Apprentissage profond et représentation latente de séquences peptidiques

Rémy Sun sous la direction de François Coste

Département d'informatique
ENS Rennes

Equipe Dyliss
IRISA

XTRA 2016

Quelles applications pour les protéines ?

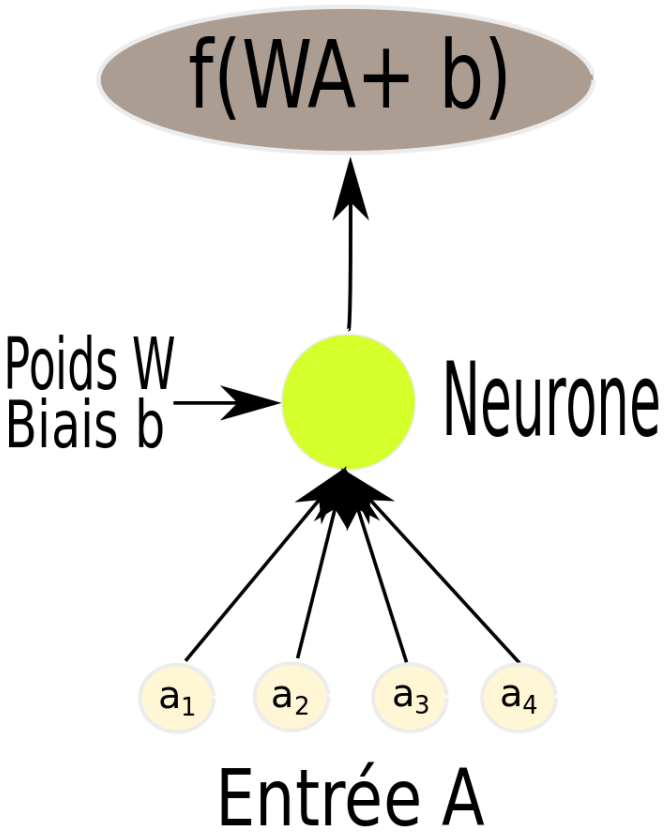
1. Apprentissage profond ?
   - Pourquoi l'apprentissage « profond » ?
   - Entraînement non-supervisé : Autoencodeurs et représentations latentes
   - Architectures standards
2. Application aux protéines
   - Qu'est-ce-qu'une protéine ?
   - Etat de l'art
   - Traiter des séquences peptidiques
   - Architectures entrainées & résultats

Quelles applications pour les protéines ?

1. Apprentissage profond ?
   - Pourquoi l'apprentissage « profond » ?
   - Entraînement non-supervisé : Autoencodeurs et représentations latentes
   - Architectures standards
2. Application aux protéines
   - Qu'est-ce-qu'une protéine ?
   - Etat de l'art
   - Traiter des séquences peptidiques
   - Architectures entrainées & résultats

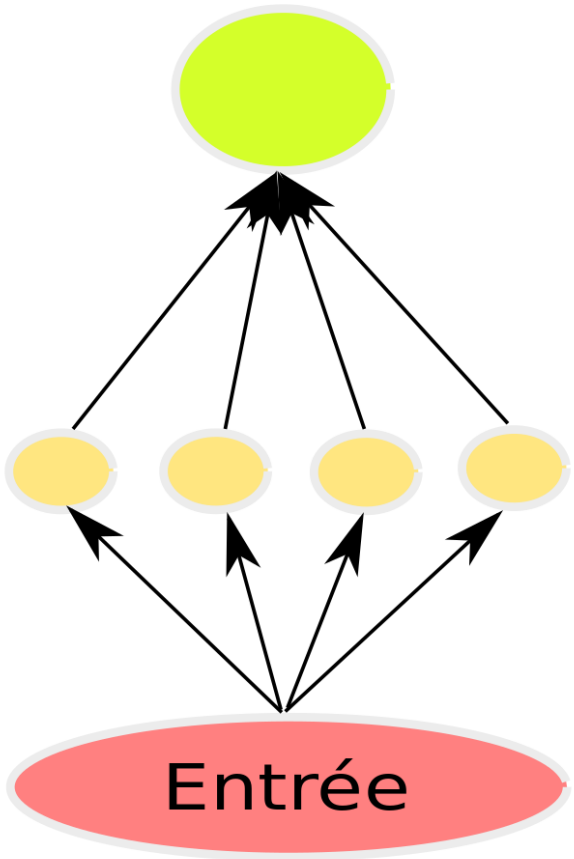
Plan

Une unité de calcul à paramètres optimisables



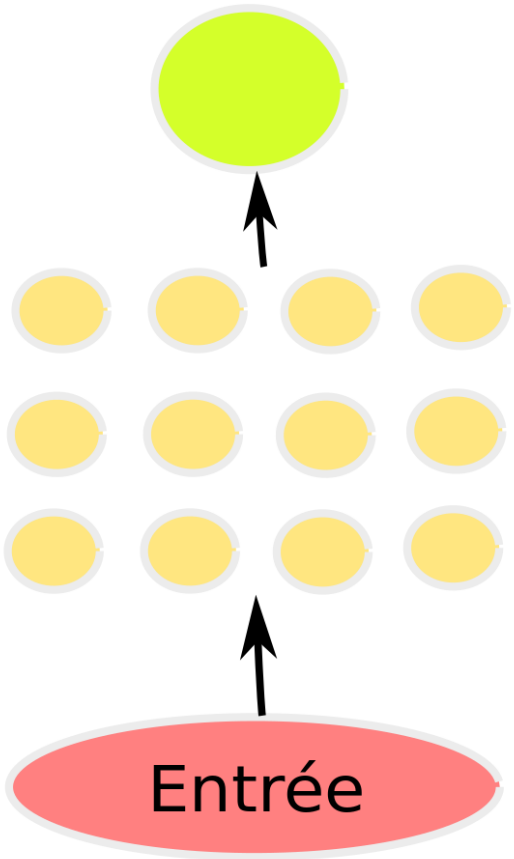
- Entrée A, poids W, biais b
- Transformation linéaire $WA + b$
- Activation non-linéaire f
- Apprentissage de W et b
  - Par rétropropagation sur la distance à l'objectif

Représentation hiérarchiques par couches



- Plusieurs couches de neurones
- Hiérarchie : plusieurs niveaux de représentations
- Evanouissement de gradient
- Grands ensembles d'entraînement

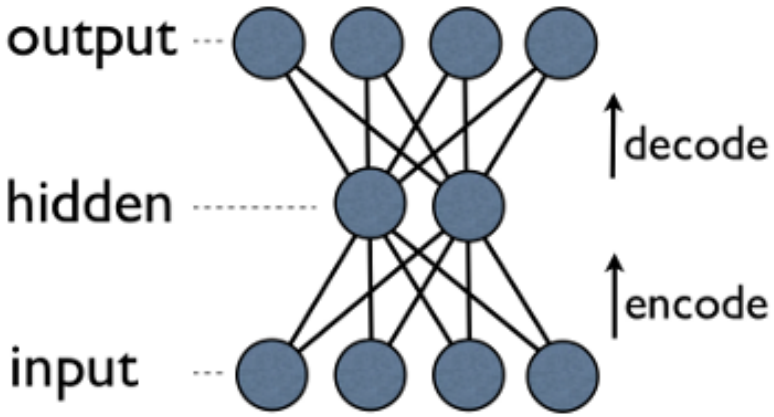
Représentation hiérarchiques par couches



