

Stage de recherche

Réseaux de Neurones Multi-Échelles
pour la Modélisation de la Langue
à partir de Grands Volumes de Données

Esteban Marquer

Année 2017–2018

Projet réalisé pour l'équipe SYNALP du laboratoire LORIA

Maître de stage : Christophe Cerisara

Parrain universitaire : Jeanine Souquière

Stage de recherche

Réseaux de Neurones Multi-Échelles
pour la Modélisation de la Langue
à partir de Grands Volumes de Données

Esteban Marquer

Année 2017–2018

Projet réalisé pour l'équipe SYNALP du laboratoire LORIA

Esteban Marquer
marquer.esteban@etu.univ-lorraine.fr

Institut des Sciences du Digital Management & Cognition
193 avenue Paul Muller,
CS 90172, VILLERS-LÈS-NANCY
+33 (0)3 72 74 16 18
idmc-contact@univ-lorraine.fr

LORIA
Campus scientifique
BP 239
54506, Vandoeuvre-lès-Nancy Cedex
+33 (0)3 83 59 20 00



Encadrant : Christophe Cerisara

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 et M. Samuel Cruz-Lara avec qui j'ai brièvement collaboré au sein du projet PAPUD.

Je remercie aussi que 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 chargé 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 offert un accueil chaleureux et qui font vivre ce laboratoire.

Table des matières

Remerciements	iii
1 Introduction	3
2 Cadre théorique	5
I Projet GMSNN	11
3 Présentation du laboratoire et de l'équipe	12
4 Architecture innovante de réseau de neurones pour l'élaboration de modèle du langage	15
5 Données disponibles	19
6 Description de l'architecture proposée	21
7 Réalisation	25
8 Conclusions sur le projet GMSNN	39
II Projet PAPUD	41
9 Présentation du projet ITEA3-PAPUD, cas d'utilisation BULL	42
10 Projet PAPUD	45
11 Données disponibles	49
12 Modèle à réaliser	51
13 Réalisation	53

14 Conclusions sur le projet PAPUD	57
15 Rétrospective sur le stage	59
16 Discussion et perspectives	60
17 Rétrospective sur le stage	61
Annexes	63

1 Introduction

1.1 Contexte et enjeux du stage

D'une part, depuis quelques années, le « *Deep Learning* » et les « réseaux de neurones » 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 outils est le Traitement Automatique des Langues (TAL, *Natural Language Processing* en anglais). L'application au TAL des réseaux de neurones artificiels ou réseaux de neurones (*Neural Networks* en anglais) qui nous intéressera particulièrement dans ce mémoire est la création de Modèle de la Langue.

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 outils spécialisés. Afin de produire des outils adaptés aux échelles actuelles, nous nous intéresseront dans ce rapport à des grands volumes de données bien au-delà des volumes habituellement utilisés en apprentissage profond (*Deep Learning* en anglais).

Ainsi, nous explorerons la question de l'application des méthodes du apprentissage profond du point de vue du TAL sur des ensembles de données de grande taille.

L'axe principal de ce mémoire est l'application de telles méthodes sur des gros volumes de données. Cela implique à la fois des problématiques relativement classiques en développement de réseau de neurones artificiels d'architecture du réseau, et de choix d'algorithme d'entraînement ; mais aussi des questions plus pragmatiques d'optimisation lié au volume de données.

1.2 Objectifs du stage

Deux objectifs successifs se distinguent dans le stage.

L'objectif initial du stage est d'explorer une idée d'architecture de réseau de neurones artificiels innovante, imaginée par notre maître de stage Mr Cerisara.

À l'issue du deuxième mois du stage, au vu des résultats de l'architecture et de l'évolution du contexte, la mission du stage a aussi évolué.

Le nouvel objectif est 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 servira de base technique pour une partie du projet PAPUD.

Les tenants et aboutissants des deux objectifs seront expliqués en détails dans le chapitre 4 et le chapitre 10.

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 étudierons plus en détail le cadre théorique du rapport, afin de définir les termes et concepts principaux utilisés dans ce rapport.

Dans un troisième temps, nous allons nous attarder plus en détail sur les différentes entités impliquées, en particulier sur l'équipe SYNALP (*SYmbolic and statistical NATural Language Processing*) et ses entités parentes, ainsi que sur le projet PAPUD.

Dans un quatrième temps, nous allons décrire plus en détail les deux aspects du stage : l'idée d'architecture de réseau de neurones artificiels et l'intérêt d'une telle architecture d'un côté, et le projet PAPUD, ses implications et la portée du stage dans ce projet. Nous verrons aussi en quoi le projet PAPUD est dans la continuité de la première partie du stage.

Dans un cinquième temps, nous exposerons le déroulement pas-à-pas du stage, avec les obstacles rencontrés, la façon de les surmonter, et en quoi chaque résultat entraîne l'étape suivante.

Dans un sixième et dernier temps, nous ferons une rétrospective sur l'avancement des objectifs, la qualité des résultats obtenus, et les apports du stage.

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 étudierons plus en détail le cadre théorique du rapport, afin de définir les termes et concepts principaux utilisés dans ce rapport.

Dans un troisième et un quatrième temps nous développerons les deux parties du stage, le projet Réseau de Neurones Récurrents Multi-Échelles Croissant (*Growing Multi-Scale Recurrent Neural Network* en anglais, GMSNN) et le projet PAPUD.

Pour cela, nous présenterons le contexte, les enjeux, et les entités impliquées dans le projet. Nous décrirons ensuite les modèles implantés, avant de dérouler le travail effectué. Enfin, une conclusion résumera les points majeurs du projet.

Dans un cinquième et dernier temps, nous ferons une rétrospective sur l'ensemble du travail réalisé.

2 Cadre théorique (terminologie et concepts fondamentaux)

2.1 Introduction

Ce chapitre est dédié à la présentation et l'explication des théories, termes et concepts nécessaires à la compréhension du présent mémoire. Cependant, il ne s'agit pas d'explications approfondies mais d'une première approche L'objectif n'est pas de fournir des explications approfondies, mais plus de fournir les connaissances minimales à la compréhension du contenu et des enjeux du rapport.

De nombreux extraits provenant de Wikipédia seront utilisés dans cette partie. L'intérêt desdits extraits n'est pas leur exactitude scientifique, mais leur capacité à cerner le sens global des termes en quelques phrases simples. D'autres sources plus fiables seront utilisées en complément, afin d'affiner les explications.

Dans un premier temps, les domaines scientifiques dans lequel s'inscrit le stage seront brièvement définis. Dans un second temps, les concepts et termes fondamentaux utilisés dans ce mémoire seront expliqués. Enfin, pour situer le stage dans son contexte scientifique actuel, un aperçu de l'état de l'art dans la littérature sera donné.

