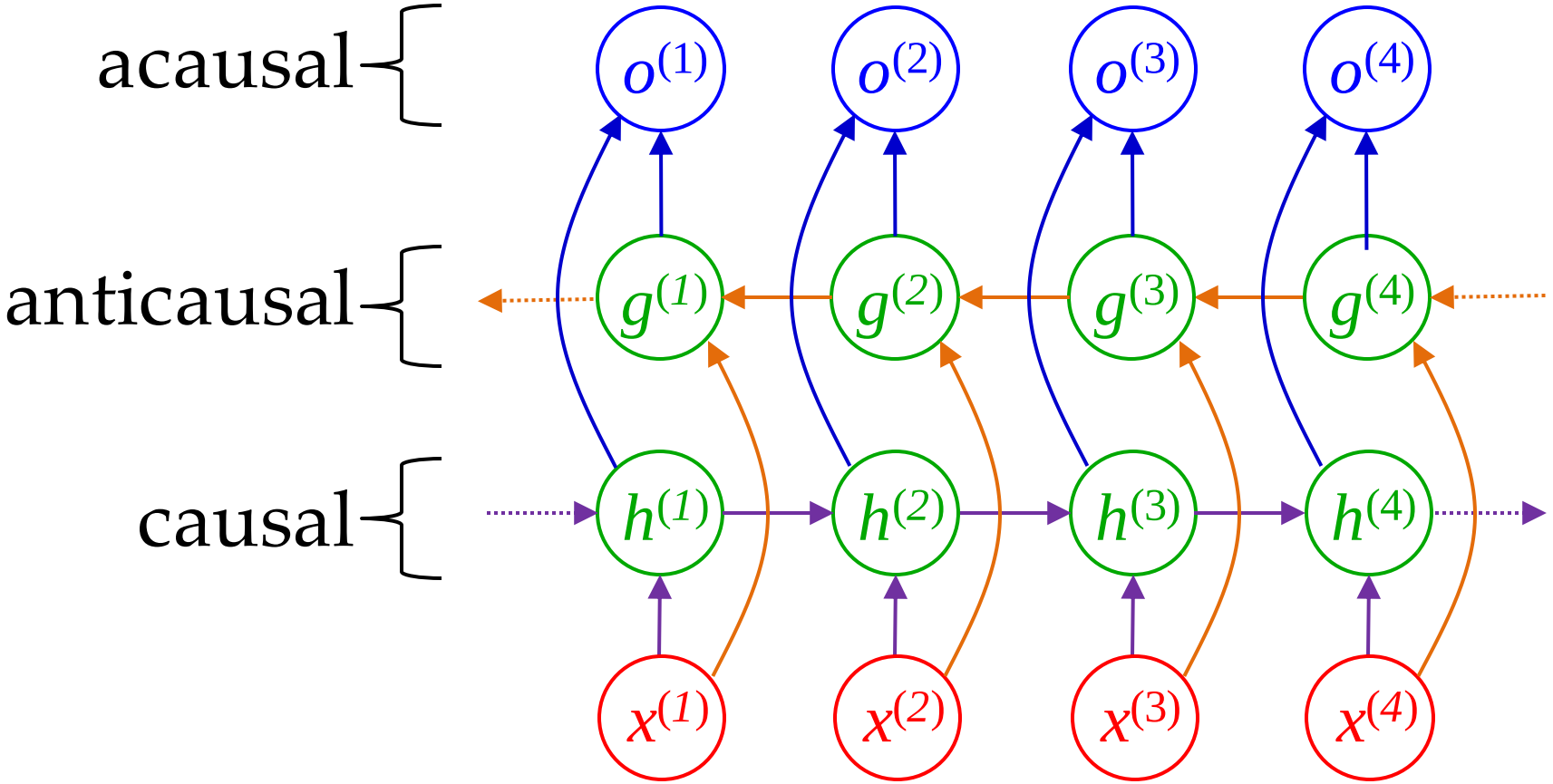
Longueur de sortie $o^{(\tau)}$

- Lors de la génération, on doit s'avoir quand arrêter d'échantillonner le RNN
- **3 stratégies** :
  1. Symbole spécial (**<END>**)
  2. Sortie supplémentaire 0-1 (via sigmoïde), qui prédit la fin
  3. Sortie qui prédit τ directement (régression)

RNN bi-directionnel

- Sortie $o^{(t)}$ peut dépendre de toute la séquence (1 à τ)
- Information pertinente parfois après une entrée x
  - ordre des mots dans une langue
    - adjectif avant ou après mot
    - langue SVO, SOV
  - reconnaissance de la voix
    - coarticulation
  - bio-informatique