- Plusieurs couches de neurones
- Hiérarchie : plusieurs niveaux de représentations
- Evanouissement de gradient
- Grands ensembles d'entraînement

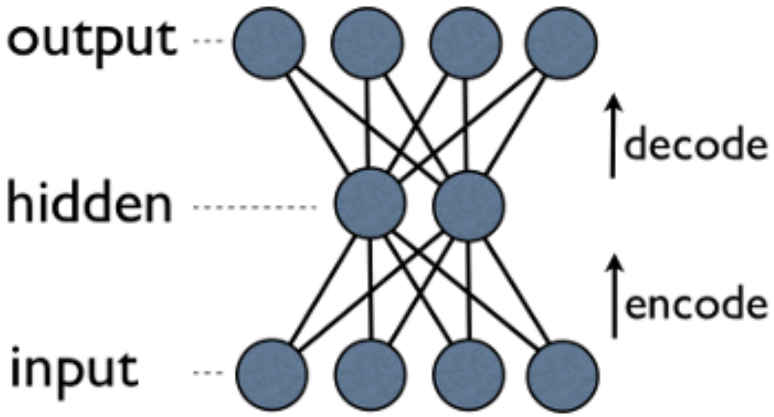
Plan

Acquérir des représentations latentes intéressantes



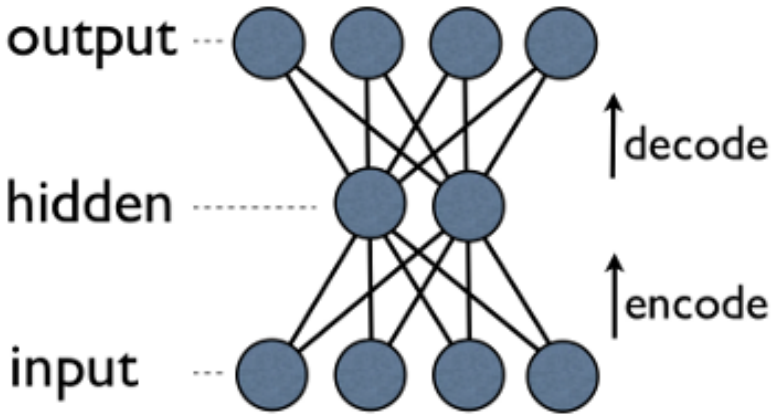
- Non supervisé
- Encodage/Décodage
- **Représentation latente**
- Eviter d'encoder l'identité
  - Compression
  - Bruitage
  - Régularisation

Acquérir des représentations latentes intéressantes



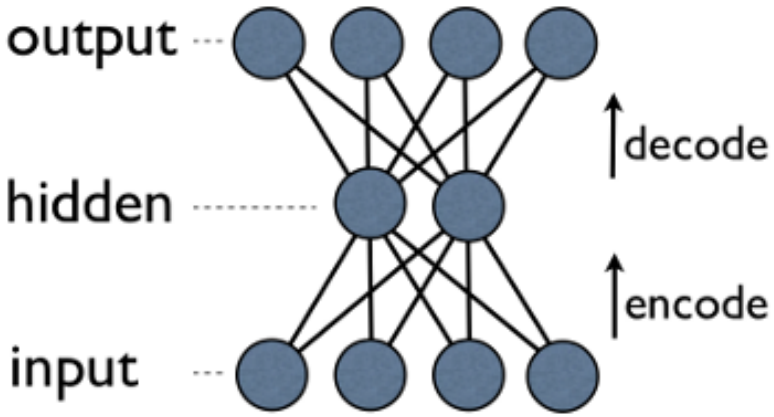
- Non supervisé
- Encodage/Décodage
- **Représentation latente**
- Eviter d'encoder l'identité
  - Compression
  - Bruitage
  - Régularisation

Acquérir des représentations latentes intéressantes



- Non supervisé
- Encodage/Décodage
- **Représentation latente**
- Eviter d'encoder l'identité
  - Compression
  - Bruitage
  - Régularisation

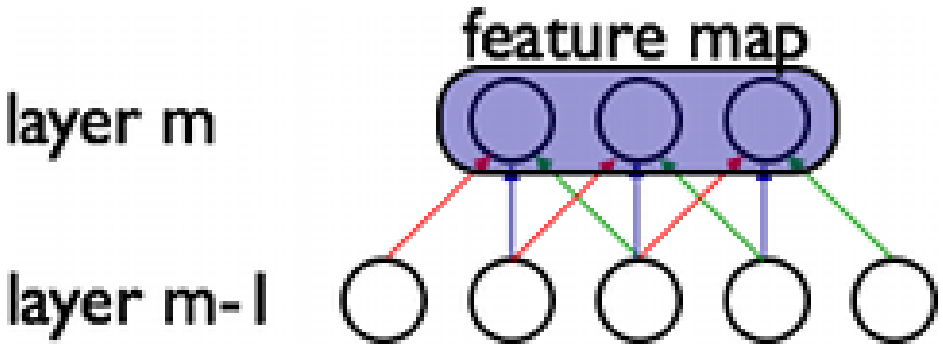
Acquérir des représentations latentes intéressantes



- Non supervisé
- Encodage/Décodage
- **Représentation latente**
- Eviter d'encoder l'identité
  - Compression
  - Bruitage
  - Régularisation