2.2 Domaines scientifiques et techniques concernés

Ce stage s'inscrit dans deux principaux domaines :

- l'apprentissage profond, une des branches de l'apprentissage automatique (*Machine Learning* en anglais) ;
- le Traitement Automatique des Langues (TAL, *Natural Language Processing* en anglais).

2.2.1 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 Wikipédia¹.

1. [wiki_ml](#).

2.2.2 Apprentissage profond

L'apprentissage profond représente un ensemble de techniques de apprentissage automatique, basés sur les techniques appelées réseaux de neurones.

Faisant partie des méthodes du 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.

2.2.3 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 Concepts et termes fondamentaux

2.3.1 Définitions générales

Réseau de neurones artificiels Un réseau de neurones artificiels ou réseau de neurones (*Neural Network* en anglais) est un « ensemble de neurones formels interconnectés permettant la résolution de problèmes complexes tels que la reconnaissance des formes ou le traitement du langage naturel, grâce à l'ajustement des coefficients de pondération dans une phase d'apprentissage. » D'après Futura².

2.3.2 Concepts récurrents en Apprentissage automatique

Modèle

Un modèle en apprentissage automatique est la représentation du monde construite lors de l'apprentissage afin de répondre au problème à résoudre.

Dans le cadre de l'apprentissage profond, un modèle correspond généralement au réseau de neurone.

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. futura_nn.

Données d'entraînement

Pour entraîner un modèle, il faut lui fournir des données. Généralement, on effectue un traitement préalable des données (*preprocessing* en anglais). Cela peut être retirer les données erronées, en adapter le format, les anonymiser, ou associer le résultat attendu aux données correspondants.

2.3.3 Modèle de la Langue

2.3.4 Principe des réseaux de neurones artificiels

[**statsoft-nn**] Nous allons voir maintenant les principes de base des réseaux de neurones, sans nous attarder sur les principes mathématiques et statistiques sous-jacents.

Le neurone formel

Un neurone forme un objet mathématique qui synthétise des informations puis qui les transforme [**datamine**]. C'est la brique fondamentale du réseau de neurone.

Il est essentiellement caractérisé par :

- **des entrées** : la source des informations à synthétiser ;
- **une fonction de combinaison** : la fonction dirigeant la synthèse des informations ;
- **une fonction d'activation** : la fonction de transformation des informations synthétisées.

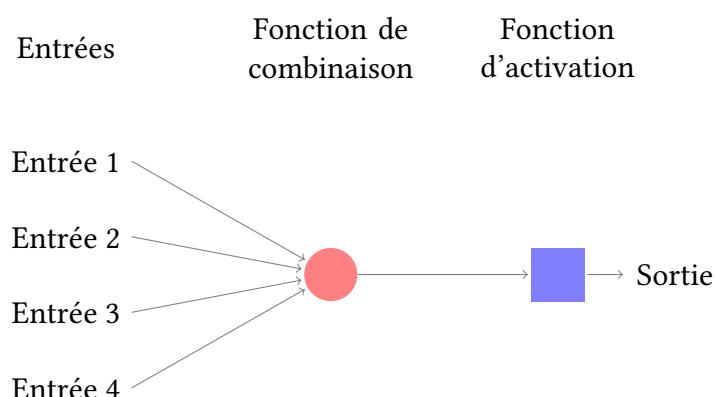


FIGURE 2.1 – Neurone formel basique

Généralement en apprentissage profond :

- les neurones formels manipulent des nombres ;
- la fonction de combinaison est une somme pondérée des entrées, alors **caractérisée par les poids** attribués à chaque entrée ;
- la fonction d'activation est généralement une fonction non linéaire (comme tanh), alors appelée **non-linéarité** ;
- les entrées sont soit les sorties d'autres neurones, soit les entrées du réseau. Cela nous mène à la partie suivante.

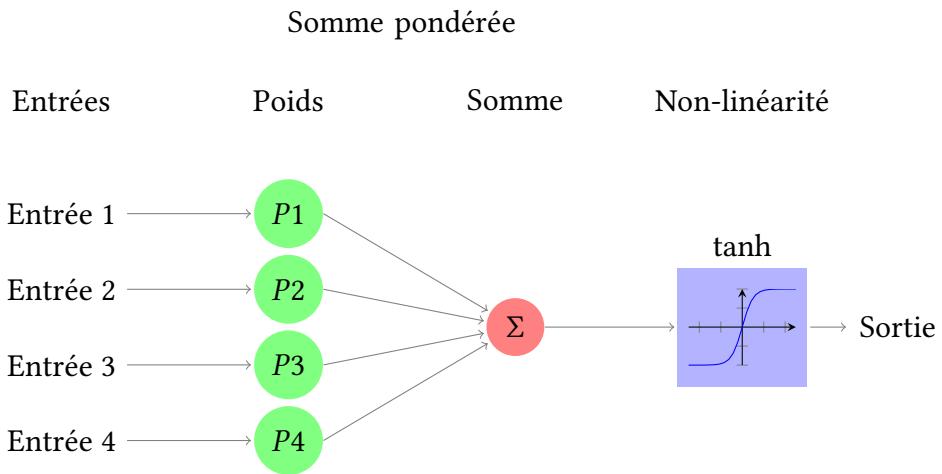


FIGURE 2.2 – Exemple de neurone formel en apprentissage profond, avec une \tanh pour non-linéarité.

Les réseaux et les couches

Les réseaux de neurones sont des réseaux composés de neurones formels, dont les sorties des uns servent d'entrées aux autres.

Il en existe de nombreuses architectures, mais toutes celles dont nous parlerons dans ce rapport sont organisées en couches. En apprentissage profond, on parle de couches cachées pour toutes les couches situées entre la couche d'entrée et celle de sortie du réseau. La Figure 2.3 représente un réseau de neurones artificiels en trois couches.

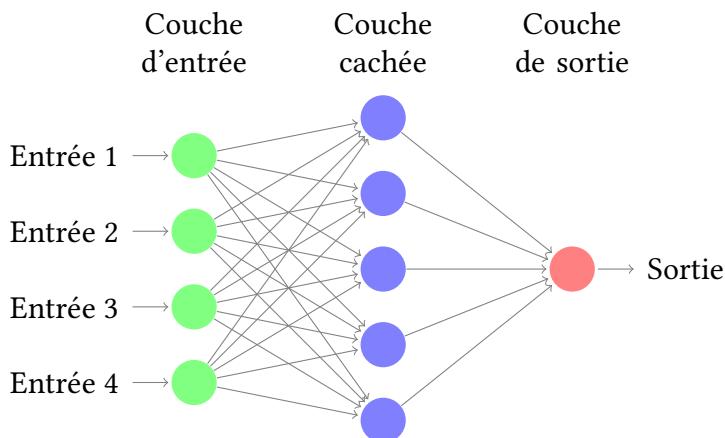


FIGURE 2.3 – Réseau de neurones en 3 couches, de respectivement 4, 5 et 1 neurones. Les couches sont complètement connectées entre elles.

