Institut de Science Financière et d'Assurances (ISFA)

MASTER 2 - ECONOMETRIE, STATISTIQUES : EQUADE

Projet Apprentissage Statistique - Réseaux de Neurones

Comparaison des modèles LSTM et CNN pour classer des paires de phrases selon les implications

Etudiant :
Crésus KOUNOUDJI

Professeure :
Esterina MASIELLO

Avril 2023

Table des matières

Chapitre 1

Cadre général

Le papier publier par [4], sur le *''Neurone formel"* inspirés par le fonctionnement des neurones biologiques marque le fondement d'un champs aujourd'hui très important de la Méthodes statistique et de l'Intelligence Artificielle avec les découverts des premiers modèles de neurones artificiels, inspirés du fonctionnement des neurones biologiques.

Ces dernières années, avec les enjeux du Big Data dû à disponibilité croissante de données massives, l'explosion des réseaux sociaux et des objets connectés la valorisation et exploitation de cette mine d'information a conduit à un essor fulgurant de ce champ de recherche booster par les avancées en matière de puissance de calcul et de stockage.

Le Deep Learning qui est une branche de l'intelligence artificielle s'appuie sur ces réseaux de neurones artificiels multicouches pour apprendre à partir de données. Ainsi, le Deep Learning a de nombreuses applications comme le computer vision (pour la reconnaissance d'images, la segmentation d'images, la détection d'objets, la classification d'images, la reconnaissance faciale) et le Natural Language Processing (NLP pour la reconnaissance vocale, la classification de textes, la génération de textes, ...).

D'ailleurs, le NLP et le Deep Learning sont étroitement liés, car les algorithmes de Deep Learning sont largement utilisés dans les tâches de traitement du langage naturel. Les réseaux de neurones profonds ont la capacité d'apprendre des représentations de haut niveau à partir de données brutes, ce qui est essentiel pour traiter le langage naturel. Dans le cadre de notre projet de Deep Learning, nous avons comparé deux modèles de réseaux de neurones pour la classification de paires de phrases en fonction de leurs implications. Dans de ce projet on cherche à déterminer quel modèle est le plus efficace pour la classification. Dans ce rapport, nous présenterons le contexte méthodologique de notre étude, l'état de l'art de la classification de phrases, les modèles utilisés (LSTM et CNN), les résultats et les perspectives.

1.1 Objectifs du projet

La classification de phrases est une tâche importante en NLP qui peut être utilisée pour de nombreuses applications, telles que la classification de spam, la détection de la fraude ou la détection des sentiments. Ce projet s'inscrit dans le cadre du cours de Méthodes statistiques et Réseaux de neurones et nous nous intéressons à la classification de paires de phrases selon leurs implications, c'est-à-dire leur relation de causalité. L'objectif spécifique de notre projet est de comparer deux modèles de réseaux de neurones, le LSTM et le CNN, pour cette classification.

Chapitre 2

Contexte Méthodologique

2.1 État de l'Art

La classification de phrases a fait l'objet de plusieurs études en NLP et diverses approches ont été proposées. Les approches traditionnelles utilisent souvent des caractéristiques lexicales et syntaxiques, telles que les parties du discours ou les n-grammes. On peut également faire de la classification de phrases avec des réseaux de neurones grâce à leur capacité à apprendre des représentations de mots et de phrases. En effet, le Deep Learning a démontré de puissantes capacités d'apprentissage de fonctionnalités et a atteint des performances remarquables en computer vision, en reconnaissance vocale et dans le NLP [1]. Originellement dédié à la computer vision, les modèles CNN se sont ensuite révélés efficaces pour le NLP et ont obtenu d'excellents résultats dans l'analyse sémantique, l'extraction de requêtes de recherche, la modélisation de phrases et d'autres tâches traditionnelles du NLP [2].