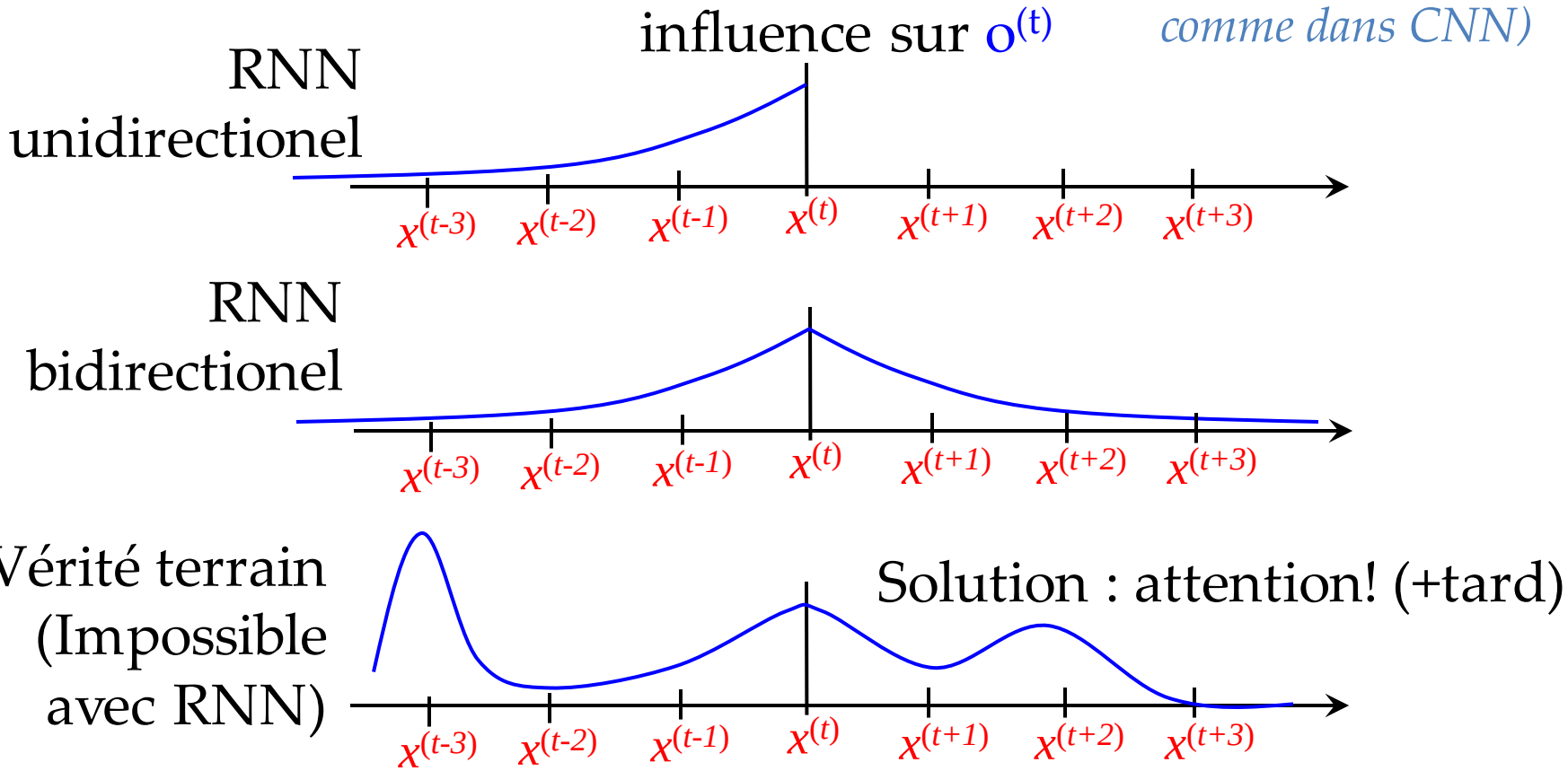
RNN bi-directionnel



Longue portée

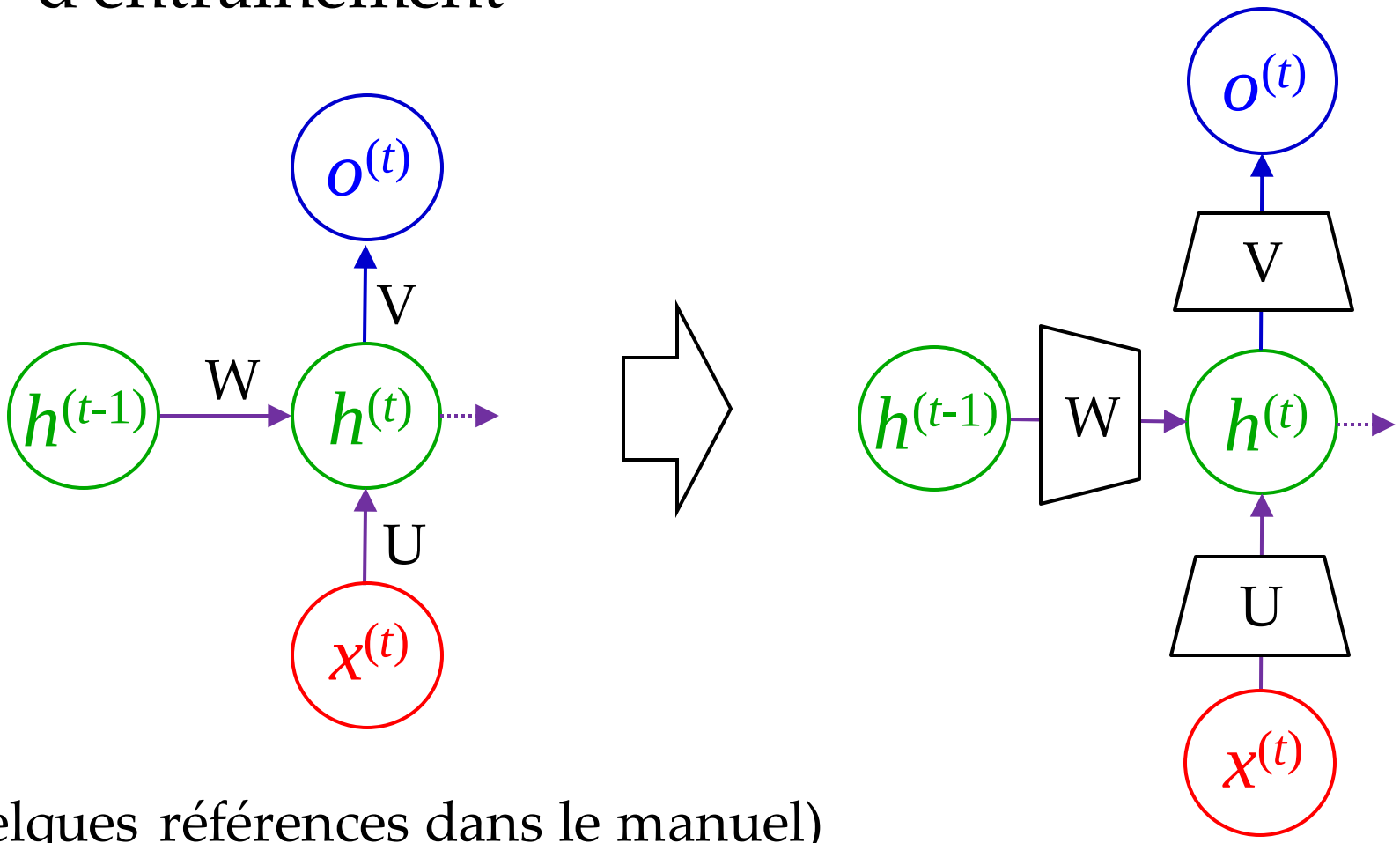
- Influence à longue portée difficile dans RNN
- CNN : champ récepteur croissant en profondeur
- RNN : décroissance exponentielle de l'influence

