Remerciements

Je tiens à remercier M. Christophe Cerisara, qui a élaboré un sujet passionnant pour ce stage, et qui a su m'accompagner tout au long de cette aventure, malgré un emploi du temps chargé et des responsabilités nombreuses en tant que chef d'équipe.

Je remercie Mme Nadia Bellalem, Mme Christine Fay-Varnier et M. Samuel Cruz-Lara avec qui j'ai brièvement collaboré au sein du projet PAPUD.

Je remercie également Mme Jeanine Souquières, qui m'a conseillé lors de l'élaboration de ce rapport.

Je tiens tout particulièrement à remercier M. Maxime Amblard, sans qui je n'aurai pas obtenu ce stage.

Mais je remercie aussi tous les stagiaires et doctorants qui ont supporté mon humour pendant plus de 3 mois, ainsi que les stagiaires de l'IDMC qui le supportent depuis bien plus longtemps.

Enfin, je remercie Annick Jacquot, qui s'est chargée de toutes les questions administratives de mon stage ; Caroline et toute l'équipe de la cafétéria, qui m'ont offert un cadre inoubliable ; et tous les gens du LORIA qui m'ont accueilli chaleureux et qui font vivre ce laboratoire.

Une mention spéciale pour mon PC agonisant, qui m'a supporté tout au long du stage et de la rédaction de ce rapport.

Avant-propos

La lecture du présent mémoire ne nécessite aucune connaissance préalable en traitement automatique de la langue, en apprentissage automatique ou en apprentissage profond.

Il est cependant préférable d'avoir des connaissances en informatique pour comprendre les enjeux du stage.

Par ailleurs, de nombreux termes techniques sont utilisés.

Bien qu'habituellement utilisés sous leur forme anglaise dans la littérature du domaine, dans ce rapport tous les termes qui possèdent une traduction française ont été traduits. Les seules exceptions sont certains sigles, qui sont utilisés tels quels dans les textes français.

C'est pourquoi tous les termes techniques seront toujours introduits en français, accompagnées d'une explication, de la traduction anglaise et du sigle anglais si nécessaire.

De plus, une partie du rapport est dédiée à l'explication des termes et concepts utilisés, et un glossaire est présent en fin d'ouvrage.

Enfin, quelques notes de bas de page fournissent des précisions ponctuelles sur certains concepts utilisées. Elles sont annoncées en exposant.

Les références aux sources sont marquées entres crochets, et numérotées par ordre d'utilisation.

Ce rapport a été réalisé avec LaTeX et les diagrammes présentés ont été produits par nos soins avec l'outil Ti*k*Z.

Table des matières

1 Introduction

1.1 Contexte et enjeux du stage

D'une part, depuis quelques années, l'apprentissage profond (*Deep Learning* en anglais) et les réseaux de neurones artificiels, ou plus simplement réseaux de neurones (*Neural Networks* en anglais) ont connu une explosion de popularité. Ce qui se cache derrière cet engouement est la combinaison de théories relativement anciennes et d'avancées technologiques permettant la mise en œuvre desdites théories.

Un des domaines exploitant les performances de ces nouveaux outils est le traitement automatique des langues (TAL, *Natural Language Processing* en anglais). En particulier, nous nous intéresserons dans ce rapport à la création de modèles de la langue (*Language Models* en anglais, LM).

D'autre part, nous sommes à l'ère du *« Big Data »*, et les quantités de données produites de nos jours sont bien au-delà de ce que nous pouvons gérer sans l'aide d'outils spécialisés. Afin de produire des outils adaptés aux échelles actuelles, nous nous intéressons dans ce rapport à des grands volumes de données bien au-delà des volumes habituellement utilisés en apprentissage profond.

Ainsi, l'axe principal de ce rapport est l'application des méthodes de l'apprentissage profond sur de grands volumes de données, de la perspective du TAL. Cela implique des problématiques relativement classiques en développement de réseau de neurones artificiels : le choix de l'architecture du réseau, de l'algorithme d'entraînement, mais aussi des questions plus pragmatiques d'optimisation liées au volume de données.

1.2 Objectifs du stage

Deux objectifs se sont succédé durant le stage.

L'objectif initial du stage était d'explorer une idée d'architecture innovante de réseau de neurones artificiels, imaginée par Mr. Cerisara, le maître de stage. Il s'agit du projet GMSNN (réseau de neurones récurrents multi-échelles croissant, *Growing Multi-Scale Recurrent Neural Network* en anglais, voir chapitre 4, page 21).