L'entraînement des réseaux de neurones

Une des façons d'entraîner un réseau de neurones artificiels est de lui présenter des données, et de comparer les résultats produits par le réseau de neurones artificiels avec les résultats attendus.

On cherche ensuite à minimiser l'écart entre résultats produits et attendus. La technique de la « rétro-propagation du gradient » permet de connaître l'implication de chacun des paramètres du modèle dans l'écart des résultats, et de les mettre à jour de façon à minimiser l'écart. Cette technique est mise en œuvre informatiquement par la différentiation automatique.

La représentation sous forme matricielle

Il existe une représentation des réseaux de neurones sous forme de matrice.

Cette représentation

Pour expliquer cette représentation, nous allons → Exemple

-> GPU

Les assemblages de modules et la programmation différentielle

??

2.3.5 Les types de réseaux de neurones artificiels

Les Réseaux *feedforward*

Les Réseaux de Neurones Artificiels Récursifs

Un Réseau de Neurones Artificiels Récursifs, plus simplement RNN en anglais) est un réseau de neurones artificiels suivant une architecture dite récurrente.

Ce genre de réseau est utilisé pour travailler avec des séquences d'entrées et/ou de sorties ; il y a transmission d'information entre chaque élément de la séquence.³

Il existe de nombreuses variantes d'architectures récurrentes pour les réseaux de neurones artificiels, en plus de l'architecture de base que nous venons de présenter. En voici deux, dont nous reparlerons plus loin dans le rapport :

GRU

LSTM

Quelques autres types non détaillés

CNN

3. [wiki_rnn](#).

2.4 État de l'art

2.4.1 Projet GMSNN

2.4.2 Projet PAPUD

-> next parts

Première partie

Projet GMSNN

3 Présentation du laboratoire et de l'équipe

3.1 Généralités

C'est dans le laboratoire du Laboratoire Lorrain d'Informatique et ses Applications (LORIA) que le stage s'est déroulé, au sein de l'équipe SYNALP dirigée par M. Christophe Cerisara, notre maître de stage.

3.2 Le LORIA

Le Laboratoire Lorrain d'Informatique et ses Applications (LORIA) est une Unité Mixte de Recherche (UMR 7503), commune à plusieurs établissements : le Centre National de la Recherche Scientifique (CNRS), l'Université de Lorraine (UL) et l'Institut National de Recherche en Informatique et en Automatique (INRIA). Depuis sa création en 1997, le LORIA se concentre sur les sciences informatiques, que ce soit par la recherche fondamentale ou appliquée.

Structure administrative du LORIA

Il est dirigé par quatre instances [**organisation_loria**] :

- **l'équipe de direction** : composée du directeur, de son adjoint, de la responsable administrative, et de l'assistante de direction ; assiste le directeur dans la prise et la mise en œuvre des décisions ;
- **le conseil scientifique** : composé du directeur du laboratoire, des deux directeurs adjoints et des scientifiques responsables des cinq départements du laboratoire composée de membres élus pour 4 ans et de membres nommés ; assiste le directeur dans la prise et la mise en œuvre des décisions ;
- **le conseil de laboratoire** : composé de membres élus pour 4 ans et de membres nommés ; émet des avis et conseille le directeur sur toutes les questions concernant l'UMR ;
- **l'Assemblée des Responsables des Équipes (AREQ)**.

La recherche au sein du LORIA

Le LORIA est l'établissement qui héberge l'équipe SYNALP, parmi de nombreuses autres équipes.

Ce laboratoire regroupe 28 équipes de recherche, structurées en 5 départements en fonction de leur domaine d'étude.

