Réseaux de neurones opérant sur des graphes pour le traitement automatique des documents

Théorie et applications

Adrien Guille (ERIC, Université Lyon 2) @ Atelier TextMine (EGC 2023)

Réseaux de neurones usuels

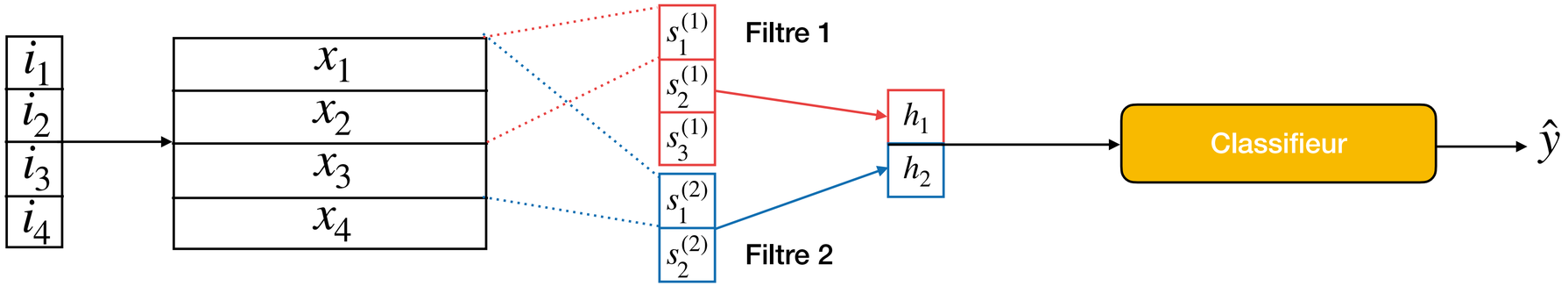
Encodage du texte

- **Séquence de représentations vectorielles des mots**
  - Soit un texte de longueur N, usuellement un réseau de neurones opère sur une séquence de la forme :
  - $$X = \begin{pmatrix} x_{1,1} \\ x_{1,2} \\ x_{1,3} \\ \vdots \\ x_{1,d} \end{pmatrix} \rightarrow \begin{pmatrix} x_{2,1} \\ x_{2,2} \\ x_{2,3} \\ \vdots \\ x_{2,d} \end{pmatrix} \rightarrow \ldots \rightarrow \begin{pmatrix} x_{N,1} \\ x_{N,2} \\ x_{N,3} \\ \vdots \\ x_{N,d} \end{pmatrix}$$

Réseaux de neurones usuels

Résolution de tâches supervisées

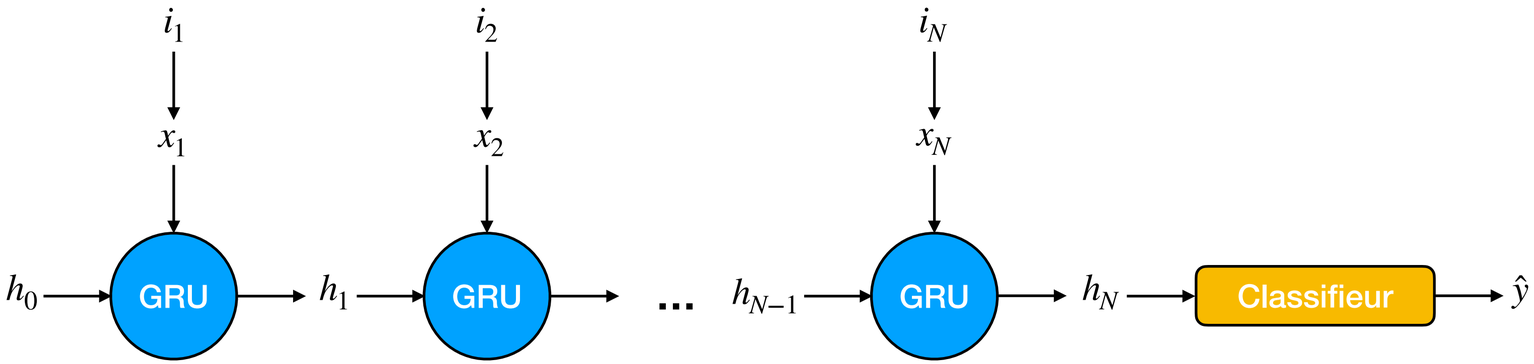
- **Réseaux convolutifs**
  - Détection de n-grammes de mots (*i.e.* carte d'attributs)
  - Sous-échantillonnage de la carte d'attributs
  - Classification