Cependant, à l'issue du deuxième mois du stage, nous avons changé d'objectif.

Le nouvel objectif a été la réalisation d'un réseau de neurones artificiels et des outils nécessaires à son utilisation, en mettant à profit les connaissances acquises durant la première partie du stage. Cette réalisation doit servir de base technique pour une partie du projet PAPUD (*Profiling and Analysis Platform Using Deep Learning*, voir chapitre 10, page 53).

1.3 Plan du rapport

Dans un premier temps, nous avons présenté à la fois le contexte, les enjeux, et les objectif généraux du stage.

Dans un second temps, nous définirons les termes et concepts principaux utilisés dans ce rapport.

La complexité du stage est qu'il est composé de deux projets, le projet GMSNN et le projet PAPUD, l'un étant la continuation de l'autre. En effet, les conclusions tirées du premier projet ont servi de base pour le second. C'est pourquoi, pour simplifier la lecture du présent rapport, chaque projet sera traité suivant le même plan.

Notamment, pour chacun d'eux, nous présenterons le contexte, les enjeux, et les entités impliquées dans le projet. Nous décrirons ensuite le modèle réalisé avant de rapporter le travail effectué. Enfin, une conclusion résumera les points majeurs du projet.

Dans un dernier temps, nous ferons un bilan de l'ensemble du travail réalisé pour en tirer les apports principaux.

2 Terminologie et concepts fondamentaux

Cette section est dédiée à la présentation et l'explication des théories, termes et concepts nécessaires à la compréhension du présent rapport.

L'objectif n'est pas de fournir des explications approfondies, mais de fournir les connaissances minimales nécessaires à la compréhension du contenu et des enjeux du stage.

Les termes présentés sont répertoriés dans le glossaire, mais aucun rappel de leur présence dans celui-ci n'est donné au long du texte.

2.1 Apprentissage automatique, modèle, entraînement et données

2.1.1 Apprentissage automatique

L'apprentissage automatique (*Machine Learning* en anglais) est un ensemble de « méthodes [statistiques] permettant à une machine (au sens large) d'évoluer par un processus systématique, et ainsi de remplir des tâches difficiles ou problématiques par des moyens algorithmiques plus classiques », d'après Wikipédia [1].

2.1.2 Modèle

Un modèle en apprentissage automatique (*Machine Learning* en anglais) est la représentation du monde construite afin de répondre au problème à résoudre.

On parle d'entrées pour désigner les données fournies au modèle, et de sorties pour désigner les données produites par ce modèle.

2.1.3 Entraînement du modèle

L'entraînement du modèle, aussi appelé apprentissage, est le processus par lequel on adapte le modèle de façon à mieux résoudre le problème.

2.1.4 Données d'entraînement

Pour entraîner un modèle, il faut lui fournir des données. Voici quelques termes courants se référant aux données d'entraînement :
- le corpus : l'ensemble des données d'entraînement ;
- un exemple : un fragment du corpus utilisé pour entraîner un modèle ;
- une époque : un cycle complet d'entraînement sur le corpus.

Généralement, on effectue un prétraitement des données (*preprocessing* en anglais) pour les préparer. Cela peut consister à retirer les données erronées, à en adapter le format, à les anonymiser, ou encore à associer le résultat attendu aux données correspondantes.

2.2 Apprentissage profond et réseaux de neurones

2.2.1 Apprentissage profond

L'apprentissage profond représente un ensemble de techniques d'apprentissage automatique consacrées aux réseaux de neurones.

Faisant partie des méthodes d'apprentissage automatique, l'apprentissage profond regroupe à la fois les méthodes de création, d'entraînement, d'optimisation et d'utilisation des modèles basés sur des réseaux de neurones.

2.2.2 Réseau de neurones artificiels

Un réseau de neurones artificiels ou plus simplement réseau de neurones (*Neural Network* en anglais) est un modèle mathématique composé d'éléments interconnectés nommés neurones formels, par analogie lointaine avec le fonctionnement des neurones biologiques.

Un réseau de neurones artificiels prend en entrée un tenseur (*tensor* en anglais, un type de matrice spécifique utilisé en apprentissage profond), et produit un tenseur en sortie. Toutes les valeurs contenues dans ces tenseurs sont des nombres.

2.2.3 Architecture et modules

Pour des raisons de concision, nous considérons les réseaux de neurones comme des modules, c'est-à-dire des boites noires, sans nous intéresser à leur conception interne.

Nous parlons d'architecture pour désigner à la fois la façon dont sont conçus les réseaux de neurones et la façon dont sont assemblés les modules pour former un modèle.