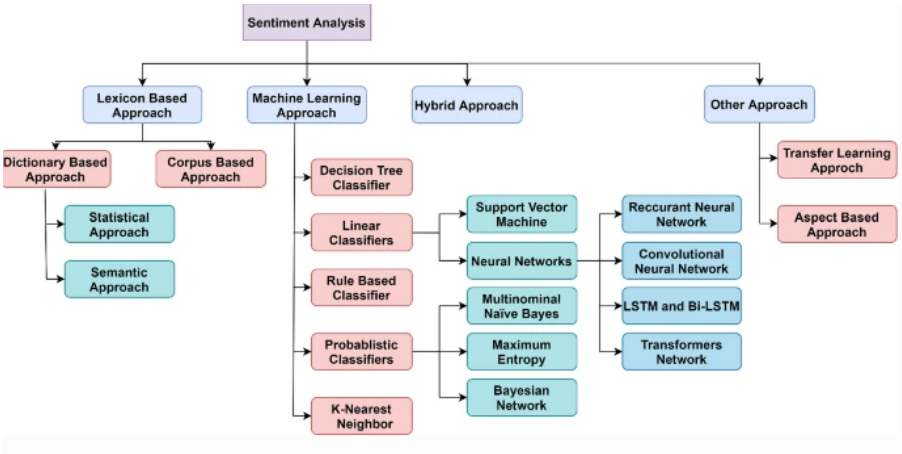

FIGURE 2.1 – Diagramme des différentes approches de l'analyse de sentiment (Wankhade et al., 2022).

Chapitre 3

Modélisation

3.1 Les modèles

Dans ce projet, on a défini un modèle de Réseau de neurones LSTM (Long Short-Term Memory) et un modèle CNN-LSTM (Convolutional Neural Network - Long Short-Term Memory). Le modèle LSTM est composé d'une couche d'embedding qui transforme les indices des mots en vecteurs de dimension 100, suivi d'une couche LSTM de 100 neurones avec un dropout de 0,2 pour éviter l'overfitting et d'une couche dense de sortie avec une fonction d'activation softmax pour la classification multi-classes.

Le modèle CNN-LSTM est similaire, avec une couche de convolution d'un filtre de taille 64 et d'une taille de noyau de 5, suivie d'une couche de pooling pour réduire la dimensionnalité de la sortie de la couche de convolution, un dropout de 0,2 pour éviter l'overfitting, une couche LSTM de 100 neurones, et une couche dense de sortie avec une fonction d'activation softmax pour la classification multi-classes.

Pour évaluer la performance des modèles, on utilise une k-fold cross-validation où les données sont divisées en 10 folds et chaque fold est utilisé une fois comme ensemble de validation et une fois comme ensemble d'entraînement (on a essayé différente valeur pour voir en 2 et 20 si on réduisait significativement l'impact des choix aléatoires lors de la division des données en augmentant ce paramètre). La précision moyenne de chaque modèle est utilisée comme mesure de la performance.

3.1.1 Modèle LSTM

Le LSTM (Long Short-Term Memory) est un type de réseau de neurones récurrents qui est très populaire pour la classification de séquences. Il est capable de mémoriser les informations à long terme et de gérer les informations qui sont espacées dans le temps. Le LSTM parvient à conserver la mémoire pendant un certain temps en ajoutant une "cellule de mémoire". En effet, dans un modèle LSTM standard, une couche est composée de quatre couches de Recurrent Neural Network (RNN) qui sont des types particuliers de réseaux de neurones constituée d'unités interconnectée et dont le graphe de connexion contient au moins un cycle [5]. L'architecture particulière RNN permet de traiter des données qui ne sont pas indépendantes. La cellule de mémoire est principalement contrôlée par "la porte d'entrée", "la porte d'oubli" et "la porte de sortie". La porte d'entrée active l'entrée d'informations dans la cellule de mémoire, et la porte d'oubli efface sélectivement certaines informations dans la cellule de mémoire et active le

stockage à l'entrée suivante. Enfin, la porte de sortie décide de l'information qui sera émise par la cellule de mémoire.