Réseaux de neurones usuels

Résolution de tâches supervisées

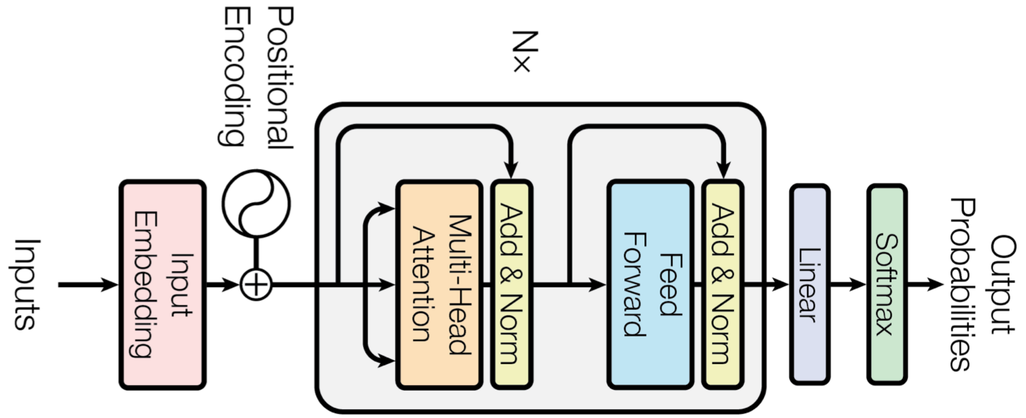
- **Réseaux récurrents**
  - Propagation d'un état caché à travers toute la séquence
    - Mise à jour de l'état caché en chaque position de la séquence
  - Classification en fonction du dernier état caché



Réseaux de neurones usuels

Résolution de tâches supervisées

- **Réseaux Transformer (encodeur)**
  - Les paramètres du réseau ($\mathcal{O}(100)$ millions) sont pré-entraînés
  - Classification en fonction de la représentation d'un token spécial en lien direct avec toute la séquence



(Figure modifiée d'après « Attention Is All You Need », Vaswani *et al.*, 2017)

Réseaux de neurones opérant sur les graphes

Données et tâches

Réseaux de neurones opérant sur les graphes

Tâches définies au niveau des sommets

- **Données**
  - Soit un graphe attribué $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X)$
    - $\mathcal{V}$: sommets
    - $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$: arêtes
    - $X \in \mathbb{R}^{|\mathcal{V}| \times d}$: attributs des sommets ; signal $\mathcal{V} \to \mathbb{R}^d$
- **Tâches**
  - Classification supervisée ou semi supervisée des sommets