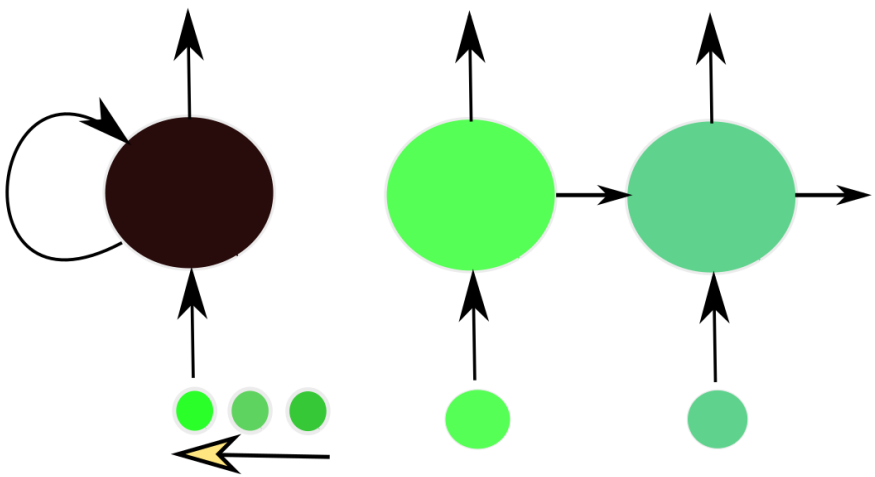
Plan

Réseaux Convolutionnels : recherche de caractéristique



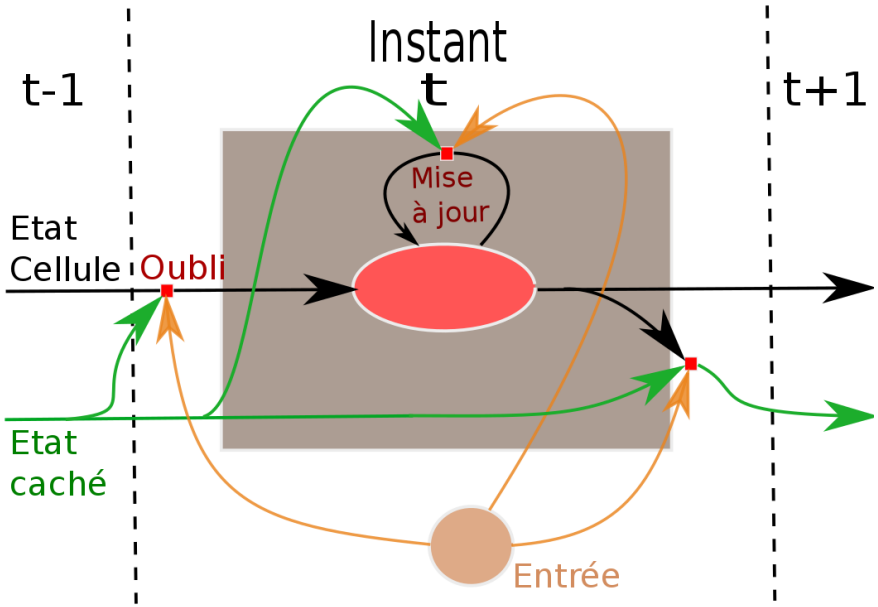
- Filtres de caractéristiques
- Permet d'isoler des caractéristiques locales

Réseaux récurrents : tenir compte de l'ordre d'apparition



- Dépendance temporelles
- Sortie + état caché persistant (boucle de rétroaction)
- Réseau « profond »à une couche
- Pas de dépendances hiérarchiques
- Unité LSTM (Long Short-Term Memory)

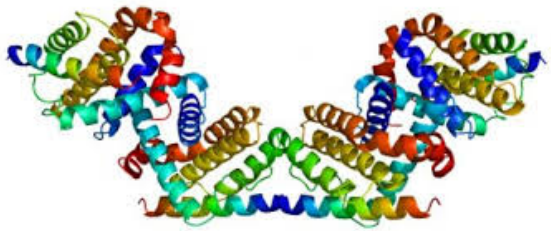
Réseaux récurrents : tenir compte de l'ordre d'apparition



- Dépendance temporelles
- Sortie + état caché persistant (boucle de rétroaction)
- Réseau « profond »à une couche
- Pas de dépendances hiérarchiques
- Unité LSTM (Long Short-Term Memory)

Plan

Une molécule chimique



- Acide aminés : molécules chimiques
- Structure primaire : chaîne d'acides aminés
- Structure secondaire : structures locales formé par les acides (hélices-α, brins-β, ...)
- Structure tertiaire : forme tridimensionnelle

Plan

Peu de travaux concernant les protéines

- Succès en :
  - Reconnaissance d'image, langages naturels, prédiction de sentiments, données bio-médicales, représentation, ...
- Protéines :
  - Prédiction de structures secondaires et locales
    - Heffernan R. et al. 2015 Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning.
    - Spencer M et al. 2015 A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction
    - ...
  - Classification de protéines selon différents critéres
    - Jian-Wei L. et al. 2013 Predicting protein structural classes with autoencoder neural networks

Peu de travaux concernant les protéines

- Succès en :
  - Reconnaissance d'image, langages naturels, prédiction de sentiments, données bio-médicales, représentation, ...
- Protéines :
  - Prédiction de structures secondaires et locales
    - Heffernan R. et al. 2015 Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning.
    - Spencer M et al. 2015 A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction
    - ...
  - Classification de protéines selon différents critéres
    - Jian-Wei L. et al. 2013 Predicting protein structural classes with autoencoder neural networks

Peu de travaux concernant les protéines

- Succès en :
  - Reconnaissance d'image, langages naturels, prédiction de sentiments, données bio-médicales, représentation, ...
- Protéines :
  - Prédiction de structures secondaires et locales
    - Heffernan R. et al. 2015 Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning.
    - Spencer M et al. 2015 A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction
    - ...
  - Classification de protéines selon différents critéres
    - Jian-Wei L. et al. 2013 Predicting protein structural classes with autoencoder neural networks

Peu de travaux concernant les protéines

- Succès en :
  - Reconnaissance d'image, langages naturels, prédiction de sentiments, données bio-médicales, représentation, ...
- Protéines :
  - Prédiction de structures secondaires et locales
    - Heffernan R. et al. 2015 Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning.
    - Spencer M et al. 2015 A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction
    - ...
  - Classification de protéines selon différents critéres
    - Jian-Wei L. et al. 2013 Predicting protein structural classes with autoencoder neural networks

Peu de travaux concernant les protéines

- Succès en :
  - Reconnaissance d'image, langages naturels, prédiction de sentiments, données bio-médicales, représentation, ...
- Protéines :
  - Prédiction de structures secondaires et locales
    - Heffernan R. et al. 2015 Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning.
    - Spencer M et al. 2015 A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction
    - ...
  - Classification de protéines selon différents critéres
    - Jian-Wei L. et al. 2013 Predicting protein structural classes with autoencoder neural networks

Peu de travaux concernant les protéines

- Succès en :
  - Reconnaissance d'image, langages naturels, prédiction de sentiments, données bio-médicales, représentation, ...
- Protéines :
  - Prédiction de structures secondaires et locales
    - Heffernan R. et al. 2015 Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning.
    - Spencer M et al. 2015 A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction
    - ...
  - Classification de protéines selon différents critéres
    - Jian-Wei L. et al. 2013 Predicting protein structural classes with autoencoder neural networks

Peu de travaux concernant les protéines

- Succès en :
  - Reconnaissance d'image, langages naturels, prédiction de sentiments, données bio-médicales, représentation, ...
- Protéines :
  - Prédiction de structures secondaires et locales
    - Heffernan R. et al. 2015 Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning.
    - Spencer M et al. 2015 A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction
    - ...
  - Classification de protéines selon différents critéres
    - Jian-Wei L. et al. 2013 Predicting protein structural classes with autoencoder neural networks