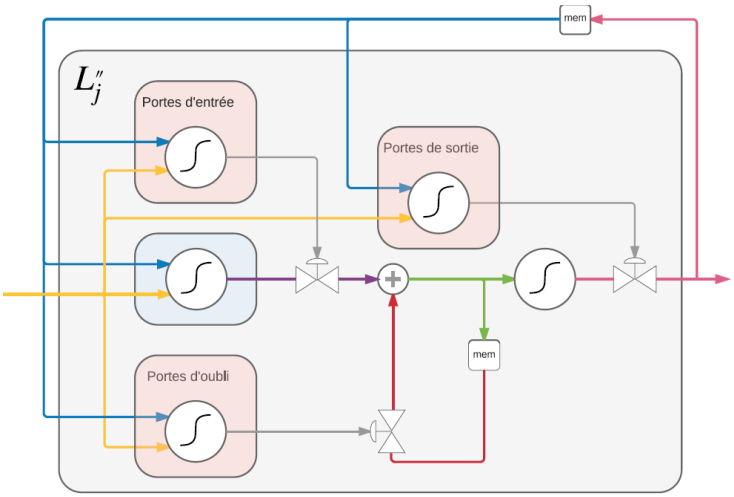

FIGURE 2.2 – Visualisation synthétique de la propagation de l'information lors de la passe avant dans une couche LSTM (sans "peepholes" pour faciliter la lisibilité). Les 3 portes agissent directement sur les vannes qui contrôlent le débit d'information qui provient de l'entrée, celui qui provient de la mémoire interne, mais également l'information qui sort de la couche.

FIGURE 3.1 – Diagramme des différentes approches de l'analyse de sentiment Illustration synthétique de la propagation de l'information dans une couche LSTM (Gregory Gelly 2017).

3.1.2 Modèle CNN

Pour citer [2] et [3], le Convolutional Neural Network (CNN) est un type de réseau de neurones qui est souvent utilisé pour la classification d'images. Cependant, il a également été appliqué à la classification de phrases en NLP. Les CNN peuvent être utilisés pour extraire des caractéristiques locales à partir d'une séquence. D'une manière générale, un CNN se compose d'une séquence de couches convolutives, chacune d'entre elles étant suivie d'une couche de mise en commun et, enfin, de couches entièrement connectées. Les couches convolutives sont des sortes de couches cachées où chaque neurone n'est connecté qu'à un sous-ensemble de neurones d'entrée. Étant donné qu'une couche de convolution est constituée de plusieurs filtres, la répétition de l'opération de convolution pour chaque noyau augmente la dimension de la sortie des couches, et les couches de mise en commun qui suivent les couches de convolution réduisent cette dimension en effectuant une opération d'agrégation (comme le maximum de la moyenne) sur la sortie. Cette sortie peut ensuite être aplatie et introduite dans les couches entièrement connectées, qui ne sont que des Multi-Layer Perceptron, et la prédiction sera effectuée. Dans ce processus, les couches convolutives agissent comme des méthodes d'extraction de caractéristiques et ont montré leur efficacité dans de nombreuses tâches prédictives.

3.2 Le jeu de données

Les données SICK dataset (NLP - Sentence Textual Similarity) utilisées pour ce projet sont issues de la plateforme Kaggle et sont constitué d'une table de 5 colonnes et 4500 lignes. La première colonne *paire ID* contient des identifiants uniques pour chaque paire de phrases

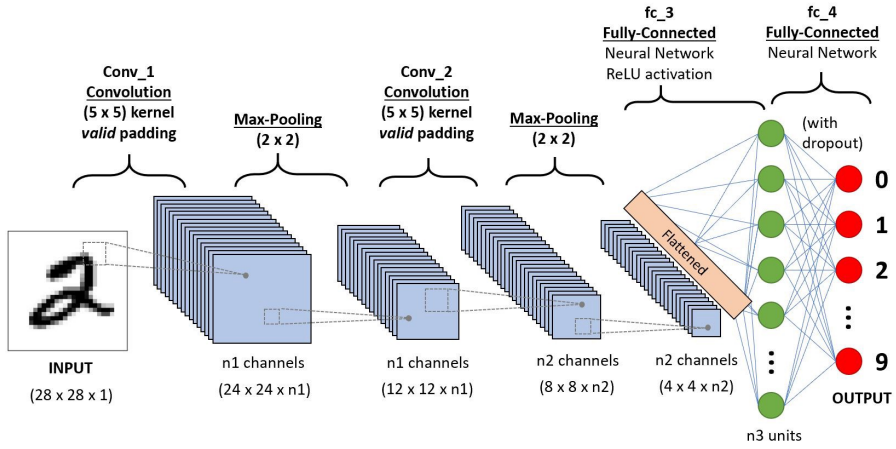

FIGURE 3.2 – Illustration de l'architecture d'un CNN (Phani Ratan, 2020).