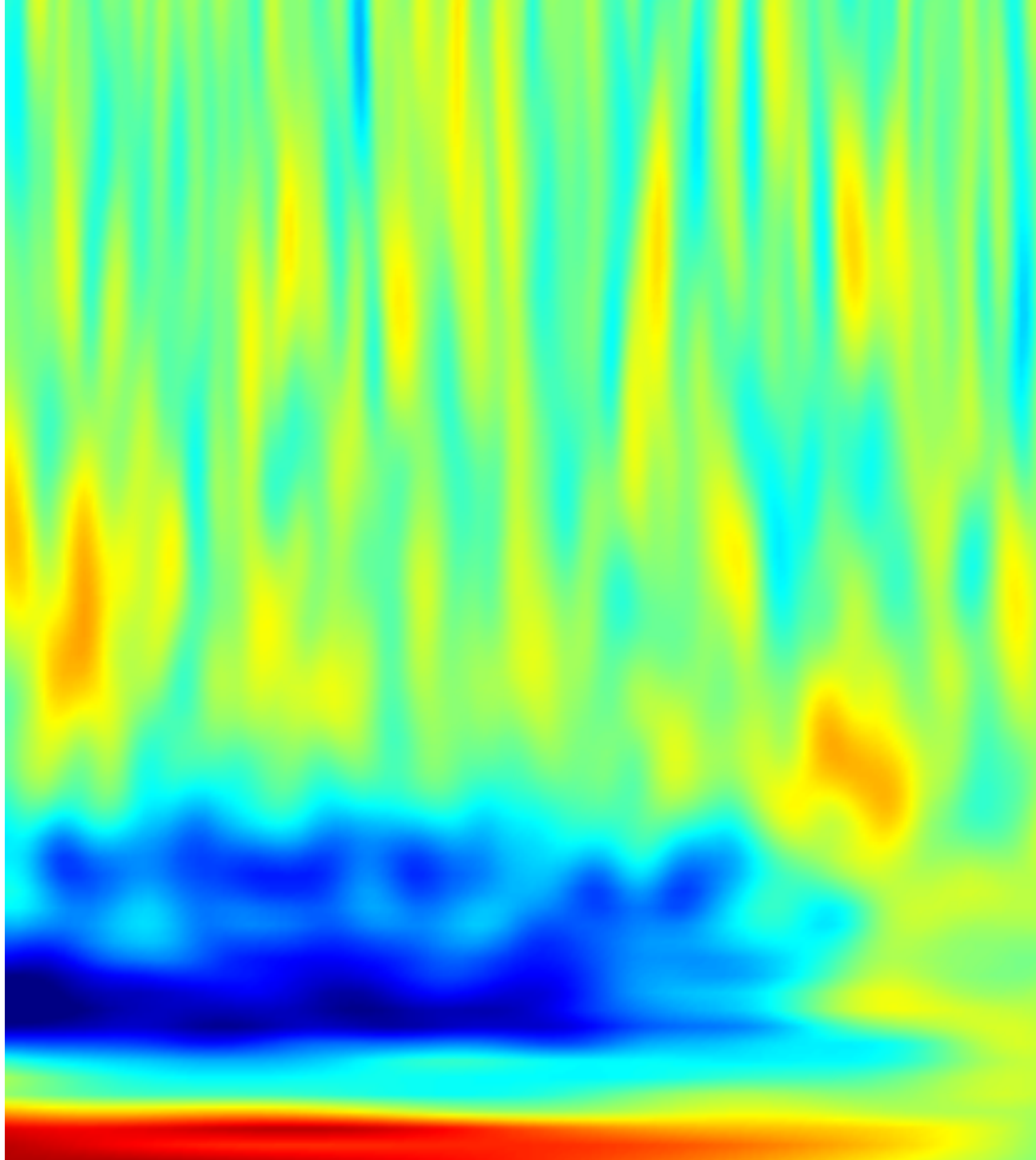
Réseaux de neurones opérant sur les graphes

Tâches définies au niveau des graphes

- **Données**
  - Soit une collection de graphes attribués $\{\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i, X_i)\}$
    - $\mathcal{V}_i$: sommets
    - $\mathcal{E}_i \subset \mathcal{V}_i \times \mathcal{V}_i$: arêtes
    - $X_i \in \mathbb{R}^{|\mathcal{V}_i| \times d}$: attributs des sommets ; signal $\mathcal{V}_i \to \mathbb{R}^d$
- **Tâches**
  - Classification supervisée de graphes

Réseaux de neurones opérant sur les graphes

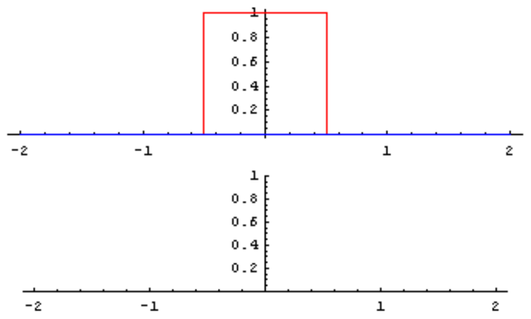
Formulation spectrale



Réseaux de neurones opérant sur les graphes

Formulation spectrale

- **Convolution d'un signal temporel par un noyau**
  - Soit un signal temporel $x \in \mathbb{R}^N$ et un noyau $g \in \mathbb{R}^N$
  - La convolution $x * g$ est une série $c \in \mathbb{R}^N$
    - Somme de x pondérée par g, avec décalage progressif de g autour de l'origine



(Lautaro Carmona, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=3066394)

Réseaux de neurones opérant sur les graphes

Formulation spectrale

- **Convolution d'un signal temporel par un noyau**
  - Dans le domaine temporel la convolution s'obtient via le produit matriciel entre la matrice circulante G obtenue d'après $g^\top$ et le signal x
  - $$c = \begin{pmatrix} g_1 & g_2 & g_3 & \cdots & g_N \\ g_N & g_1 & g_2 & \cdots & g_{N-1} \\ g_{N-1} & g_N & g_1 & \cdots & g_{N-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_2 & g_3 & g_4 & \cdots & g_1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{pmatrix}$$