Peu de travaux concernant les protéines

- Succès en :
  - Reconnaissance d'image, langages naturels, prédiction de sentiments, données bio-médicales, représentation, ...
- Protéines :
  - Prédiction de structures secondaires et locales
    - Heffernan R. et al. 2015 Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning.
    - Spencer M et al. 2015 A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction
    - ...
  - Classification de protéines selon différents critéres
    - Jian-Wei L. et al. 2013 Predicting protein structural classes with autoencoder neural networks

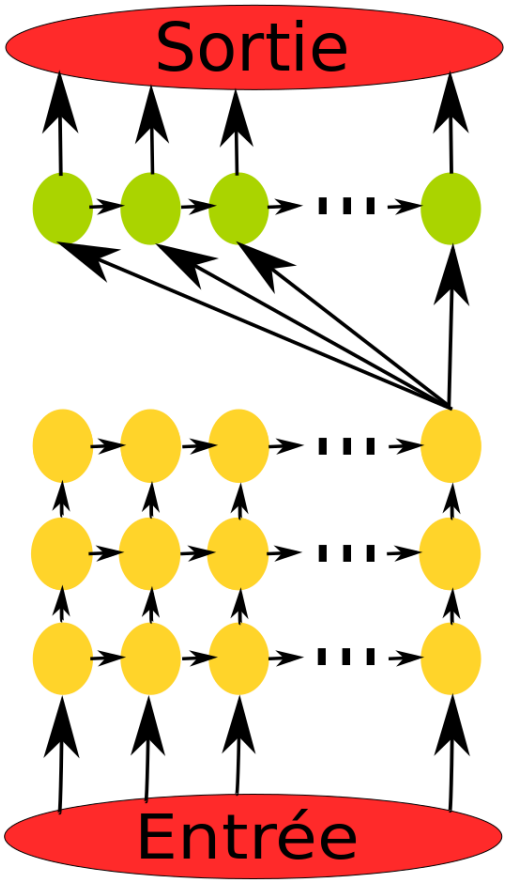
Plan

Traiter des fragments courts pour étudier des chaînes longues

- Tâche sur une chaîne longue : prédiction de classe structurale (SCOPe 2.6, 40%)
  - Travaux usuels : représentation par vecteur de fréquence des protéines augmenté
  - Découpage de la chaîne en fragments courts
- Etude sur les séquences peptidiques
- Représentation de l'acide a_i par $V = (v_k)$ où $v_i = 1$ et $v_k = 0 (k \neq i)$