dans la base de données. Les colonnes suivantes *sentence A* et *sentence B* contiennent respectivement la première phrase de la paire et la deuxième phrase de la paire. Dans la colonne *relatedness score* on a score de proximité entre les *phrases A* et *B*, qui mesure leur similarité sémantique. Le score est compris entre 1 et 5, où 1 indique une faible similarité sémantique et 5 une forte similarité. La dernière colonne est *entailment judgment* indique avec les modalités *ENTAILMENT*, *CONTRADICTION* ou *NEUTRAL* ou si la *phrase A* implique, contredit ou n'a aucun lien avec la *phrase B*.

	pair_ID	sentence_A	sentence_B	relatedness_score	entailment_judgment
0	1	A group of kids is playing in a yard and an ol...	A group of boys in a yard is playing and a man...	4.5	NEUTRAL
1	2	A group of children is playing in the house an...	A group of kids is playing in a yard and an ol...	3.2	NEUTRAL
2	3	The young boys are playing outdoors and the ma...	The kids are playing outdoors near a man with ...	4.7	ENTAILMENT
3	5	The kids are playing outdoors near a man with ...	A group of kids is playing in a yard and an ol...	3.4	NEUTRAL
4	9	The young boys are playing outdoors and the ma...	A group of kids is playing in a yard and an ol...	3.7	NEUTRAL

FIGURE 3.3 – Cinq première ligne de la table de données SICK dataset.

3.3 Résultats

3.3.1 Analyse descriptive

Il s'agit d'analyser les données, voir s'il y a des anomalies comprendre pourquoi ces valeurs sont différentes de la moyenne et voir si elles des résultats valides mais inattendus ou des erreurs de prédiction qui peuvent biaisés les estimations. Ainsi, on pourra ajuster les hyperparamètres pour avoir des modèles performant, en collectant plus de données d'entraînement ou en utilisant un algorithme différent. Si elles sont des résultats valides mais inattendus on peut regarder pourquoi ces cas sont différents et comment vous pouvez les traiter différemment pour améliorer les performances. Mais en règle général, on utilise des ensembles de modèles, ou plusieurs modèles ensemble pour améliorer les performances et réduire l'impact des valeurs aberrantes.

La figure 3.4 ci-dessous représente le nombre de pairs de phrase par Classe d'implication. On remarque que les modalités "contradiction" son relativement mal représentée dans la base. Cela implique que qui la répartition des base "test" et "train" on peut avoir un nombre différent de modalité dans chaque groupe te donc être source de biais des prédictions des modèles entraînés.

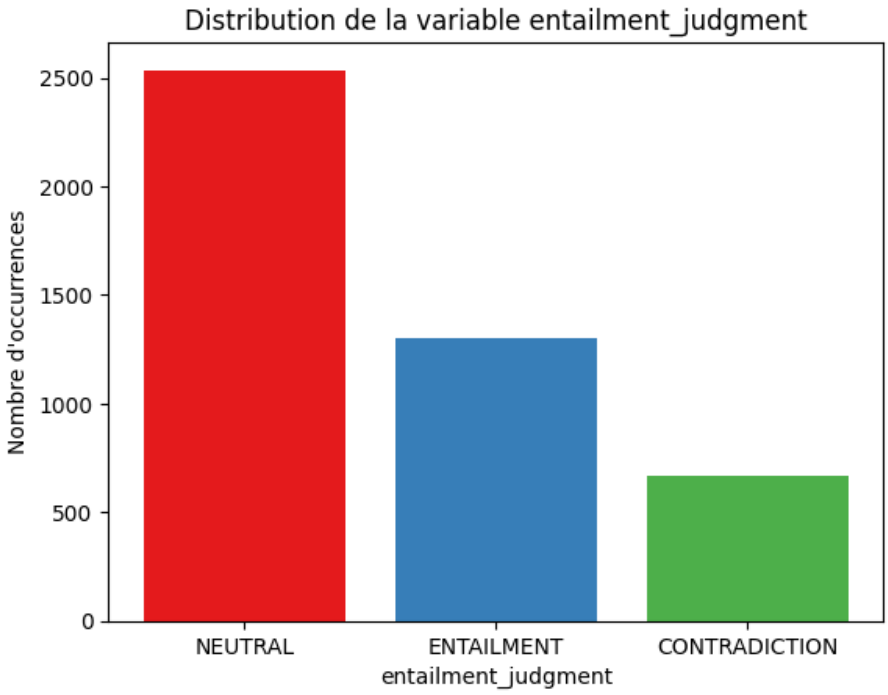

FIGURE 3.4 – Distribution de la variable de Jugement de l'implication.