Réseaux de neurones opérant sur les graphes

Formulation spectrale

- **Convolution d'un signal temporel par un noyau**
  - La matrice G est diagonalisable dans la base de Fourier
    - $G = Q \Lambda Q^\top$, avec $Q \in \mathbb{R}^{N \times N}$ les vecteurs propres de G, qui correspondent aux modes de Fourier du noyau
  - D'après le théorème de la convolution la convolution dans le domaine temporel équivaut à la multiplication dans le domaine de Fourier
    - $x * g = \mathcal{F}^{-1}(\mathcal{F}(x) \otimes \mathcal{F}(g)) = Q(Q^\top x \otimes Q^\top g) = Q \, \mathrm{diag}(g) \, Q^\top x$

Réseaux de neurones opérant sur les graphes

Formulation spectrale

- **« Convolution » d'un signal défini sur un graphe par un noyau**
  - On considère maintenant un signal $x \in \mathbb{R}^{|\mathscr{V}|}$, un signal $\mathscr{V} \rightarrow \mathbb{R}$
  - On formule une sorte de « convolution » par analogie, dans le domaine de Fourier, la notion de translation n'étant pas définie sur un graphe
  - La représentation de $\mathscr{G}$ dans le domaine de Fourier dépend de la matrice laplacienne associée
    - Matrice laplacienne normalisée symétrique : $L_{sym} = I - D^{-1/2}AD^{-1/2}$

Réseaux de neurones opérant sur les graphes

Formulation spectrale

- **« Convolution » d'un signal défini sur un graphe par un noyau**
  - Décomposition spectrale de la matrice laplacienne
  - $L_{sym} = Q \Lambda Q^\top$, avec Q les vecteurs propres du laplacien, *i.e.* les modes du graphe
  - Pour un noyau $g \in \mathbb{R}^{|\mathscr{V}|}$, on définie une opération imitant la formulation spectrale de la convolution $x \underset{\mathscr{G}}{*} g = Q(Q^\top x \otimes Q^\top g) = Q \, \mathrm{diag}(g) \, Q^\top x$
    - Problème : le nombre de paramètres est dépendant de la taille de $\mathscr{G}$

Réseaux de neurones opérant sur les graphes

Formulation spectrale

- **« Convolution » d'un signal défini sur un graphe par un noyau**
  - Approximation polynomiale du filtre
    - $g_\Theta = \sum_{k=0}^{K-1} \Theta_k \Lambda^k$: fixe le nombre de paramètres à K
  - Approximation par polynôme de Tchebychev
    - $\hat{g}_{\Theta'} = \sum_{k=0}^{K-1} \Theta'_k T_k(\tilde{\Lambda})$: avec $\tilde{\Lambda} = \frac{2}{\lambda_{\text{max}}} \Lambda - I$ et où $T_0(\tilde{\Lambda}) = \vec{1}$, $T_1(\tilde{\Lambda}) = \tilde{\Lambda}$ et les termes suivants définis par récurrence

Réseaux de neurones opérant sur les graphes

Formulation spectrale

- **Graph Convolutional Network : GCN (Kipf *et al.* 2017)**
  - On fixe $K = 2$ et on considère que $\lambda_{\text{max}} = 2$
  - On substitue A par $\tilde{A} = A + I$ et on corrige la matrice des degrés, $\tilde{D}$
    - Kipf nomme cela le « renormalization trick » car les valeurs propres sont translatées de $[0; 2]$ vers $[-1; 1]$: concrètement on ajoute des boucles
  - Couche GCN
    - $H^{(n)} = \text{ReLU}(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(n-1)}W^{(n)})$ avec $H^0 = X$, et où $W^{(n)}$ est une matrice de poids propre à la n-ième couche GCN

Réseaux de neurones opérant sur les graphes

Formulation spectrale

- **Graph Convolutional Network : GCN (Kipf *et al.* 2017)**
  - Définition d'un réseau GCN à deux couches pour la classification de sommets
    - $\hat{y} = \text{softmax}\Big(\tilde{L}\Big(\text{ReLU}(\tilde{L}XW^{(1)})\Big)W^{(2)}\Big)$, avec $\hat{y} \in [0;1]^C$
  - Estimation, semi supervisée, des paramètres
    - Soit $y : \mathscr{V}' \subset \mathscr{V} \rightarrow \{0;1\}^C$ les vraies classes d'un sous-ensemble de sommets
    - On minimise l'entropie croisée $-\sum_{v \in \mathscr{V}'} \sum_{c} y_{vc} \log \hat{y}_{vc}$