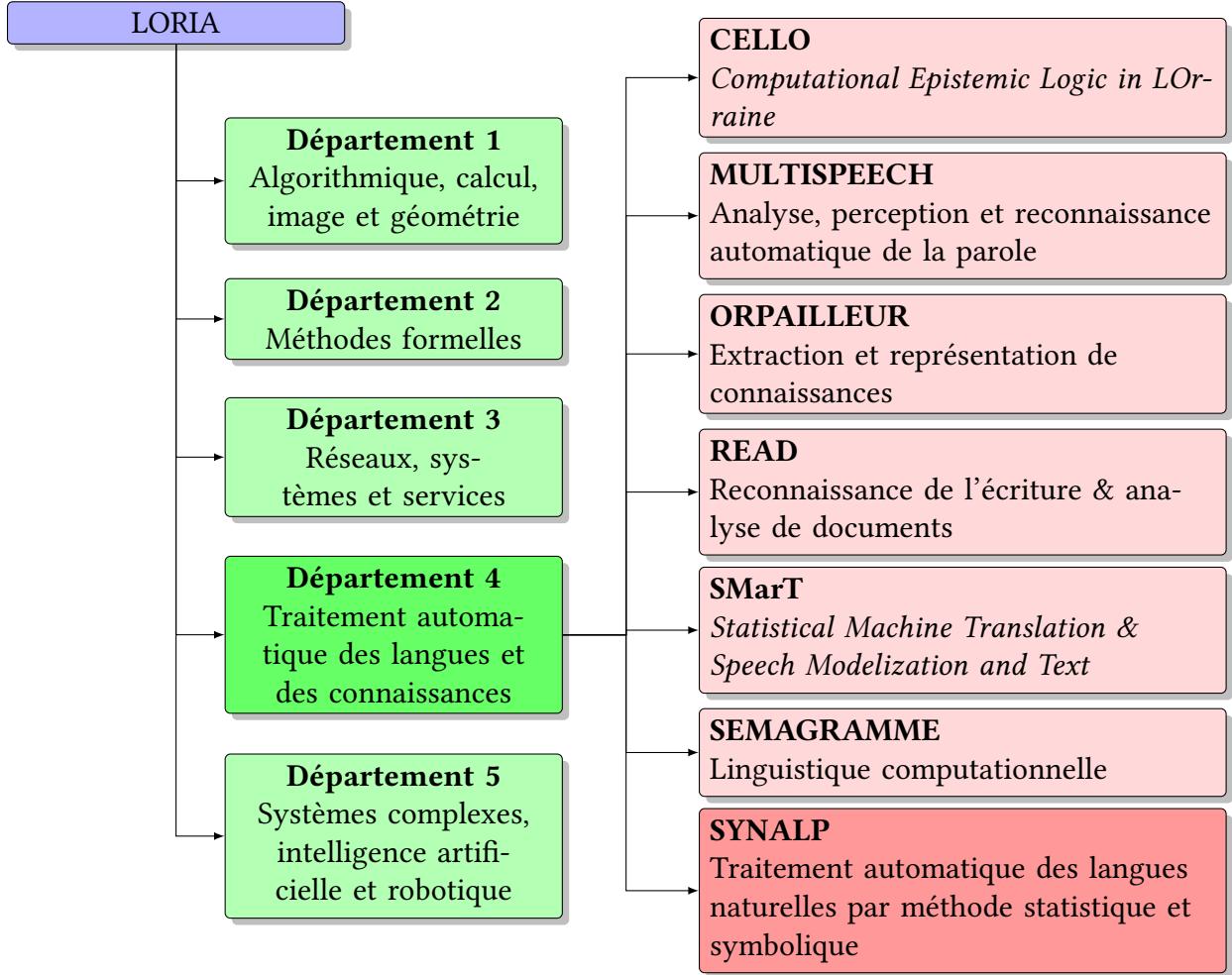


FIGURE 3.1 – Organigramme des départements du LORIA, et des équipes du département 4

La structure générale du LORIA en départements et plus en détail du département 4 est représentée sur l'organigramme de la Figure 3.1. Les thématiques générales de chaque département et des équipes du département 4 y sont présentées brièvement. Un organigramme complet du LORIA est disponible sur le site du laboratoire [[org_loria](#)].

3.3 L'équipe SYNALP

L'équipe SYNALP (*SYmbolic and statistical NATural Language Processing*) est une équipe de recherche affiliée à la fois au CNRS et à l'Université de Lorraine. Elle fait partie, avec 6 autres équipes, du département 4, dédié au traitement automatique des langues (TAL) et des connaissances.

Membres

L'équipe SYNALP est sous la direction de M. Christophe Cerisara, et comporte actuellement 12 membres permanents, une dizaine de doctorants et d'ingénieurs, et approximativement 6 stagiaires à l'heure de l'écriture de ce mémoire.

Thématiques de recherche

La recherche dans SYNALP se concentre sur les approches hybrides, symboliques et statistiques du TAL, ainsi que sur les applications de ces approches.

Ainsi, les principaux sujets de recherche de l'équipe sont les Modèle de la Langue, les grammaires formelles, la sémantique computationnelle, le traitement de la parole, et les outils et ressources utilisés en TAL.

Ce stage s'inscrit en particulier dans la réalisation de Modèle de la Langue, et l'élaboration d'outils et ressources utilisés en TAL. Nous verrons en détail pourquoi dans le chapitre ??.

Pour en savoir plus

Des informations plus détaillées sur le LORIA sont disponible sur le site du laboratoire [[about_loria](#)]. Par ailleurs, la liste complète des membres de l'équipe, ainsi que des informations plus détaillées sont disponible sur le site de SYNALP (en anglais) [[about_synalp](#)].

4 Architecture innovante de réseau de neurones pour l'élaboration de modèle du langage

4.1 Contexte

Nous avons vus dans le chapitre 2 du apprentissage profond et en particulier des RNN

Nous nous plaçons ici dans le contexte de la réalisations de Modèle de la Langue. Les modèles actuels, généralement basés sur les RNN, atteignent de très bonnes performances. Les modèles basés sur les caractères se montrent particulièrement flexibles, car ces modèles "apprennent" les mots, à la place de se reposer sur des dictionnaires très volumineux, qui ont des difficultés à gérer les fautes et les mots nouveaux.

4.1.1 Manque d'utilisation des gros volumes de données

Cependant, ces modèles sont souvent développés et entraînés avec peu de données. Les raisons envisageables sont principalement : le manque de données brutes ou préparées ; et le peu d'amélioration de performance malgré des coûts largement augmentés.

4.1.2 Problèmes de mémoire

Une des raison du manque d'augmentation de performance, typique des RNN, est la limite de rappel d'informations en « mémoire » (dans ses états cachés).

Pour avoir un ordre d'idée, on peut considérer qu'un RNN basique conserve en mémoire des informations datant d'au plus 20 entrées auparavant ; un GRU peut se rappeler d'informations vieilles d'une 100^{aine} d'entrées ; et un LSTM dépasse difficilement les 200 entrées.

Il est donc difficile d'apprendre des dépendances entre des éléments très écartés.

De nombreuses tentatives ont été faite de résoudre ce problème, par exemple en changeant l'architecture du réseau de neurones artificiels (ex : GRU, LSTM), ou en augmentant le réseau avec des mécanismes comme ce que l'on appelle mécanismes attentionnels, ou avec de la mémoire explicite.

4.2 Solution proposée

L'architecture proposée par notre maître de stage vise à la fois à tirer partie des grands volumes de données, et à permettre au modèle d'établir dépendances de haut niveau, voir des connaissances contextuelles externes.

L'architecture et ses caractéristiques sont décrites en détail chapitre 6.

Nous avons nommé cette architecture GMSNN. Ainsi, nous désignerons ce projet par « projet GMSNN » dans le reste du rapport.

4.3 projet GMSNN

La mission qui nous a été confiée est la création d'un apprentissage automatique basé sur l'architecture GMSNN.

L'implémentation devait se réaliser à partir d'une base de code sur laquelle notre maître de stage avait commencé à travailler (plus de détails sont disponibles sous-section 7.2.1).

À cela s'ajoute l'exploration du potentiel de l'architecture, par le biais d'une amélioration du modèle créé à l'aide d'optimisations classiques et de changements de l'architecture.

Enfin, la réintégration des optimisations déjà contenues dans la base de code devait conclure le stage.

4.4 Organisation du travail

Durant ce projet, nous avons travaillé individuellement.

Un fonctionnement en rapport réguliers (disponibles en annexes), complémentés d'une occasionnelle correspondance électronique, a permis de tenir notre maître de stage informé de l'avancement du stage.

À cela s'ajoutent des réunions hebdomadaires avec M. Cerisara, afin de faire le point sur les résultats obtenus et de décider de la marche à suivre.

Le code et les rapports sont stockés sur les serveurs Gitlab de l'Inria, avec le système de gestion de version Git.

4.4.1 Organisation initiale du travail

Nous avons prévu l'organisation temporelle du travail dès la prise de connaissance du sujet définitif du stage (la réalisation de l'architecture proposée).

La première semaine était dédiée à l'acquisition des connaissances nécessaires, à la lecture d'articles et à la prise en main des outils. Ensuite, 3 semaines étaient consacrées à la prise en main de la base de code fournie et à l'implémentation d'un prototype. Les 4 semaines suivantes devaient permettre d'améliorer l'architecture et d'intégrer de nouvelles fonctionnalités. Enfin, les optimisations étaient de l'art contenue dans la base de code devaient être intégrées durant les 4 dernières semaines.