(pas une fenêtre précise, comme dans CNN)



Deep RNN?

- Ne semble pas être (encore) commun
- Profondeur accentue les difficultés d'entraînement



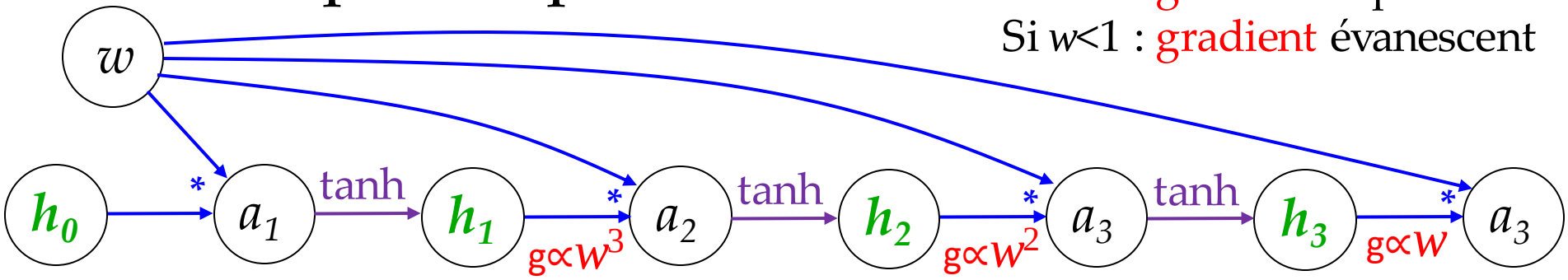
(quelques références dans le manuel)

Gradient et entrainement

Exploding/vanishing gradient

- Poids W partagés
- Exemple simplifié :

Si $w>1$: gradient explose
Si $w<1$: gradient évanescent



Pour un réseau récurrent linéaire (simplification) : $h^{(t)} = W^{\mathrm{T}} h^{(t-1)}$

Décomposition en éléments propres $W = Q \Lambda Q^{\mathrm{T}}$

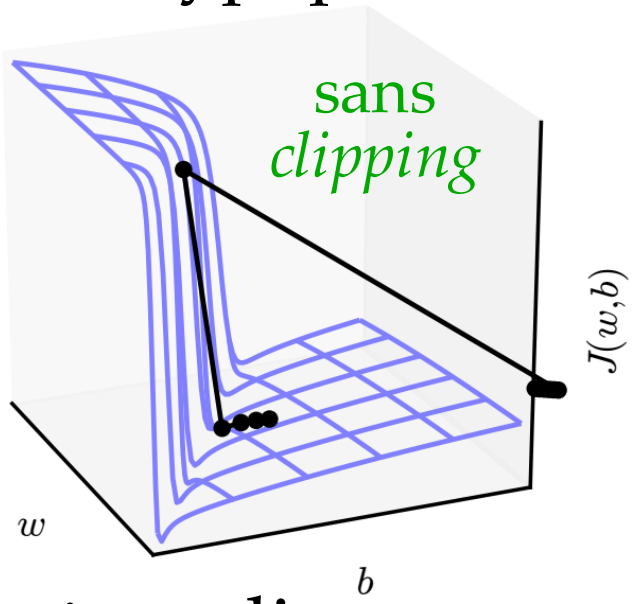
$h^{(t)} = Q^{\mathrm{T}} \Lambda^{t} Q h^{(0)}$ ⇨ si valeur propre $\lambda > 1$: vecteur propre explose