Réseaux de neurones opérant sur les graphes

Formulation par passage de messages

Réseaux de neurones opérant sur les graphes

Passage de messages

- **Un cadre général pour décrire les GNN**
  - Chaque couche réalise 2 étapes
    - 1) Chaque sommet envoie un message à ses voisins
    - 2) Chaque sommet agrège (*e.g.* somme pondérée) les messages reçus
- **Une vision beaucoup plus élémentaire du GCN**
  - $h_i^{(n)} = \sum_{v_j \in \mathcal{N}(i)} \frac{1}{\sqrt{d(i)}\sqrt{d(j)}} h_j^{(n-1)} W^{(n)}$, avec $\mathcal{N}(i) \subset \mathcal{V}$ les voisins du sommet i, ainsi que lui-même (boucle)

Réseaux de neurones opérant sur les graphes

Passage de messages

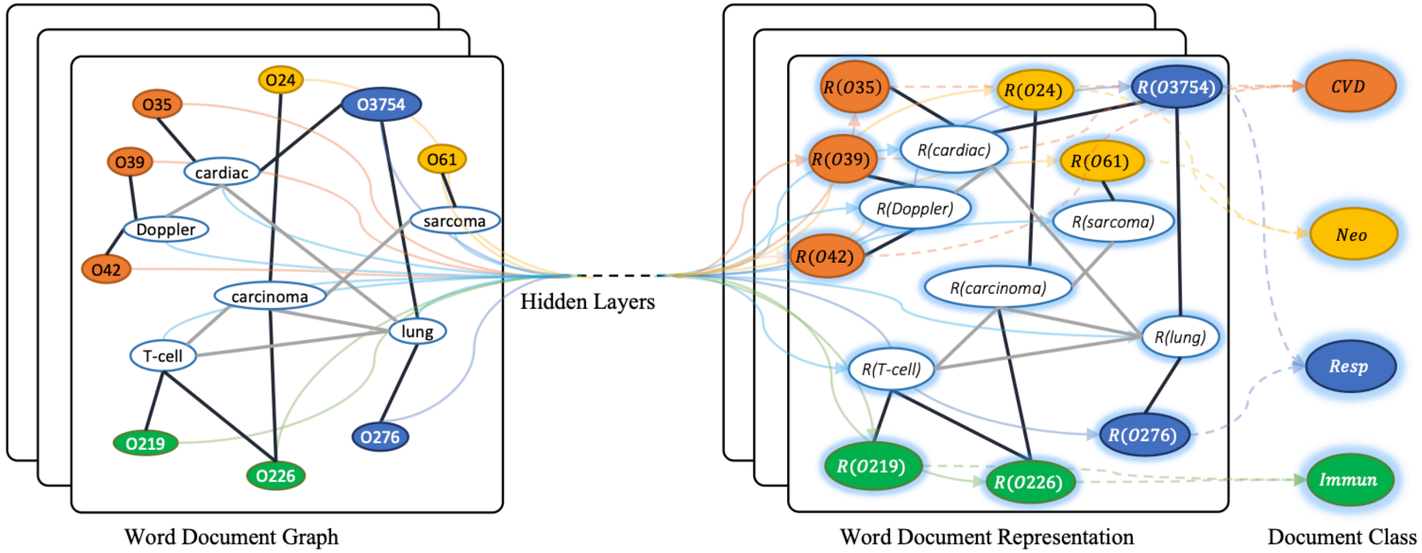
- **Graph Attention Network : GAT (Veličković, 2018)**
  - $h_i^{(n)} = \sum_{v_j \in \mathcal{N}(i)} \alpha_{ij}^{(n)} h_j^{(n-1)} W^{(n)}$, avec $\mathcal{N}(i) \subset \mathcal{V}$ les voisins du sommet i, ainsi que lui-même (boucle) et α_{ij} un poids qui quantifie l'attention qu'accorde le sommet i à la représentation du sommet j
  - Le poids d'attention est proportionnel à une somme pondérée de $W^{(n)} h_i^{(n-1)}$ et $W^{(n)} h_j^{(n-1)}$
    - $\alpha_{ij}^{(n)} = \frac{\exp\left(\text{LeakyReLU}\left(a \cdot \left[W^{(n)} h_i^{(n-1)}, W^{(n)} h_j^{(n-1)}\right]\right)\right)}{\sum_{k \in \mathcal{N}(i)} \exp\left(\text{LeakyReLU}\left(a \cdot \left[W^{(n)} h_i^{(n-1)}, W^{(n)} h_k^{(n-1)}\right]\right)\right)}$, avec $a \in \mathbb{R}^{2d}$ un paramètre supplémentaire de la couche et où $[\cdot, \cdot]$ désigne la concaténation