Les figures 3.5 et 3.6c ci-dessous, représente respectivement la distribution des scores de parenté et la distribution des scores de parentés par classe. Dans le premier cas on remarque que les scores de parentés sont globalement étalés vers ma gauche et que dans le second cas, il y a une certaine hétérogénéité dans la distribution à l'intérieure des classes.

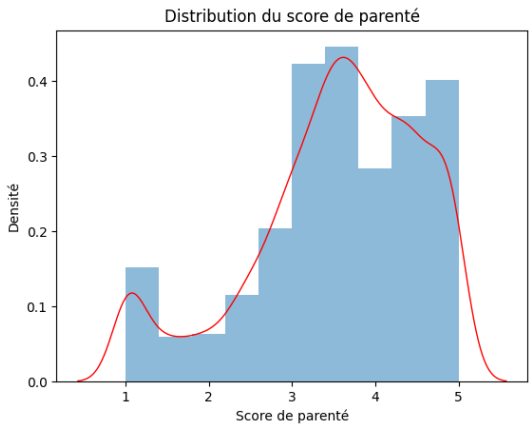

FIGURE 3.5 – score de parenté.

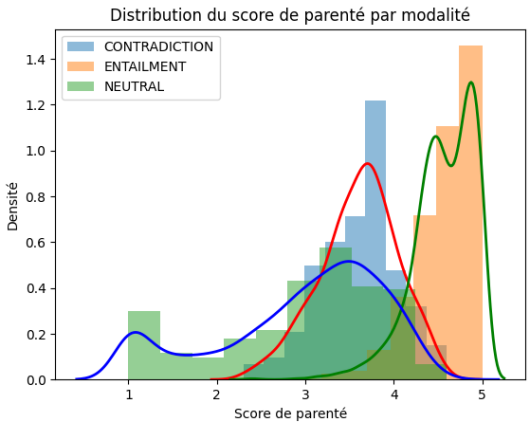

FIGURE 3.6 – score de parenté par implication

3.3.2 Modèles de prédiction

La cross-validation a permis de construire les matrices de confusion 3.7 ci-dessous et de calculer les performance des modèles. On obtient 63.00% en moyenne de prformance pour le modèle LSTM et 66.53% pour le CNN.