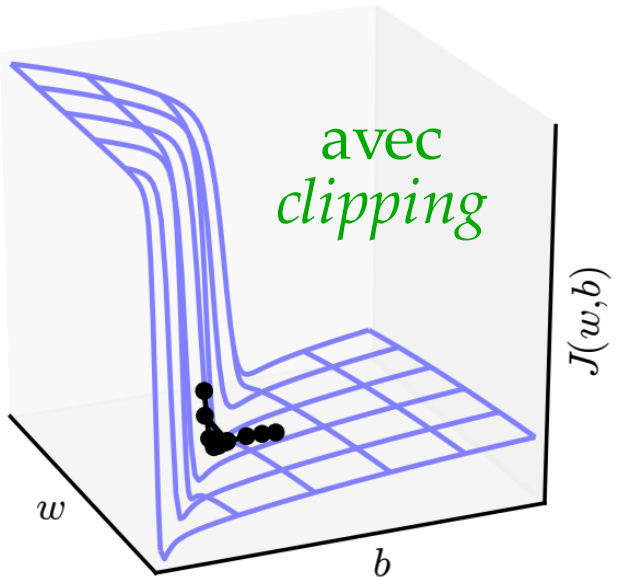
si valeur propre $\lambda < 1$: vecteur propre évanescent

Gradient clipping pour entraînement RNN

tiré de : I. Goodfellow et al. Deep Learning

- Ravins typique dans les RNN :





- Solution, clipper :

la norme du gradient

$$if \ \|\vec{g}\| > v$$

$$\vec{g} \frac{v}{\|\vec{g}\|}$$

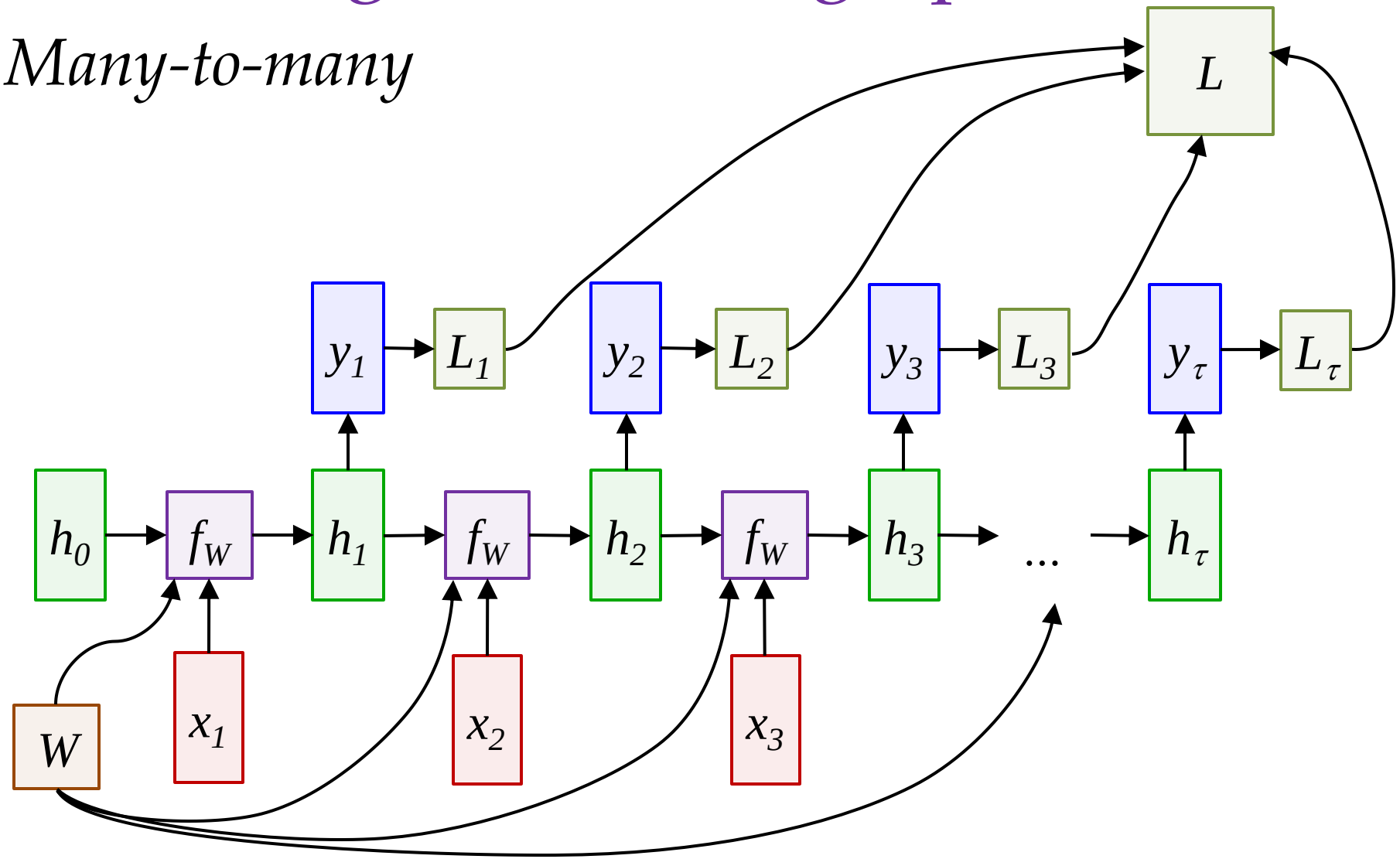
les entrées du gradient, individuellement

$$\vec{g} = \begin{bmatrix} 88493.4 \\ -0.3 \\ ... \\ -9948423 \end{bmatrix} = \begin{bmatrix} v \\ -0.3 \\ ... \\ -v \end{bmatrix}$$