Application au traitement automatique des documents

Traitement automatique des documents

Classification transductive

- **Graph Convolutional Networks for Text Classification : TextGCN (Yao *et al.* 2019)**
  - Graphe au niveau corpus mêlant sommets-mots et sommets-documents

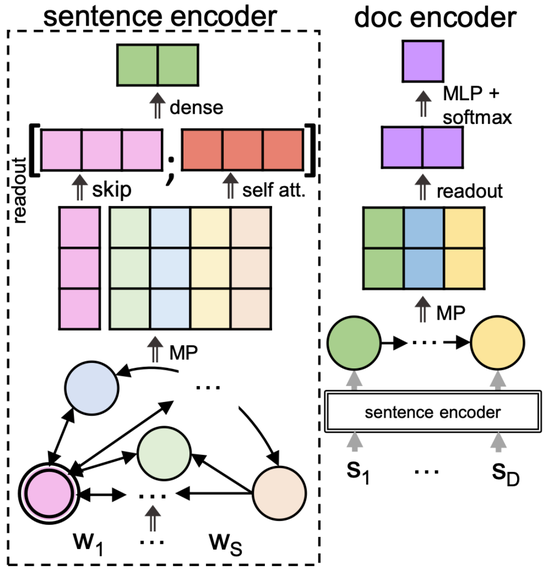


(Figure tirée de « Graph Convolutional Networks for Text Classification », Yao *et al.*, 2019)

Traitement automatique des documents

Classification inductive

- **Message Passing Attention Networks for Document Understanding : MPAD (Nikolentzos *et al.* 2020)**
  - Graphes au niveau document mêlant sommets-mots et sommets-phrases

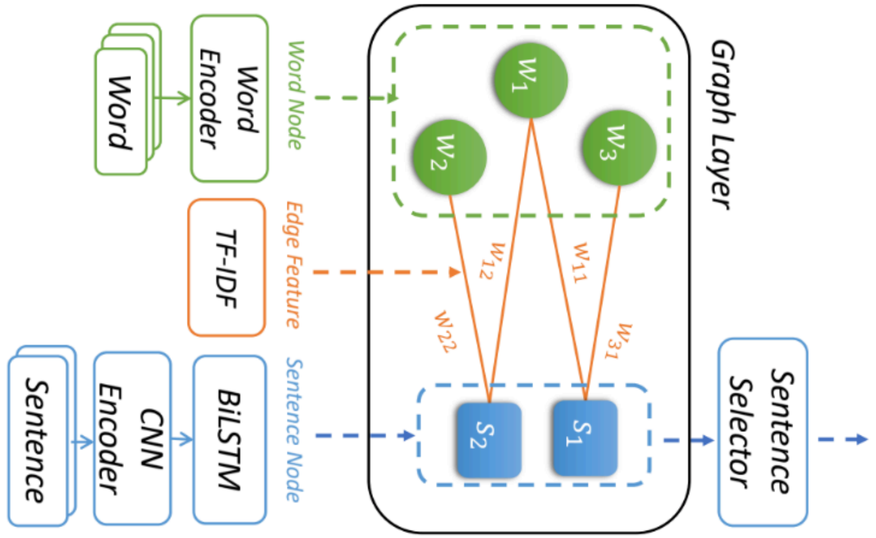


(Figure tirée de « Message Passing Attention Networks for Document Understanding », Nikolentzos *et al.*, 2020)

Traitement automatique des documents

Résumé extractif

- **Heterogeneous Graph Neural Networks for Extractive Document Summarization (Wang *et al.* 2020)**
  - Graphes mêlant sommets-mots et sommets-phrases