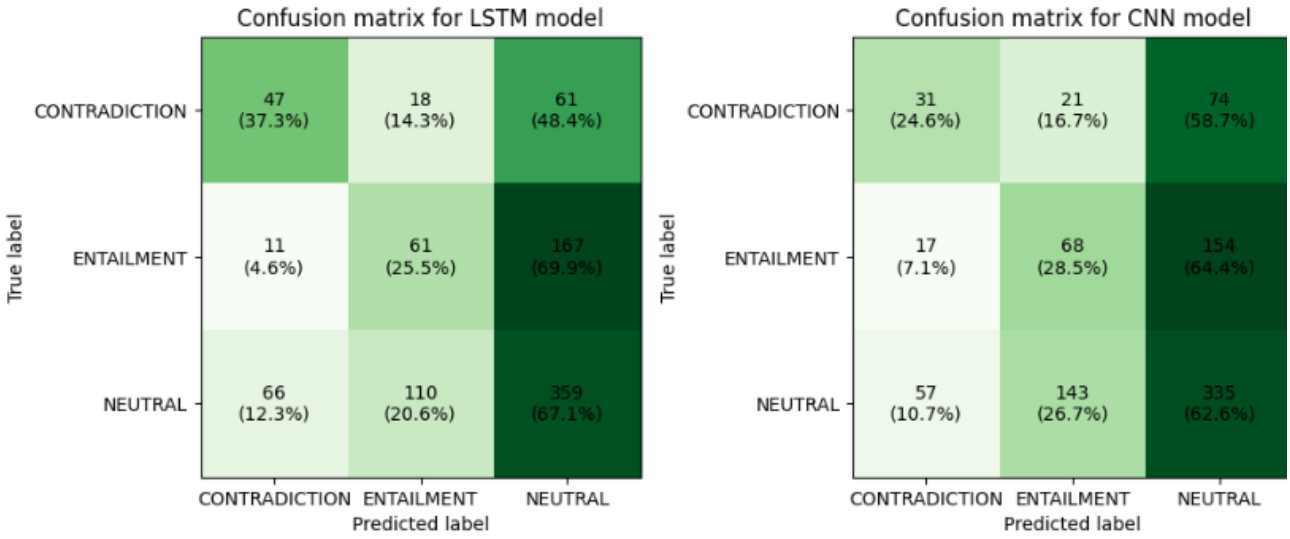

FIGURE 3.7 – Matrice de confusion des modèles LSTM et CNN .

De plus le graphique 3.8 ci-après compare les performances des deux modèles. Sur les boîtes à moustache, nous avons la distribution des données de performance, où la ligne médiane est la médiane des données, la boîte inférieure représente le premier quartile des données (25%), la boîte supérieure représente le troisième quartile des données (75%), et les points à l'extérieur de la boîte sont des données aberrantes. Sur ce graphique ci-dessus, la boîte à moustaches du LSTM est beaucoup plus écrasée sur la médiane que celle du CNN, indiquant que les performances du LSTM sont plus cohérentes que celles de CNN. Ainsi, les performances du modèle CNN sont un peu meilleures et plus cohérentes que celles de LSTM à priori.

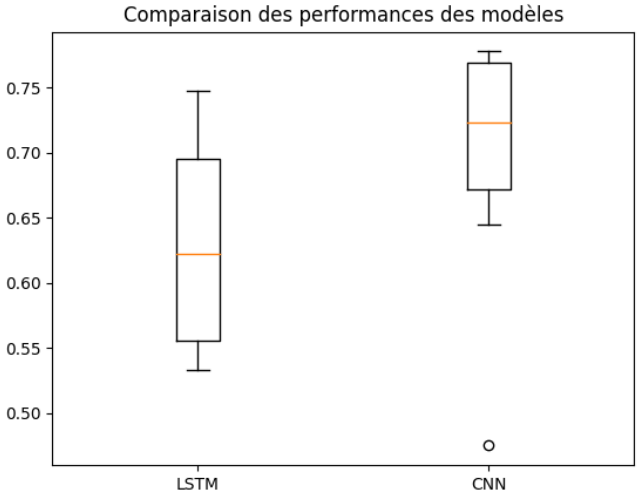

FIGURE 3.8 – comparaison graphique de la performance du modèle LSTM et CNN .

Cependant, après plusieurs tests, il est important de noter que la cohérence du modèle LSTM dépends beaucoup du paramètre n_split du $K - fold$ cross validation.

Chapitre 4

Conclusion et perspectives

Dans ce projet et avec nos données, les performances du modèle CNN sont un peu meilleures et plus cohérentes que celles de LSTM à priori. La cohérence du modèle LSTM dépends beaucoup du paramètre n_split du $K - fold$ cross validation.

Pour améliorer les modèles, on peut jouer sur les hyperparamètres tels que le dropout, le nombre de neurones, le nombre d'époques, le batch size, le nombre de filtres de la couche de convolution, etc. On peut également ajouter ou de supprimer des couches pour changer l'architecture du modèle.

Bibliographie

[1] Yahui Chen. "Convolutional Neural Network for Sentence Classification". In : (2015). URL : http://hdl.handle.net/10012/9592 (visité le 02/05/2023).

[2] Yoon Kim. "Convolutional neural networks for sentence classification". In : (2014). (Visité le 02/05/2023).

[3] Annavarapu Chandra Sekhara Rao Chaitanya Kulkarni Mayur Wankhade. "A survey on sentiment analysis methods, applications, and challenges". In : (7 fév. 2022). URL : https://link.springer.com/article/10.1007/s10462-022-10144-1 (visité le 02/05/2023).

[4] Warren S. McCulloch Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In : (déc. 1943). URL : https://link.springer.com/article/10.1007/bf02478259 (visité le 02/05/2023).

[5] Warren S. McCulloch Walter Pitts. "Reseaux de neurones recurrents pour le traitement automatique de la parole". In : (). URL : https://theses.hal.science/tel-01615475 (visité le 02/05/2023).