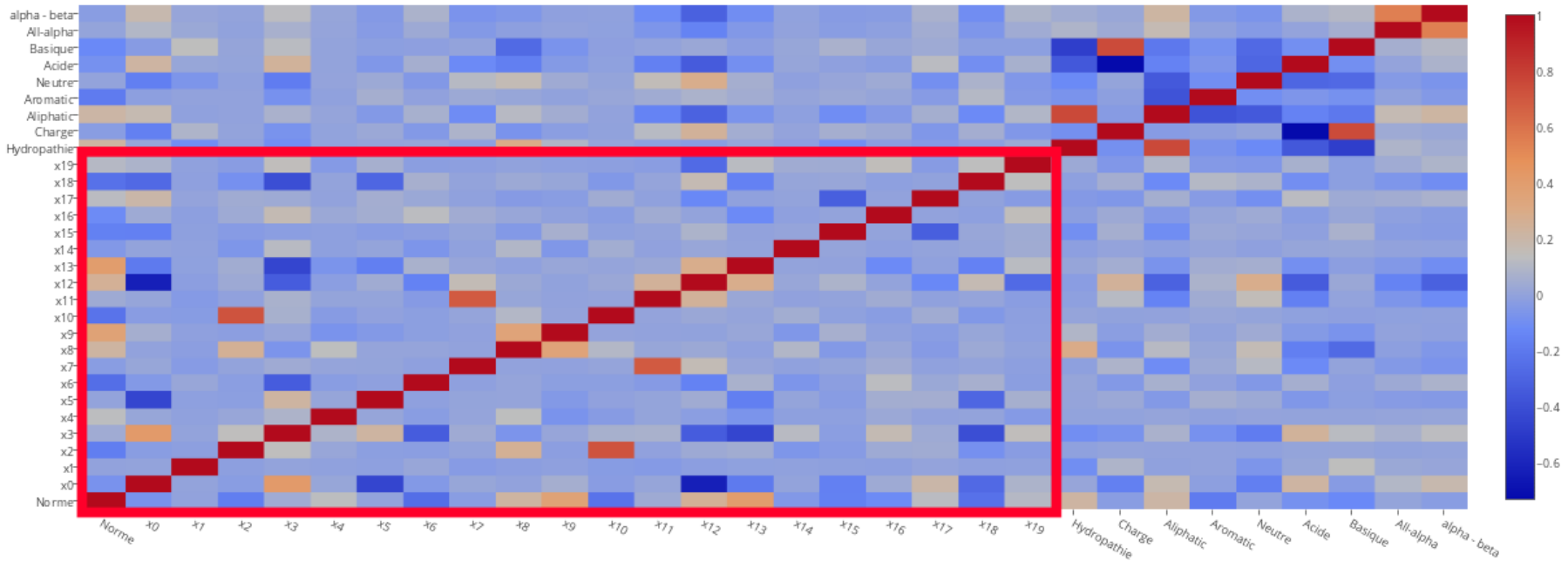
Plan

Autoencodeurs



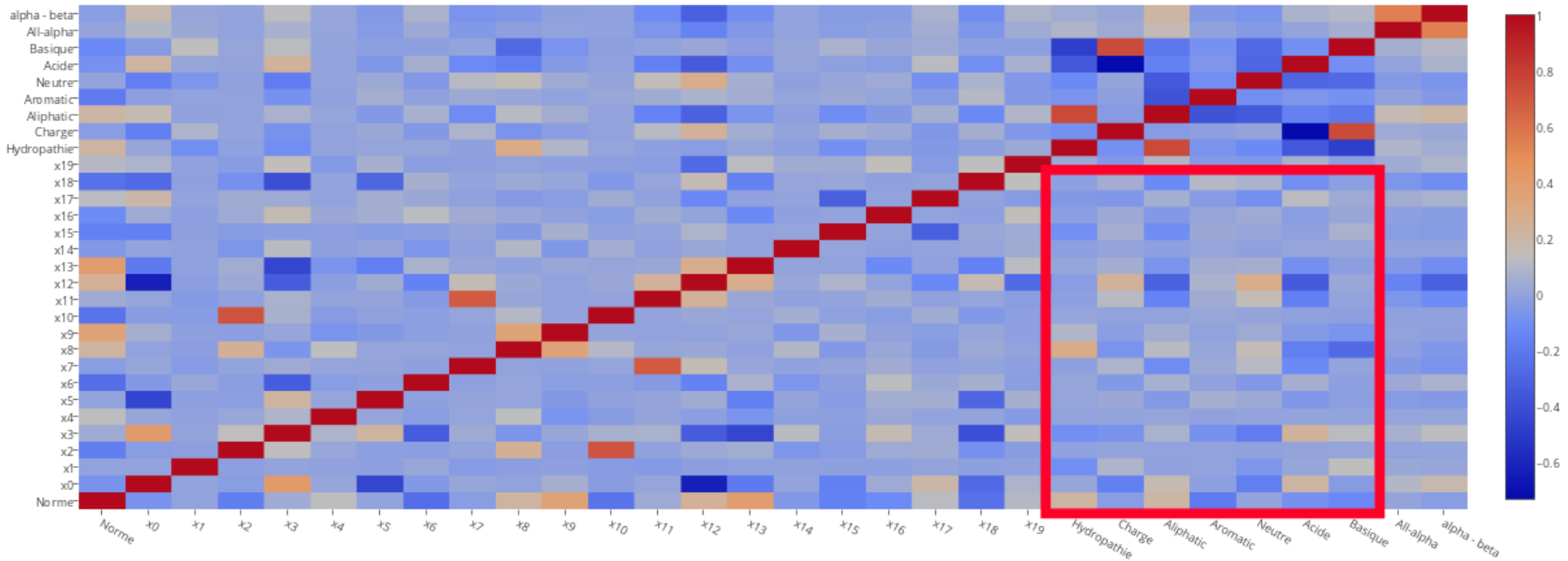
- Entraînement sur des fragments de taille 11
- Augmentation de la taille de l'ensemble d'entraînement de 13500 à 700 000
- Espace latent à 20 dimensions
- Encodeur récurrent à 3 couches
- Décodeur récurrent

Les représentation latentes présentent des corrélations remarquables



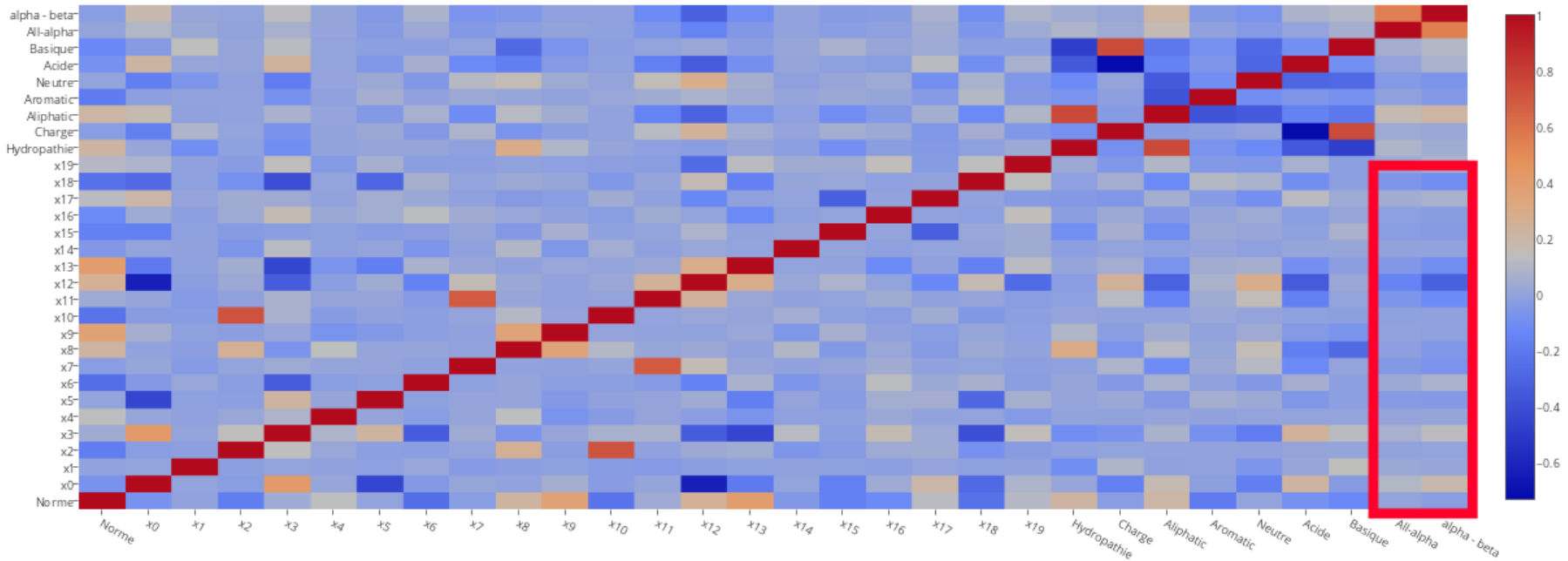
- Dimensions liées dans l'espace latent
- Corrélation de coordonnées à l'hydropathie, à la charge ...
- Pas de corrélation à la structure secondaire observées

Les représentation latentes présentent des corrélations remarquables



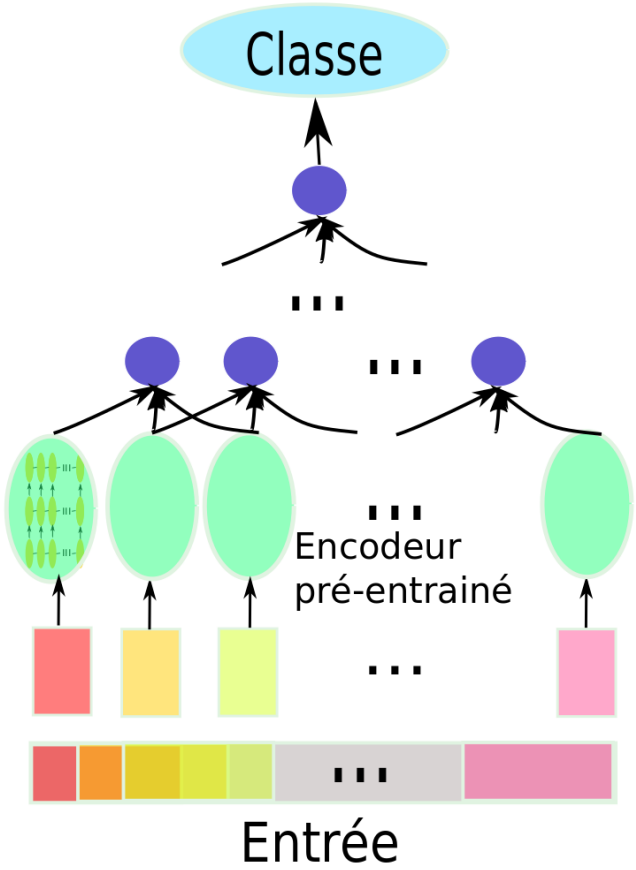
- Dimensions liées dans l'espace latent
- Corrélation de coordonnées à l'hydropathie, à la charge ...
- Pas de corrélation à la structure secondaire observées