(similaire comme scaling par paramètre, lors de l'optimisation)

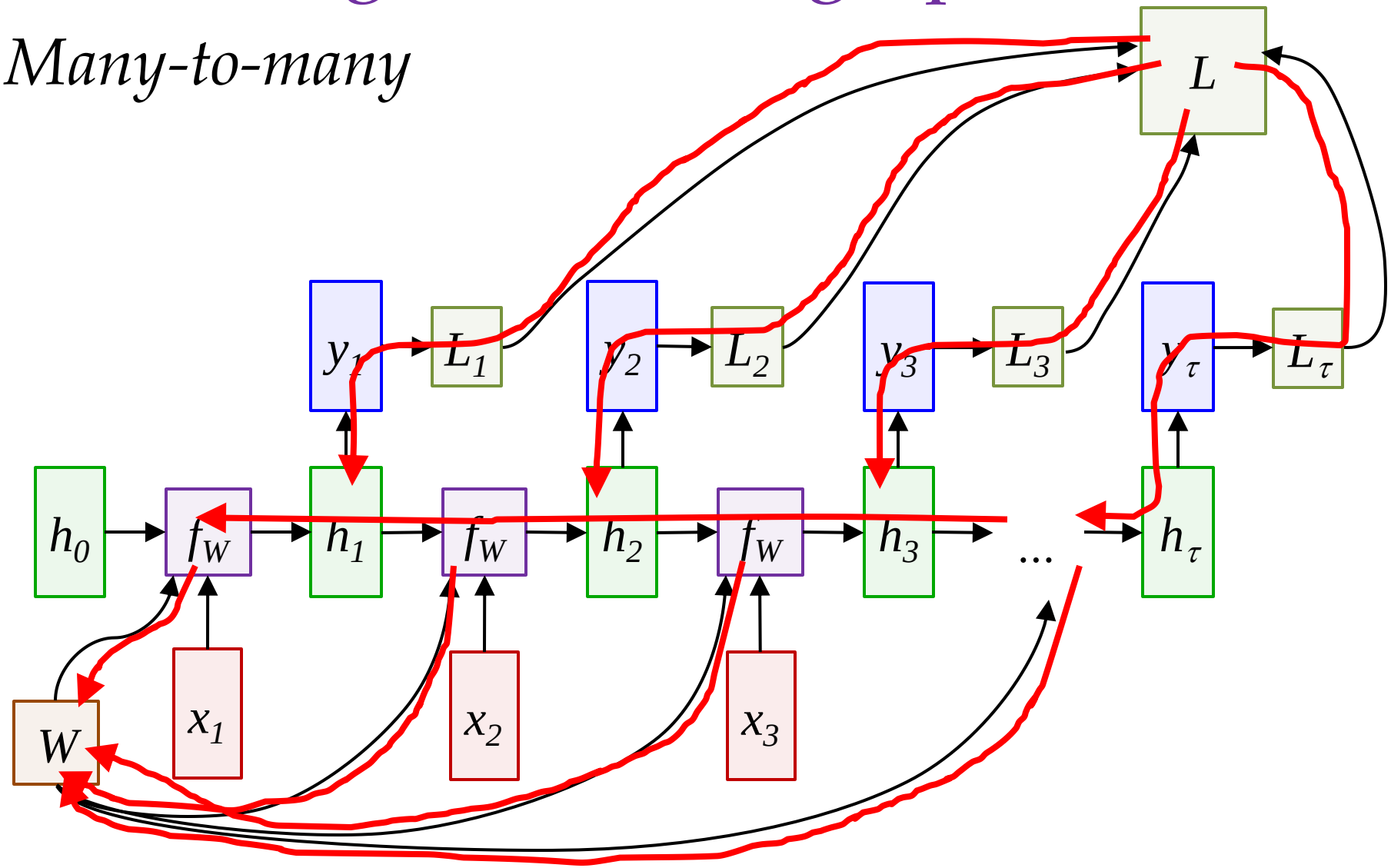
- Si NaN, bouger au hasard d'une magnitude v