La Figure 4.1 représente cette répartition prévue du travail.

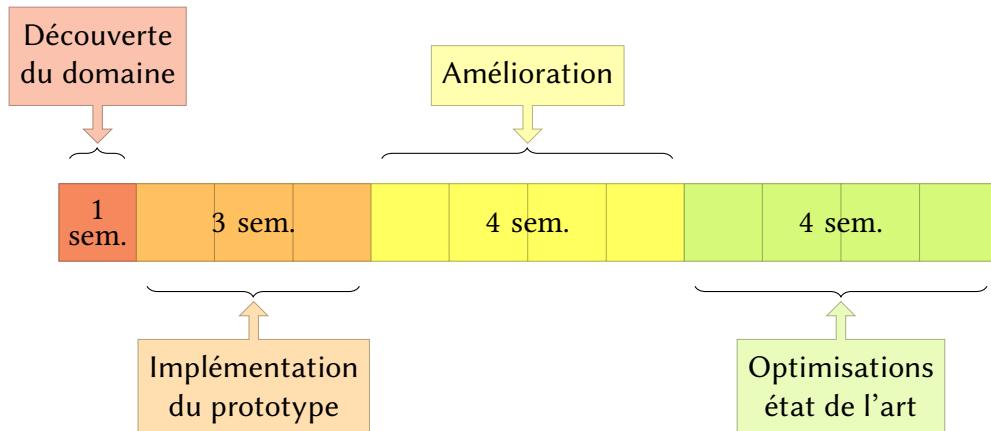


FIGURE 4.1 – Répartition prévue du travail

4.4.2 Déroulement réel du projet

Le projet s'est déroulé comme prévu jusqu'à la fin de la période d'amélioration.

Cependant, comme décrit section 7.7, nous avons décidé d'interrompre ce projet pour nous consacrer au projet PAPUD.

La Figure 4.2 représente la répartition réelle du travail.

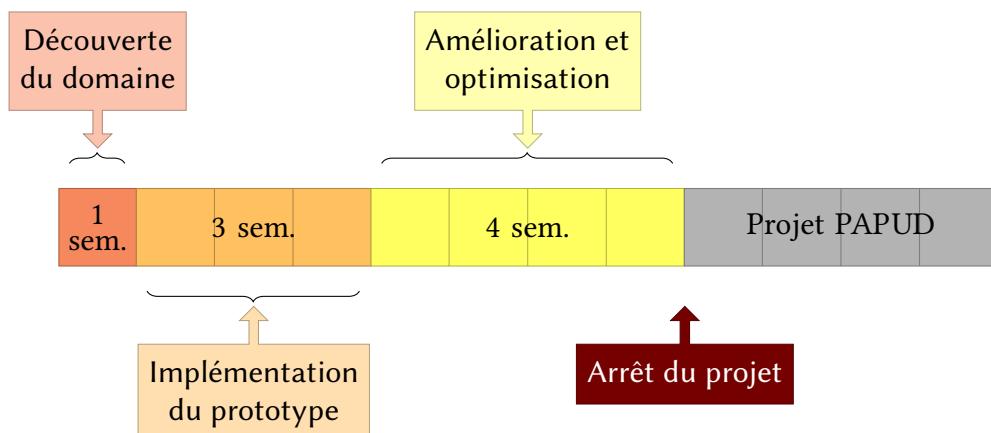


FIGURE 4.2 – Répartition réelle du travail

4.5 Outils

4.5.1 Langage et librairie

Le langage choisi pour l'implémentation est Python, qui largement fourni en outils et librairies d'apprentissage profond.

Parmi ces librairies, notre choix c'est porté sur PyTorch, qui contrairement à d'autres librairies telles que Caffe ou Keras, permet de modular l'architecture du réseau au cours de l'apprentissage. Cette propriété est très importante, étant donné la nature « croissante » de l'architecture proposée pour la première partie du projet.

4.5.2 Grid5000 et les machines distantes

Pendant le déroulement du projet, le modèle à été testé et entraîné sur des machines distantes.

Ces machines font partie du réseau Grid5000. « Grid'5000 est un banc d'essai à grande échelle et polyvalent pour la recherche expérimentale dans tous les domaines de l'informatique, avec un accent sur l'informatique parallèle et distribuée, y compris Cloud, HPC et Big Data. » D'après le site de Grid5000??.

Un des avantage de cet outil est la présence de machines spécialisé dans les calculs GPU, qui sont celles que nous avons utilisées.

5 Données disponibles

Les données d'entraînement utilisées sont tirées du Wikipédia anglais. Ces données sont tirée du fichier « enwik8 »[enwik8] composé d'environ 100 000 000 caractères.

Ces données sont composées de texte balisé structuré en paragraphes. Quelques fragment de XML sont aussi présents, mais ils sont minoritaires dans les données.

Deux versions alternatives de ce corpus ont été utilisées.

1. la première est composée des 10 000 000 premiers caractères de « enwik8 » ; cette version à servi aux entraînements et à la plupart des tests du modèle ;
2. la seconde est composée des 1 000 000 premiers caractères de « enwik8 » ; elle à servi pour le débogage du modèle.

5.1 Extrait des données d'entraînement

Voici un extrait des données brutes avant le découpage en caractères.

```
1 While anarchism is most easily defined by what it is against, anarchists also
  offer positive visions of what they believe to be a truly free society.
  However, ideas about how an anarchist society might work vary considerably
  , especially with respect to economics; there is also disagreement about
  how a free society might be brought about.
2
3 == Origins and predecessors ==
4
5 [[ Peter Kropotkin | Kropotkin ]], and others, argue that before recorded [[
  history ]], human society was organized on anarchist principles.&lt;ref&gt;
  ;[[ Peter Kropotkin | Kropotkin ]], Peter. ''&quot;[[ Mutual Aid: A Factor of
  Evolution]]&quot;'', 1902.&lt;/ref&gt; Most anthropologists follow
  Kropotkin and Engels in believing that hunter-gatherer bands were
  egalitarian and lacked division of labour, accumulated wealth, or decreed
  law, and had equal access to resources.&lt;ref&gt;[[ Friedrich Engels |
  Engels ]], Freidrich. ''&quot;[ http://www.marxists.org/archive/marx/works
  /1884/origin-family/index.htm Origins of the Family, Private Property, and
  the State]&quot;'', 1884.&lt;/ref&gt;
6 [[ Image : WilliamGodwin.jpg | thumb | right | 150px | William Godwin ]]
```

Fragment de code 5.1 – Extrait des premières lignes du fichier enwik8, correspondant à l'article sur l'anarchisme.