Les représentation latentes présentent des corrélations remarquables



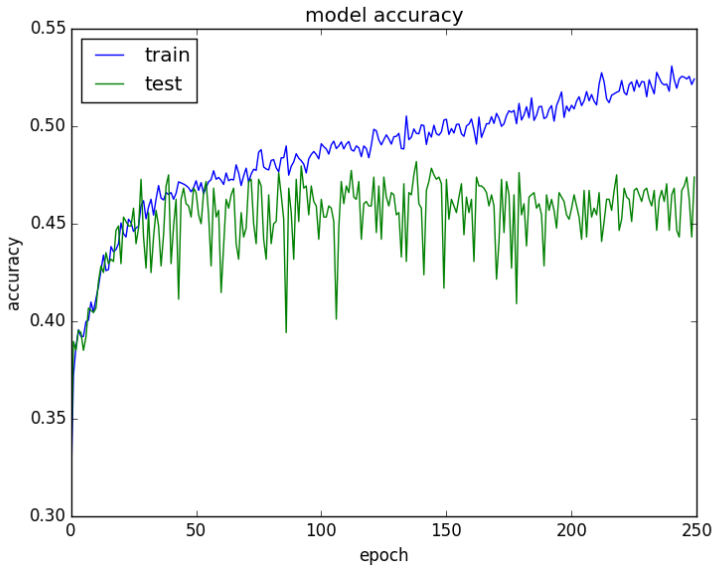
- Dimensions liées dans l'espace latent
- Corrélation de coordonnées à l'hydropathie, à la charge ...
- Pas de corrélation à la structure secondaire observées

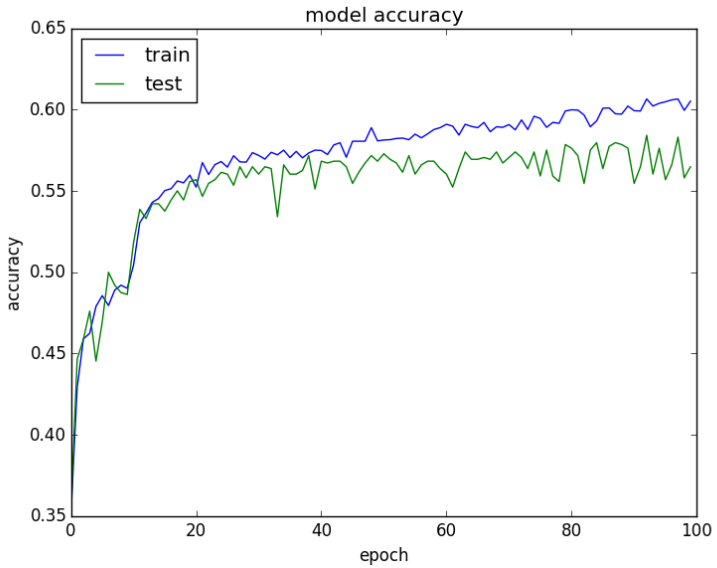
Classificateur de classe structurales



- Tâche : classifier les classes structurales des protéines
- Classificateur convolutionnel
- Premières couches pré-entraînées
- Validation de la représentation latente acquise

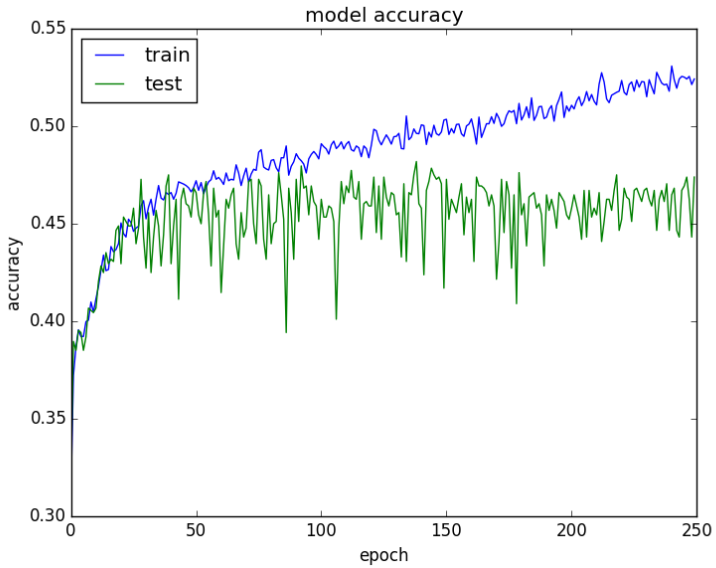
Les représentations latentes sont exploitables par un classificateur structural

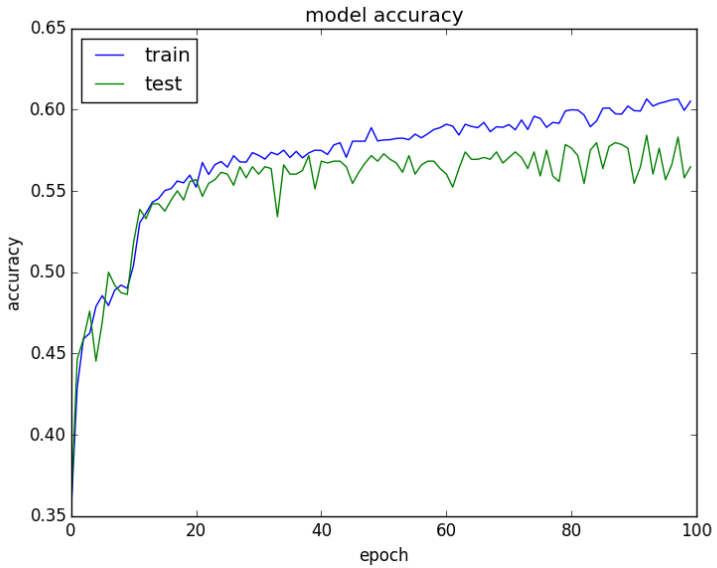




- Comparaison favorable au même classificateur non pré-entrainé :
  - Atteinte plus rapide de la précision maximale
  - Précision maximale plus élevée
- Pertinence de la représentation latente

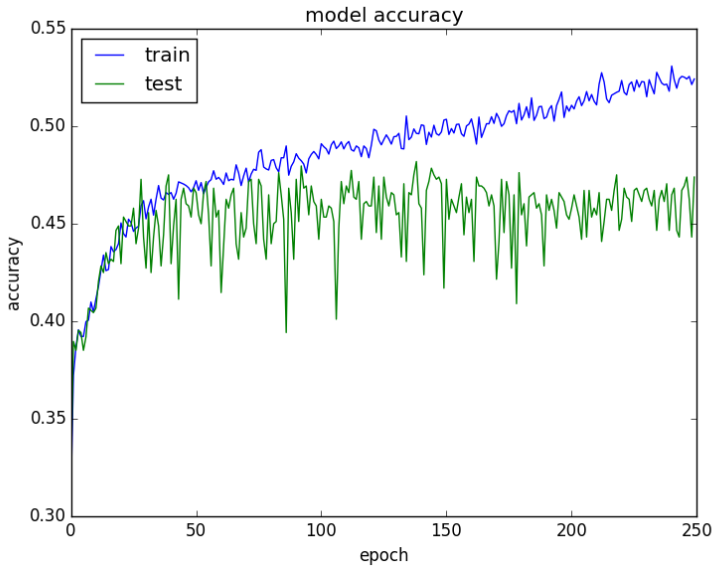
Les représentations latentes sont exploitables par un classificateur structural