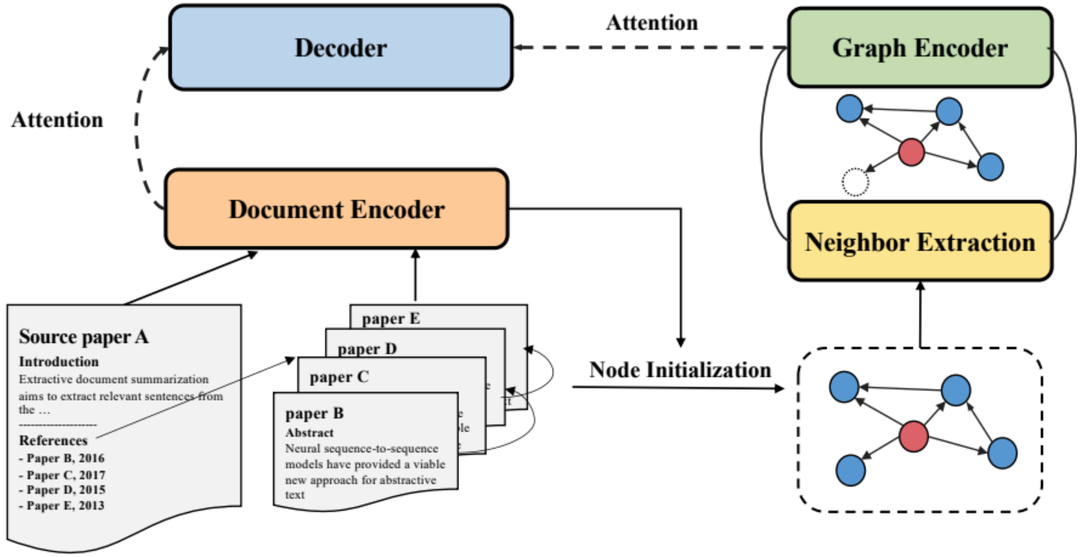


(Figure tirée de « Heterogeneous Graph Neural Networks for Extractive Document Summarization » (Wang et al. 2020)

Traitement automatique des documents

Résumé extractif

- **Enhancing Scientific Papers Summarization with Citation Graph (An *et al.* 2021)**
  - Graphe de sommets-documents



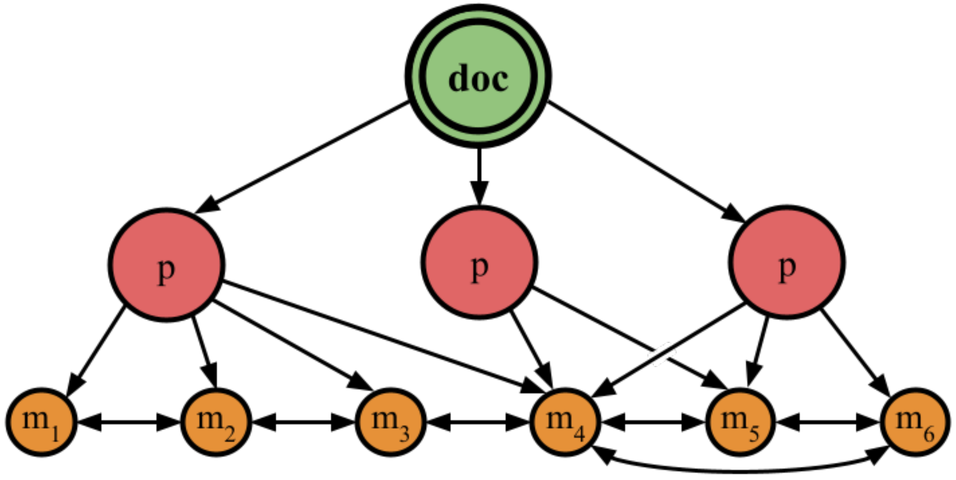
(Figure tirée de « Enhancing Scientific Papers Summarization with Citation Graph » (An et al. 2021)