5.2 Prétraitement des données

Le prétraitement des données est composé du découpage du document, et du remplacement des caractères

Comme mentionné dans la sous-section 7.6.2, un défaut dans le prétraitement a mené à la disparition des espaces.

En effet, le prétraitement d'origine du corpus utilisait les espaces en tant que séparateurs pour le stockage des données. Par la suite, au moment d'utiliser les données pré-traitées, l'intégralité des espaces étaient supprimés, y compris ceux du texte d'origine.

Malheureusement, ce défaut a été découvert à la fin du projet, et n'a pas pu être corrigé à temps.

6 Description de l'architecture proposée

6.1 Propriétés du modèle

Pour rappel, l'architecture proposée à pour but d'établir un Modèle de la Langue.

Elle est caractérisée par trois propriétés majeures :

- la structure récurrente ;
- l'utilisation de plusieurs échelles ;
- la croissance du modèle.

Nous avons nommé cette architecture GMSNN en considérant ses principales caractéristiques.

6.1.1 Récurrence du modèle

Comme souvent dans la réalisation de Modèle de la Langue, on peut considérer les données sous forme de séquence.

Dans notre cas, le caractère à prédire est dépendant de la suite de tous les caractères précédents.

En apprentissage profond, le type de réseau de neurones artificiels considéré le plus adapté à la manipulation de séquences est le RNN (voir ??).

C'est pour ces raisons que l'architecture à été conçue à partir de RNN.

6.1.2 Passer à l'échelle

Comme décrit sous-section 4.1.2, les RNNs ont un problème inhérent de capacité mémoire, qui limite la distance des dépendances apprises par le modèle.

Afin de compenser ce défaut, l'architecture GMSNN s'appuie sur des couches de plus en plus vastes, appelées « échelles » dans ce rapport. Chacune de ces couches est un RNN, comme montrée dans la ??.

Chaque échelle supplémentaire permet de modéliser des dépendances sur de plus grandes distances.

De plus, chaque échelle tire ses informations de l'échelle précédente. L'exemple suivant explique ce mécanisme de transfert de l'information, également illustré sur la Figure ??.

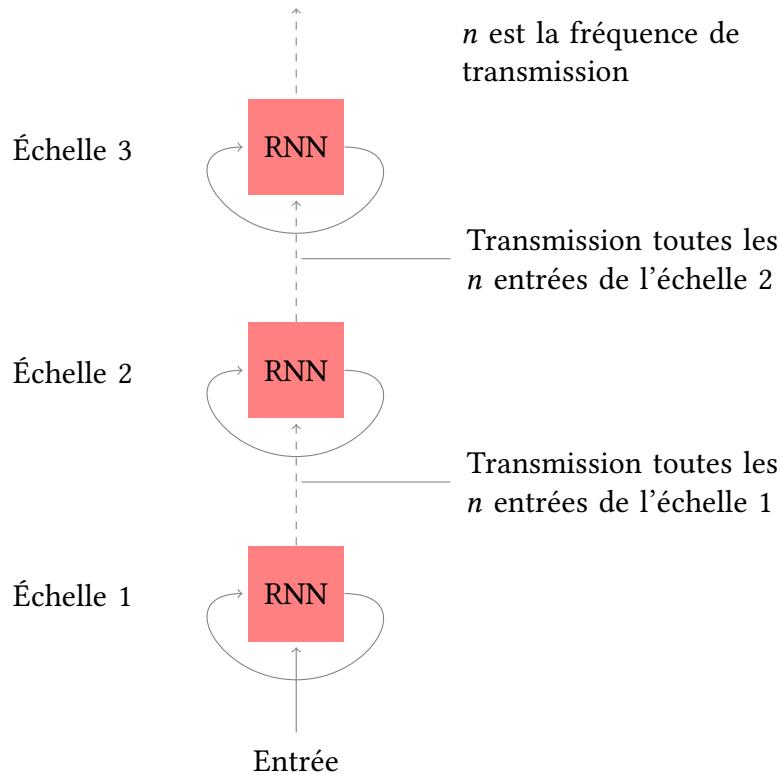


FIGURE 6.1 – Principe de transmission de l’information d’une échelle à la suivante. Ici, la fréquence de transmission est notée n .

Admettons que la capacité de mémoire d’un RNN est de 9 entrées (nombre arbitrairement défini pour l’exemple). Si l’échelle supérieure récupère des informations toutes les 3 entrées de la couche inférieure, sa capacité de mémoire devient $9 * 3 = 27$ entrées. L’échelle encore au dessus aura une capacité mémoire de $9 * 3 * 3 = 81$ entrées, et ainsi de suite. Ici le nombre 3 représente la fréquence de récupération des informations. On parlera par la suite de « fréquence de transmission ».

Plusieurs niveaux d’abstraction de l’information

Une autre caractéristique importante du GMSNN est qu’une échelle prend en entrée les informations **abstraites** par la couche précédente. Ainsi, on peut s’attendre à ce que chaque échelle ajoute un niveau d’abstraction supplémentaire au modèle, comme sur la Figure 6.2.

6.1.3 Adapter le modèle au volume de données et croissance du modèle

Une propriété dérivée de l’architecture est de s’adapter au nombre d’entrées.

En effet, comme décrit dans la sous-section 6.1.2, le nombre d’échelles est dépendant du nombre d’entrées totales.

On peut considérer que tant que aucune entrée ne lui est fournie, une échelle reste dans son état initial, elle n’« existe » pas ; par conséquent, les échelles qui en sont dépendantes n’« existent » pas non plus.

Ainsi, au fur et à mesure de l’entraînement, le modèle croît à la façon d’une pyramide.