Calcul du gradient sur graphe déroulé



adapté de cs231n

Calcul du gradient sur graphe déroulé

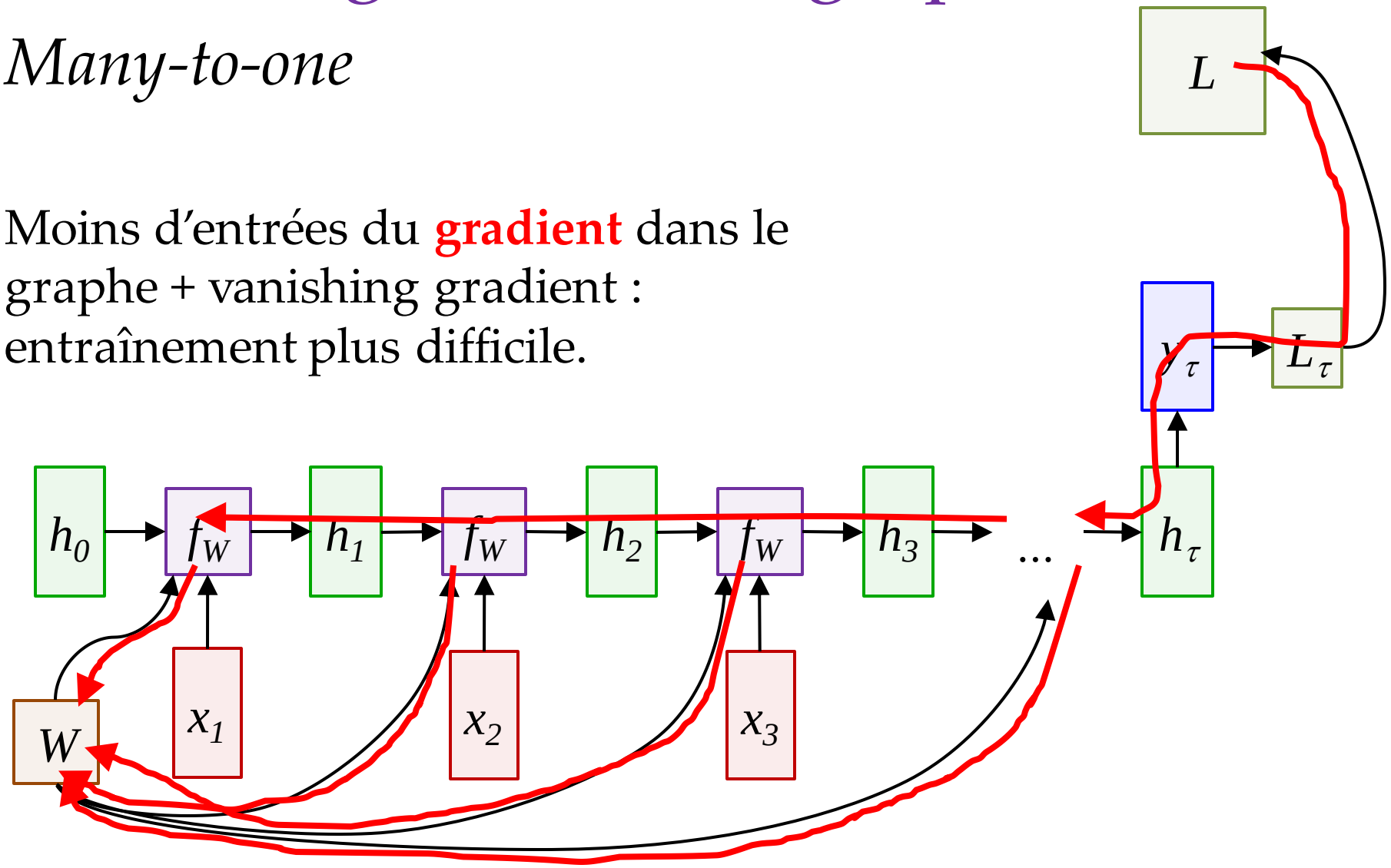


adapté de cs231n

Calcul du gradient sur graphe déroulé

Many-to-one

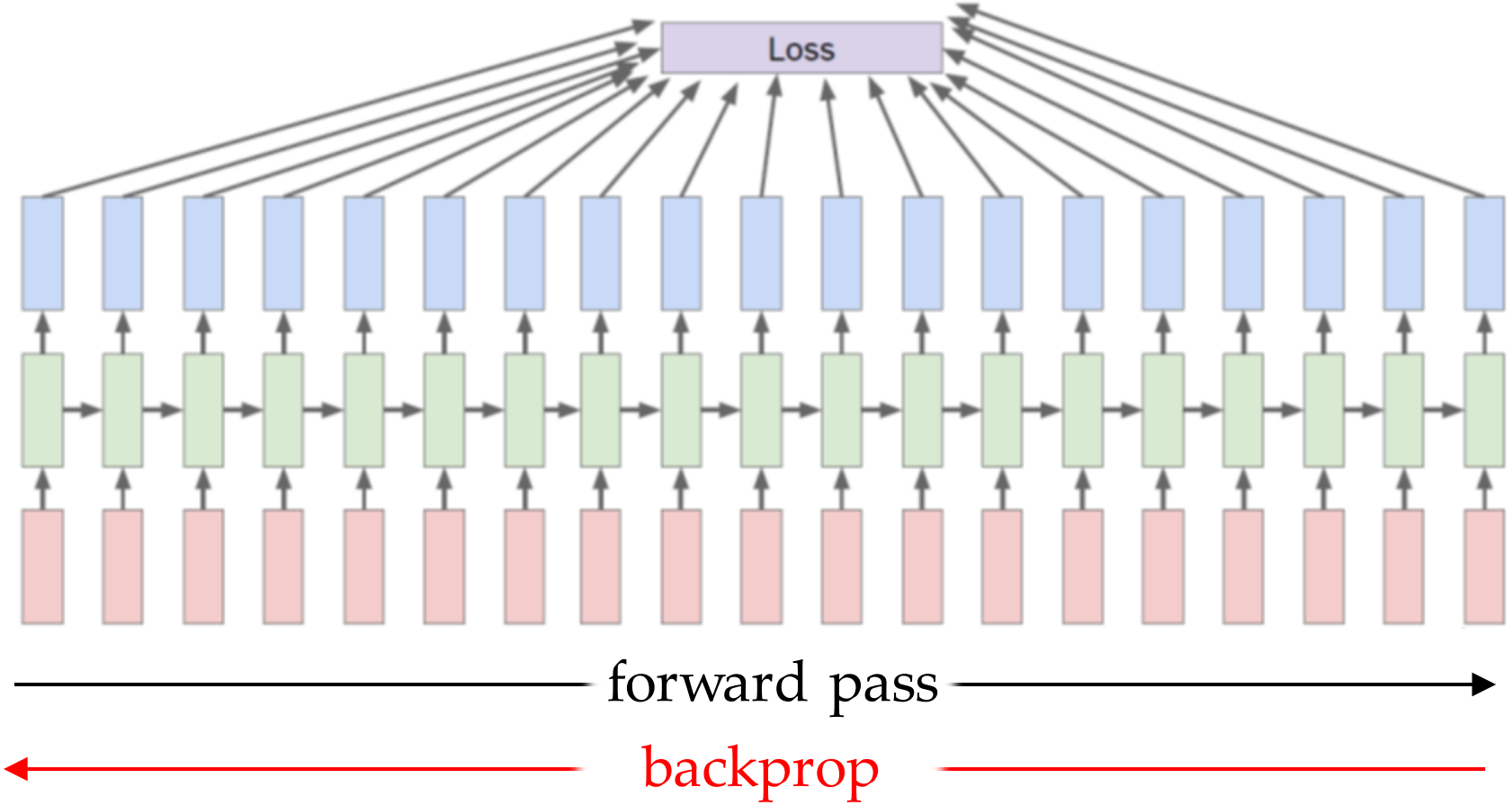
Moins d'entrées du **gradient** dans le graphe + vanishing gradient : entraînement plus difficile.