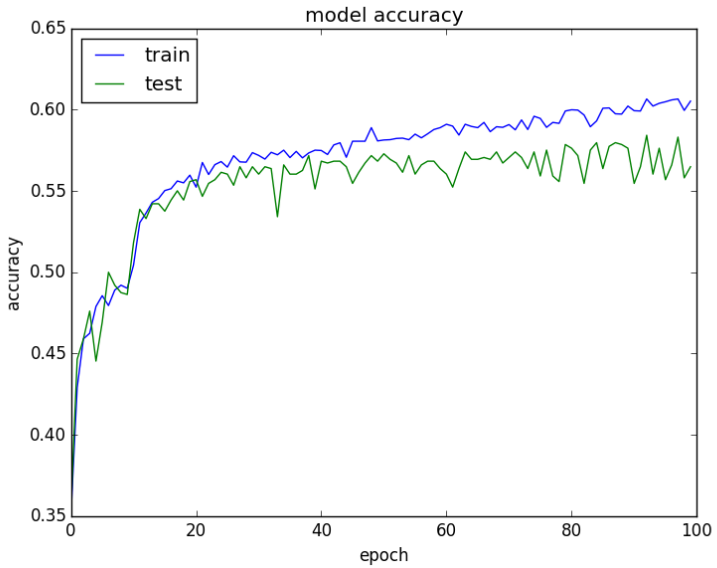




- Comparaison favorable au même classificateur non pré-entrainé :
  - Atteinte plus rapide de la précision maximale
  - Précision maximale plus élevée
- Pertinence de la représentation latente

Les représentations latentes sont exploitables par un classificateur structural





- Comparaison favorable au même classificateur non pré-entrainé :
  - Atteinte plus rapide de la précision maximale
  - Précision maximale plus élevée
- Pertinence de la représentation latente

Résumé

- L'apprentissage profond permet de détecter des structure hiérarchiques ou temporelles.
- Problème particulier : Pas assez d'exemples labélisés et chaînes très longues.
- Apparition de corrélations entre la représentation latente et des caractéristiques des séquences peptidiques.

- Perspectives
  - Utilisation d'autres architectures utilisées en langages naturels.
  - Influence des hyper paramètres.

Résumé

- L'apprentissage profond permet de détecter des structure hiérarchiques ou temporelles.
- Problème particulier : Pas assez d'exemples labélisés et chaînes très longues.
- Apparition de corrélations entre la représentation latente et des caractéristiques des séquences peptidiques.
- Perspectives
  - Utilisation d'autres architectures utilisées en langages naturels.
  - Influence des hyper paramètres.

Résumé

- L'apprentissage profond permet de détecter des structure hiérarchiques ou temporelles.
- Problème particulier : Pas assez d'exemples labélisés et chaînes très longues.
- Apparition de corrélations entre la représentation latente et des caractéristiques des séquences peptidiques.

- Perspectives
  - Utilisation d'autres architectures utilisées en langages naturels.
  - Influence des hyper paramètres.

Résumé

- L'apprentissage profond permet de détecter des structure hiérarchiques ou temporelles.
- Problème particulier : Pas assez d'exemples labélisés et chaînes très longues.
- Apparition de corrélations entre la représentation latente et des caractéristiques des séquences peptidiques.

- Perspectives
  - Utilisation d'autres architectures utilisées en langages naturels.
  - Influence des hyper paramètres.

Résumé

- L'apprentissage profond permet de détecter des structure hiérarchiques ou temporelles.
- Problème particulier : Pas assez d'exemples labélisés et chaînes très longues.
- Apparition de corrélations entre la représentation latente et des caractéristiques des séquences peptidiques.

- Perspectives
  - Utilisation d'autres architectures utilisées en langages naturels.
  - Influence des hyper paramètres.

Résumé

- L'apprentissage profond permet de détecter des structure hiérarchiques ou temporelles.
- Problème particulier : Pas assez d'exemples labélisés et chaînes très longues.
- Apparition de corrélations entre la représentation latente et des caractéristiques des séquences peptidiques.

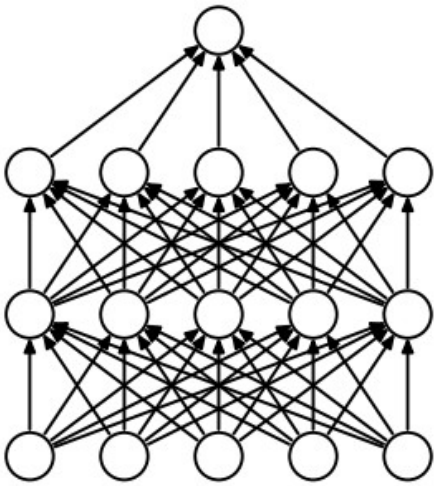
- Perspectives
  - Utilisation d'autres architectures utilisées en langages naturels.
  - Influence des hyper paramètres.

Résumé

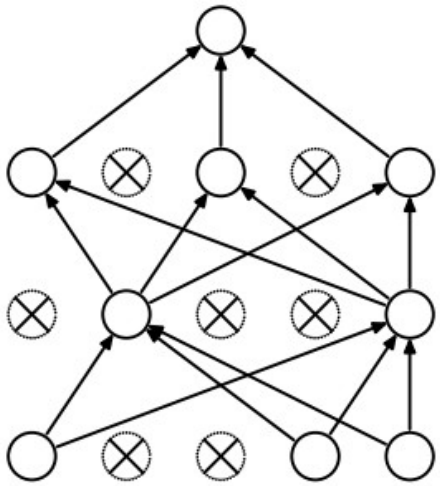
- L'apprentissage profond permet de détecter des structure hiérarchiques ou temporelles.
- Problème particulier : Pas assez d'exemples labélisés et chaînes très longues.
- Apparition de corrélations entre la représentation latente et des caractéristiques des séquences peptidiques.

- Perspectives
  - Utilisation d'autres architectures utilisées en langages naturels.
  - Influence des hyper paramètres.