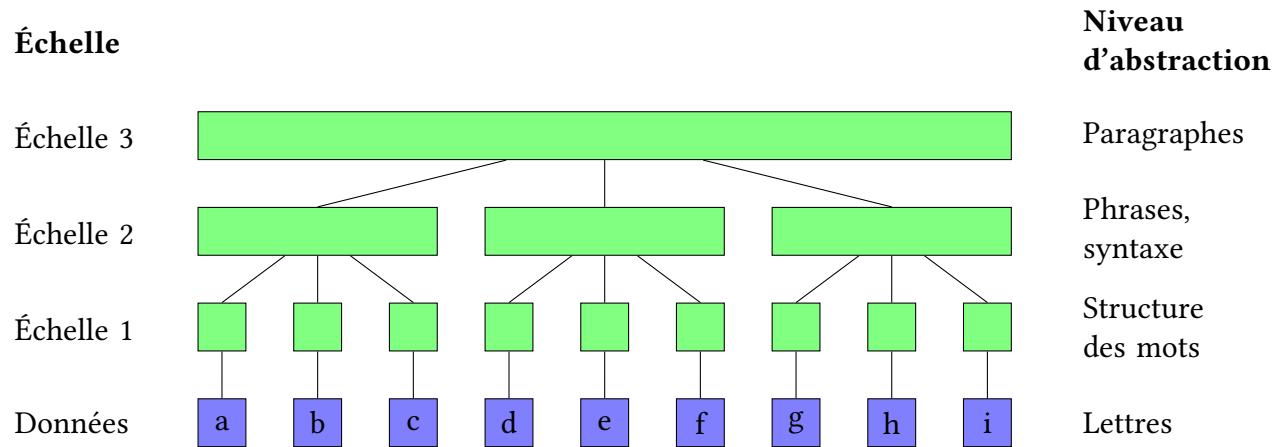


FIGURE 6.2 – Différentes échelles, et les niveaux d'abstraction que l'on pourrait attendre de celles-ci. Chaque bloc vert correspond à une entrée pour l'échelle correspondante. Ici la fréquence de transmission vaut 3. Les blocs bleus en bas du graphique correspondent aux caractères qui sont fournis en entrée au modèle.

Il existe une formule pour déterminer le nombre de couches « existantes » en fonction du nombre d'entrées présentées et de la fréquence de transmission :

$$n = \lfloor \log_f i \rfloor + 1$$

n : nombre de couches, i : nombre d'entrées, f : fréquence de transmission

Croissance potentiellement infinie du modèle

Il est envisageable d'adapter le modèle au nombre d'entrées **durant l'entraînement**, en créant réellement les échelles au fur et à mesure que l'on fournit les données.

Dans ce cas, tant que l'on lui fournit des données, la croissance du modèle est potentiellement infinie.

Comme décrit sous-section 7.5.1, cette propriété a rapidement été abandonnée.

7 Réalisation

7.1 Recherche documentaire

La première partie du projet, qui à durée à peu près une semaine, a été la recherche documentaire et la prise en main des outils. Ce travail a été effectué à partir des documents fournis par notre maître de stage, de la documentation de PyTorch[[60MinBlitzTorch](#), [ByExampleTorch](#), [Classify](#), [doc_pytorch](#)] et de Grid5000, complétés par nos recherches personnelles.

7.2 Étude et ré-implémentation simplifiée du modèle état de l'art

7.2.1 Travail effectué

La deuxième partie du projet a été la prise en main de la base de code fournie. Il s'agit d'une implémentation état de l'art d'un Modèle de la Langue au niveau du caractère, sur laquelle notre maître de stage avait commencé à travailler. Le code d'origine provient du dépôt « awd-lstm-lm »[[awd_source](#)], qui contenait un modèle état de l'art de Modèle de la Langue au niveau du caractère.

Au début du stage, la base de code contenait :

- la version d'origine du dépôt;
- un début de ré-implémentation simplifiée du modèle de la version d'origine ; cette version devait servir de base pour développer le modèle du GMSNN, ainsi que de comparaison pour les performances du nouveau modèle ; elle comportait quelques bogues et ne fonctionnait pas en l'état ;
- un début de travail sur l'architecture du GMSNN.

L'objectif de cette étape était de faire fonctionner la ré-implémentation simplifiée du modèle.

Pour cela, nous avons déchiffré et re-documenté le code, qui contenait des fragments obsolètes et peu documentés. Après le déchiffrage, il a fallu comprendre et réparer les fragments défectueux.

7.2.2 Modèle ré-implémenté simplifiée

Le modèle simplifié produit est composé d'un module encodant les caractères, d'un RNN particulier (un LSTM, voir sous-section 2.3.5), et d'un module produisant une distribution de probabilité sur les caractères connus. La Figure 7.1 représente cette architecture.

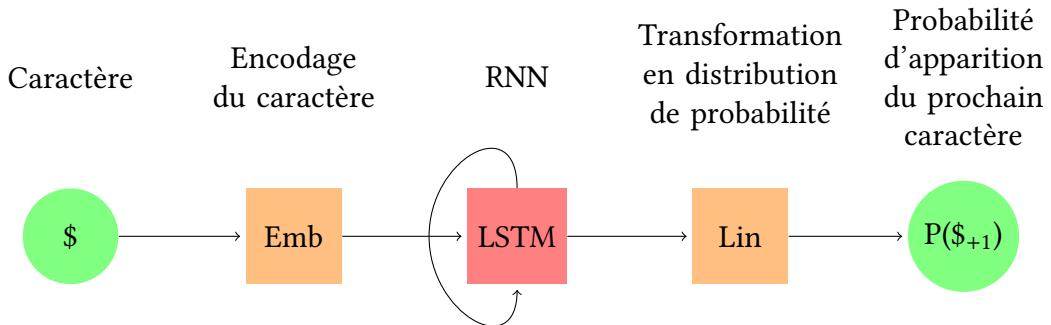


FIGURE 7.1 – Architecture du modèle réimplémenté. Le modèle prend en entrée des caractères, et produit des probabilités sur quel caractère apparaîtra ensuite.

Le module d'encodage des caractères, appelé *embedding layer* en anglais (littéralement « couche d'inclusion »), produit une représentation apprise de chaque caractère sous forme de tenseur (*tensor* en anglais). Ce tenseur est appelé *embedding*. Ce module, entraîné, peut apprendre des propriétés spécifiques à chaque caractère. Par exemple ce module peut apprendre que telle lettre est une consonne et que telle autre est un caractère de ponctuation.

Le RNN traite les caractères sous forme de séquence, et peut ainsi apprendre la structure en mots, la syntaxe et d'autres propriétés du langage.

Le module produisant la distribution de probabilité est un module linéaire. Il transforme les informations produites par le réseau de neurones en probabilité pour chaque caractère d'être le prochain caractère de la séquence.

Voir les annexes C.8, C.9 et C.10 pour les rapports sur le modèle ré-implémenté.

7.2.3 Conclusion

Cette étape, outre la préparation du code, a permis la mise en œuvre et une meilleure compréhension des concepts appris durant l'étape précédente. C'est la première pierre de l'édifice qu'est le GMSNN.

7.3 Implémentation du nouveau modèle

7.3.1 Travail effectué

La troisième partie du projet a été la réalisation d'un prototype de l'architecture GMSNN, basé sur la ré-implémentation simplifiée du modèle état de l'art.

L'architecture du GMSNN est identique à celle du modèle ré-implémenté, mis à part le RNN qui est remplacé par le module GMSNN (voir Figure 7.2). C'est sur ce nouveau module que le reste du travail au cours du projet GMSNN a été effectué.