Pour décrire les architectures, nous utilisons des diagrammes considérant les modules comme des blocs.

2.2.4 Paramètre

Un paramètre pour un module ou un modèle est une valeur qui varie au cours de l'entraînement. Généralement, plus un modèle possède de paramètres, plus son entraînement consomme de ressources mais plus la qualité de l'apprentissage est élevée.

2.2.5 Principales architectures de réseaux de neurones

La liste suivante présente les principales architectures de réseaux de neurones utilisées dans ce rapport, classées en ordre croissant de complexité et de consommation de ressources.

- Le réseau de neurones artificiels intégralement connecté ou module linéaire est un des plus simples. Il est emblématique des réseaux de neurones.
- Le réseau de neurones récurrents (*Recurrent Neural Network* en anglais, RNN) est une architecture de réseaux de neurones particulièrement adaptée au traitement de séquences. Le RNN traite des éléments les uns après les autres, et stocke dans une « mémoire » des informations au fur-et-à-mesure. Il utilise les informations stockées pour améliorer la façon dont il traite les éléments suivants.
- Le réseau récurrent à mémoire à court et long terme (*Long Short Term Memory* en anglais, LSTM) est un RNN particulier. Équipé d'une mémoire supplémentaire, il plus puissant mais plus coûteux à entraîner qu'un RNN classique.

2.3 Traitement automatique des langues (TAL) et modèle de la langue

2.3.1 Traitement automatique des langues (TAL)

Le traitement automatique des langues (TAL, *Natural Language Processing* en anglais) est une discipline qui s'intéresse au traitement des informations langagières par des moyens formels ou informatiques.

2.3.2 Modèle de la langue

Un modèle de la langue (*Language Model* en anglais, LM) est une « distribution de probabilité sur une séquence de mots [ou de caractères] » (d'après Wikipédia [2]) utilisée pour estimer la probabilité d'apparition du prochain mot ou caractère.

Autrement dit, c'est une représentation servant à prédire le mot suivant à partir des mots précédents (ou le caractère suivant à partir des caractères précédents).

2.3.3 Contexte et dépendances

Dans le cadre d'un modèle de la langue, le contexte est l'ensemble des informations disponibles hors du mot ou caractère à prédire.

On parle aussi de dépendances entre d'une part les mots ou caractères et d'autre part l'élément du contexte correspondant.

Parmi les informations contenues dans le contexte d'un mot, on peut trouver aussi bien le sens des mots environnants que la structure syntaxique de la phrase, ou encore des informations plus générales comme le fait que les chats sont des mammifères.

On considère que, plus on a d'informations contextuelles, plus le modèle de la langue est précis. Par exemple, si on nous dit « un chat », il sera plus difficile de prédire la couleur du chat que si on nous dit « un chat de couleur sombre ».

2.4 Performance et mesure

Le dernier concept important du rapport est celui de performance des modèles produits.

La performance d'un modèle est évalué par trois composantes :
— la qualité maximale des résultats produits ; dans le cas d'un modèle de la langue, il s'agit de la qualité de la prédiction ;
— le temps d'entraînement nécessaire pour atteindre cette qualité (nombre d'époques, durée, etc.) ;
— la consommation de ressources nécessaire pour atteindre cette qualité (mémoire, puissance de calcul, …).

Afin d'évaluer la performance, des mesures ont été définies :
— pour mesurer la qualité du résultat, on utilise l'écart entre le résultat produit et le résultat attendu ; une mesure définie dans la littérature et utilisée dans les annexes est le BPC [1] ;
— pour mesurer le temps d'entraînement, on le nombre d'époque et le temps nécessaire par époque, la durée totale en heures, etc. ;
— pour mesurer la consommation de ressources, on évalue l'espace mémoire occupée (en MiB ou GiB), la puissance de calcul utilisée (pourcentage de la puissance disponible), etc.

Un modèle optimal serait un modèle qui atteint d'excellents résultats (l'écart entre ce qui est attendu et ce qui est produit le plus faible possible), le plus rapidement possible, en consommant le moins de ressources possibles.

1. Le BPC (*Bits Per Character* en anglais) [50] est originalement une mesure de la qualité de compression de texte, mais plusieurs papiers ont détourné cette mesure en s'en servant d'estimation de la qualité d'un modèle de la langue au niveau du caractère. Dans cet usage, le BPC est assimilable à une mesure de la précision des résultats du modèle de la langue. Plus le BPC est proche de 0 plus la qualité du modèle de la langue est élevé.