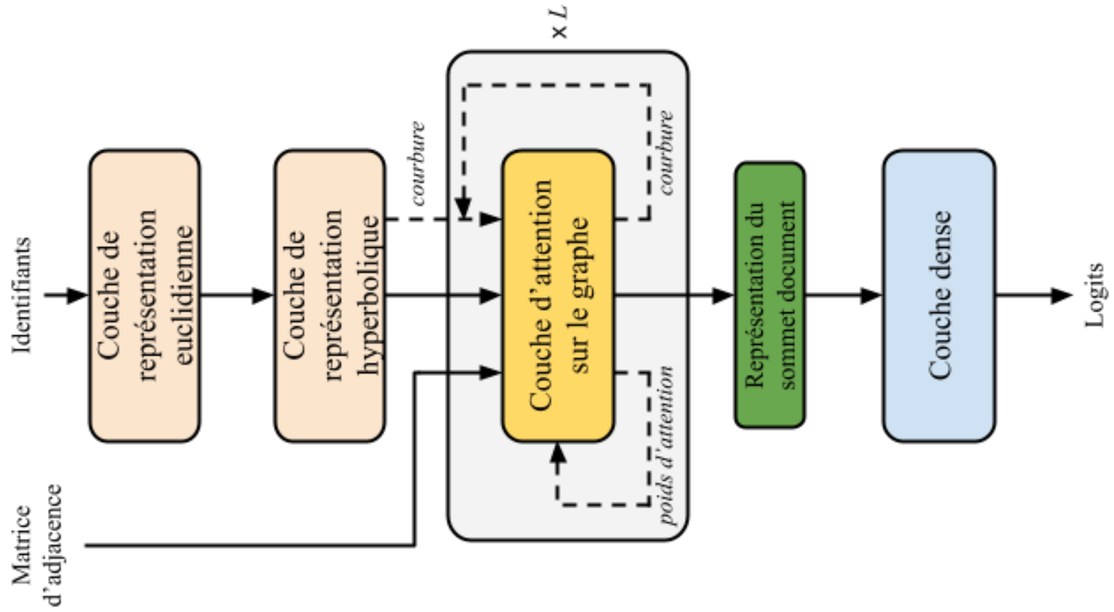
Pistes de recherche

Pistes de recherche

De l'espace euclidien à l'espace hyperbolique

- **Efficient Document Classification with Hyperbolic Hierarchical Graph Neural Networks (Guille *et al.* 2023)**
  - Chaque couche apprend des représentations sur une surface de courbure spécifique

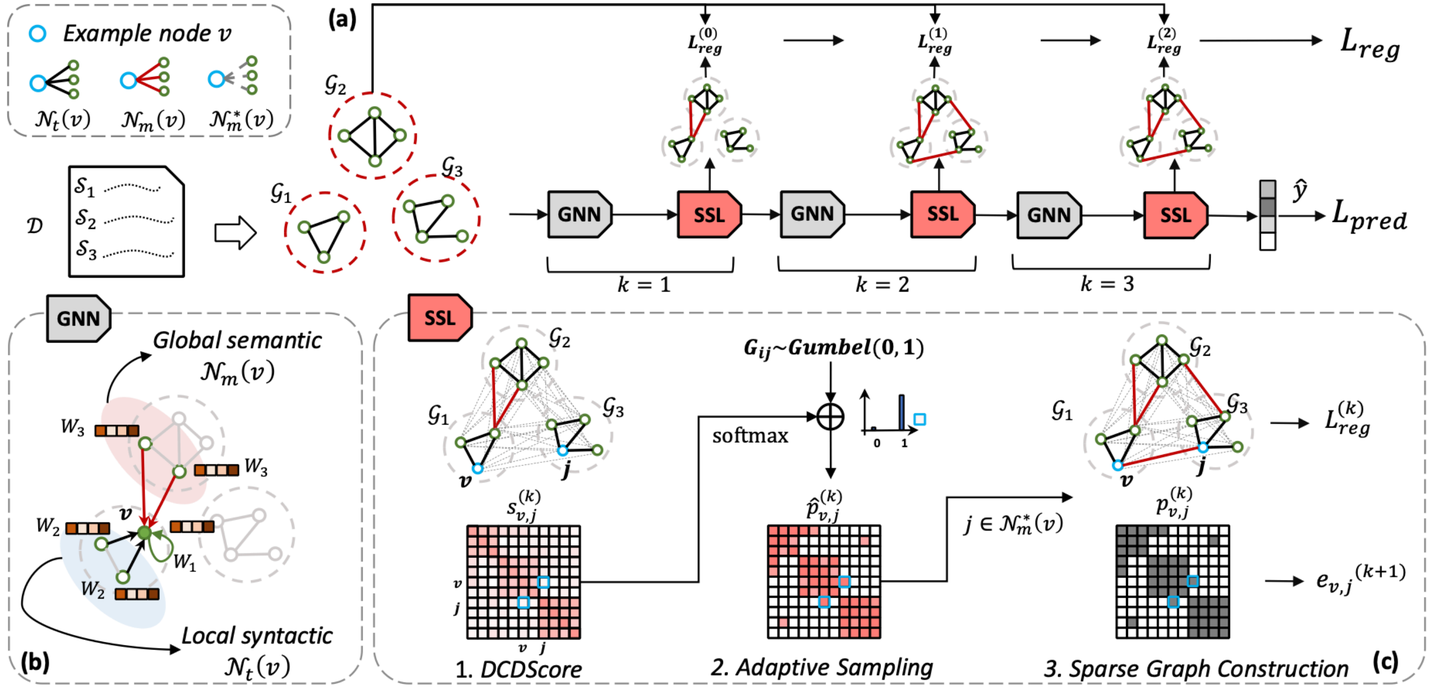




Pistes de recherche

De graphes fixes à des graphes appris

- **Sparse Structure Learning via Graph Neural Networks for Inductive Document Classification (Piao *et al.* 2022)**



(Figure tirée de « Sparse Structure Learning via Graph Neural Networks for Inductive Document Classification » Piao et al. 2022)