adapté de cs231n

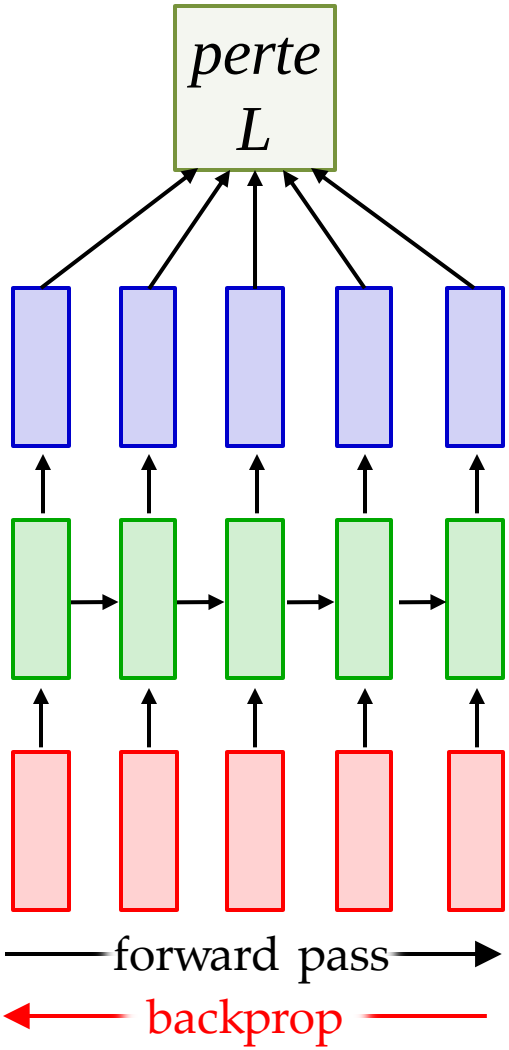
Backprop through time (BPTT)