Eviter le sur-entraînement



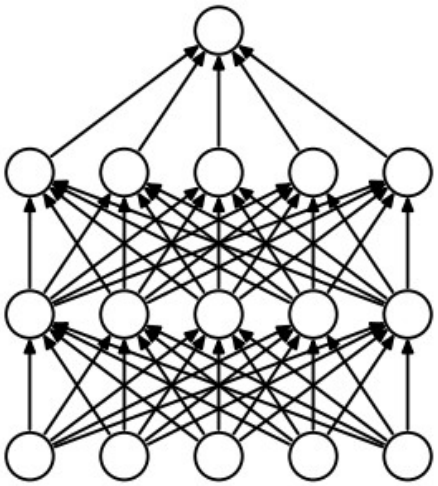
(a) Standard Neural Net



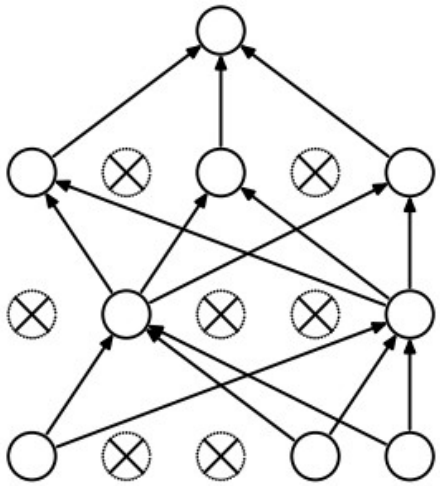
(b) After applying dropout.

- Désactiver aléatoirement des neurones
- Eliminer la concentration d'information
- Faire travailler tout le réseau
- Généraliser la représentation apprise
- Permet d'entraîner ad nauseam

Eviter le sur-entraînement



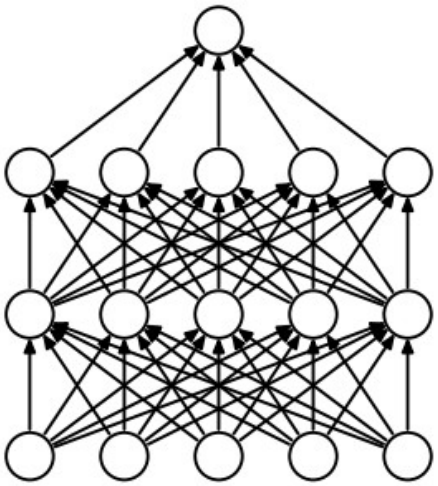
(a) Standard Neural Net



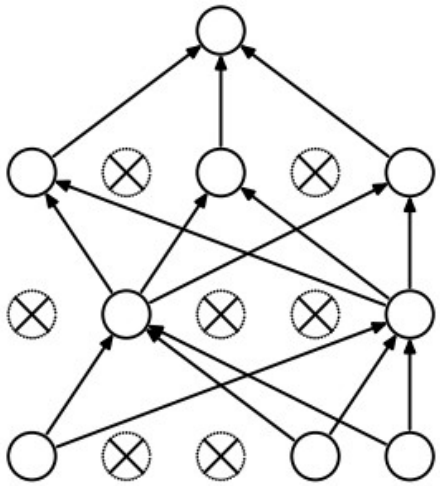
(b) After applying dropout.

- Désactiver aléatoirement des neurones
- Eliminer la concentration d'information
- Faire travailler tout le réseau
- Généraliser la représentation apprise
- Permet d'entraîner ad nauseam

Eviter le sur-entraînement



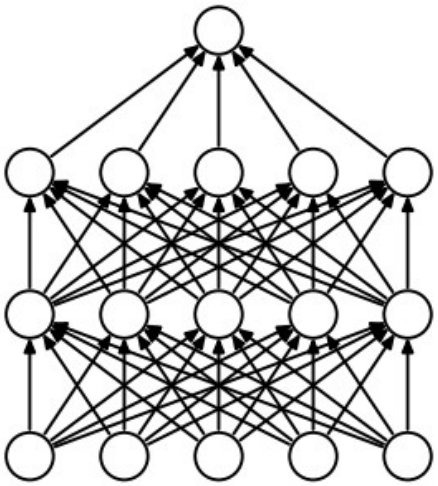
(a) Standard Neural Net



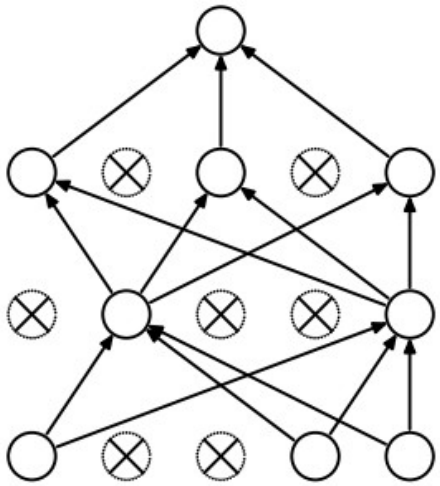
(b) After applying dropout.

- Désactiver aléatoirement des neurones
- Eliminer la concentration d'information
- Faire travailler tout le réseau
- Généraliser la représentation apprise
- Permet d'entraîner ad nauseam

Eviter le sur-entraînement



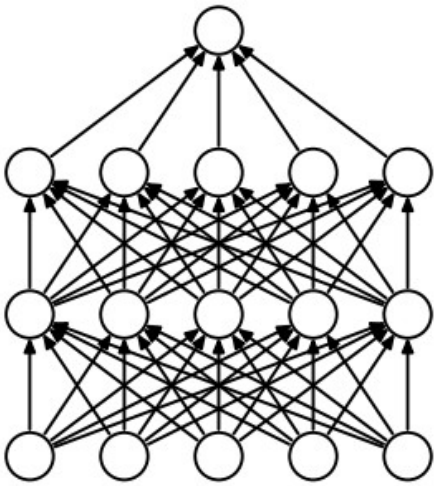
(a) Standard Neural Net



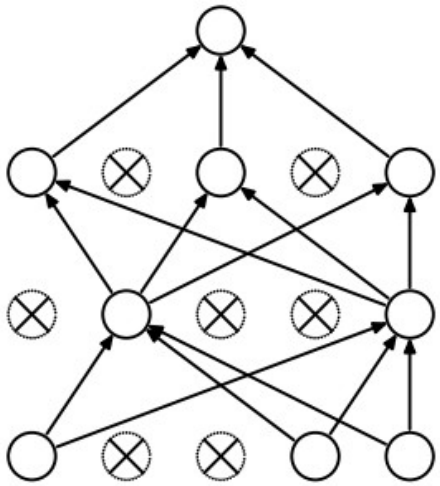
(b) After applying dropout.

- Désactiver aléatoirement des neurones
- Eliminer la concentration d'information
- Faire travailler tout le réseau
- Généraliser la représentation apprise
- Permet d'entraîner ad nauseam

Eviter le sur-entraînement



(a) Standard Neural Net



(b) After applying dropout.

- Désactiver aléatoirement des neurones
- Eliminer la concentration d'information
- Faire travailler tout le réseau
- Généraliser la représentation apprise
- Permet d'entraîner ad nauseam