- Calcule la séquence au complet



adapté de cs231n

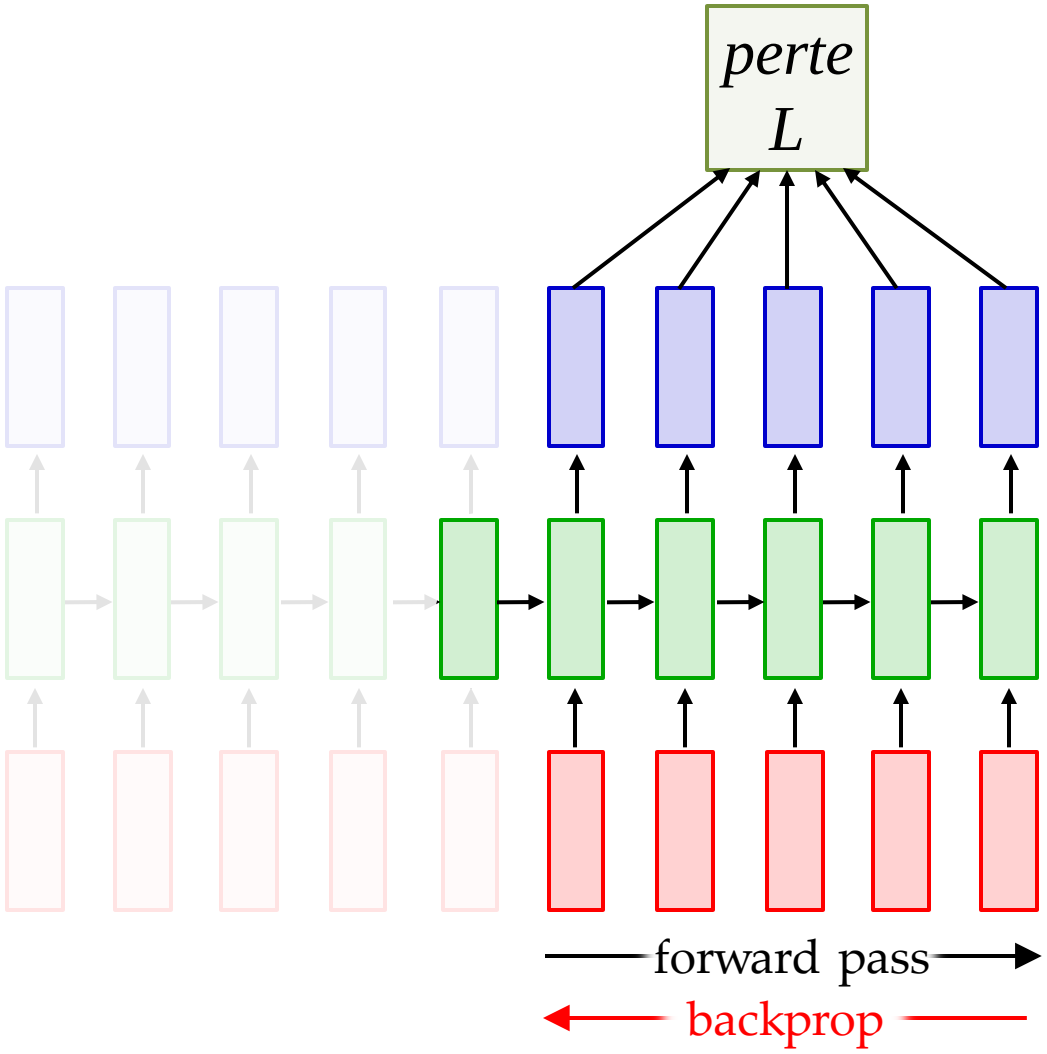
Truncated BPTT



Effectue BPTT sur des segments de la séquence

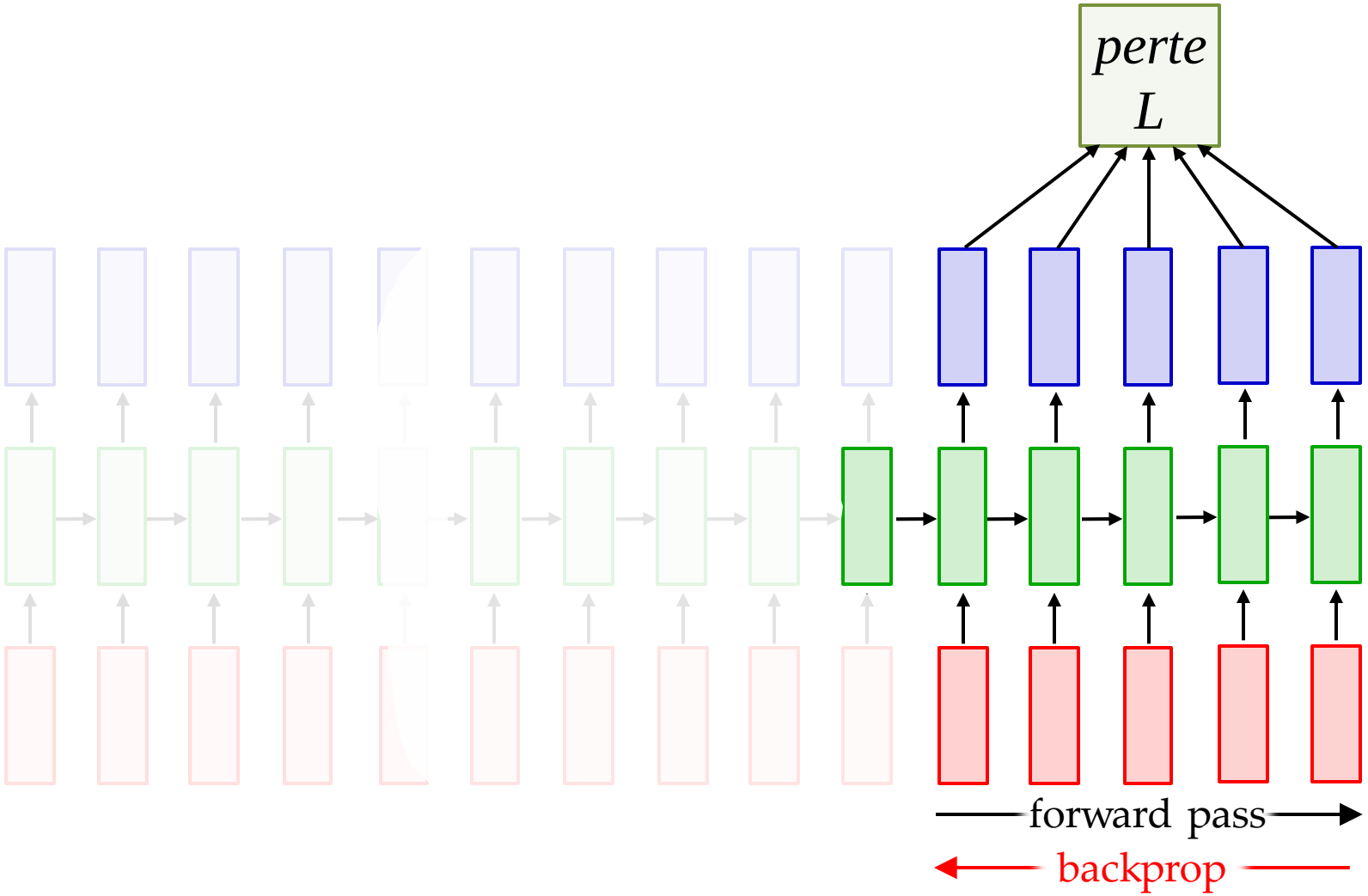
adapté de cs231n

Truncated BPTT



adapté de cs231n

Truncated BPTT



adapté de cs231n