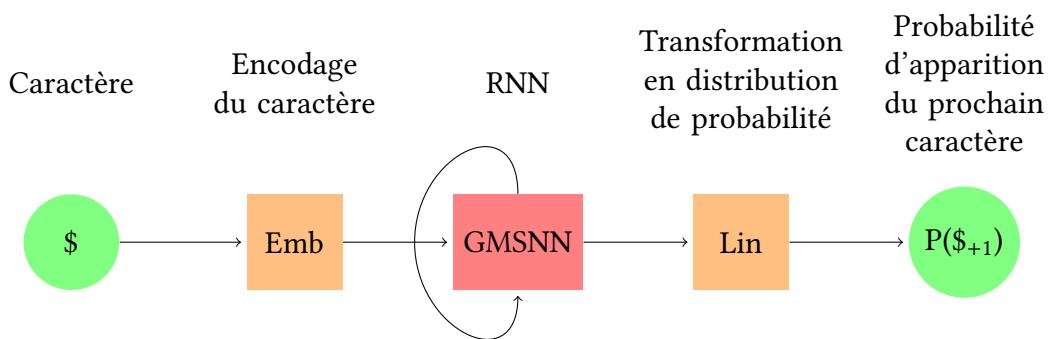


FIGURE 7.2 – Architecture du modèle réimplémenté. Le modèle prend en entrée des caractères, et produit des probabilités sur quel caractère apparaîtra ensuite.

Ce prototype est une implémentation naïve de l'architecture, permettant de mettre en place les mécanismes de base du modèle.

Durant cette étape, nous avons mis en place l'architecture multi-échelle avec deux mécanismes fondamentaux : la transmission de l'information d'une échelle à l'autre, et l'agrégation de l'information de toutes les échelles.

Chaque couche de l'architecture (chaque « échelle ») est un LSTM, comme dans le modèle d'origine.

7.3.2 Transmission d'information

La transmission d'information se fait d'une couche à la couche supérieure. Pour rappel, cette transmission se fait périodiquement, en fonction d'un nombre appelé fréquence de transmission.

Par exemple, pour fréquence de transmission de 3 :

- toutes les 3 entrées de la couche $n - 1$, la couche n reçoit de l'information de la couche $n - 1$;
- toutes les 3 entrées de la couche n (soit toutes les 3^2 entrées de la couche $n - 1$), la couche $n + 1$ reçoit de l'information de la couche n .

Dans un premier temps, il a fallu choisir quelle information transmettre d'une échelle à l'échelle supérieure. En effet, les RNNs produisent à la fois une sortie, et un état caché. L'utilisation de l'*embedding* a été écarté initialement, car elle n'est pas en accord avec l'architecture proposée.

Le choix s'est porté sur l'état caché, qui contient les informations abstraite de la séquence de caractères, contrairement à la sortie qui contient les informations sur le caractère suivant uniquement.

Ensuite, il a fallu déterminer comment regrouper les informations avant de les transmettre au module produisant la distribution de probabilité.

Voir l'annexe C.11 pour le rapport sur le prototype.

7.3.3 Conclusion

Cette partie du projet a permis la mise en œuvre de l'architecture proposée. La dernière étape est d'améliorer les performances du modèle, et de l'optimiser.

7.4 Intégration de systèmes de visualisation

Afin d'évaluer les performances du modèle dans la suite du projet, il a été nécessaire d'établir un système de visualisation des performances.

7.4.1 Utilisation de librairies

Dans un premier temps, diverses librairies permettant de visualiser l'état du réseau de neurones artificiels ont été testées, en particulier VisualDL [VisualDLSite, VisualDLGit].

Malheureusement, aucune de ces librairies ne sont pas en mesure de supporter les architectures les plus complexes (en particulier celles qui impliquent des RNN).

Ainsi, aucune des librairies testées n'a fonctionné avec notre modèle.

7.4.2 Création d'un outil personnalisé

Nous avons donc réalisé un module capable d'enregistrer des données et de réaliser des graphiques. Nous nous sommes basés sur le module « matplotlib » [matplotlib] de Python, et sur une variante française du format CSV [csv].

Ce module a évolué tout au long du projet pour s'adapter à nos besoins.

L'intégralité des graphiques produits dans les divers rapports du projet (disponibles en annexes) a été produit avec ce module.

7.5 Optimisation et amélioration du nouveau modèle

Une fois le prototype fonctionnel, il a fallu améliorer les performances. Par performances, nous entendons principalement le temps nécessaire pour que la qualité prédictive du modèle dépasse un certain seuil.

Pour améliorer ce temps de calcul, il est possible de travailler sur deux dimensions :

- la **quantité de données** traitées en un laps de temps ; c'est une **stratégie quantitative** ;
- la **qualité** de l'apprentissage pour une quantité fixée de données ; c'est une **stratégie qualitative**.

Ainsi, plusieurs options sont possibles :

- optimiser les algorithmes et le modèle pour réduire le temps nécessaire pour traiter les données ;
- améliorer le modèle en réglant les paramètres (comme la fréquence de transmission) ou en implémentant de nouvelles mécaniques ;

Les deux stratégies ont été utilisées. Il faut noter que certaines améliorations qualitatives ont un impact quantitatif négatif.

Principalement, le travail effectué pendant cette partie du projet est un travail de débogage et d'analyse et d'optimisation, avec peu d'implémentation de nouvelles mécaniques dans le modèle.

7.5.1 Agrégation des sorties des couches : d'une stratégie additive à une concaténation

La première optimisation a été de changer la façon de regrouper les informations de toutes les « échelles » avant de les transmettre au module produisant la distribution de probabilité.

Initialement, les sorties de toutes les « échelles » étaient sommées. Cela permettait de maintenir des tenseurs de dimensions uniformes quel que soit le nombre d'« échelle » (Figure 7.3a).

Après discussion avec notre maître de stage, la stratégie d'agrégation a été changé en une concaténation des sorties.

Comme montré dans la Figure 7.3b, la taille du tenseur concaténé change en fonction du nombre d'entrées. La manipulation de tenseurs de taille non fixée est très ardue dans ce cas précis, bien que nous ne développerons pas plus avant les raisons de cette difficulté.

Cela a nécessité l'abandon de la propriété de croissance à l'infini de l'architecture (décrise sous-section 6.1.3), au profit d'un nombre maximal d'échelles défini à l'avance ou déterminée à l'aide d'une formule en fonction des données disponibles (décrise sous-section 6.1.3).

La stratégie par concaténation est plus lente en terme de temps de calcul que la stratégie additive, cependant pour le même temps de calcul elle permet d'obtenir de meilleurs résultats (Figure 7.4).

Voir l'annexe C.12 pour plus de détail sur le choix de la stratégie d'agrégation.

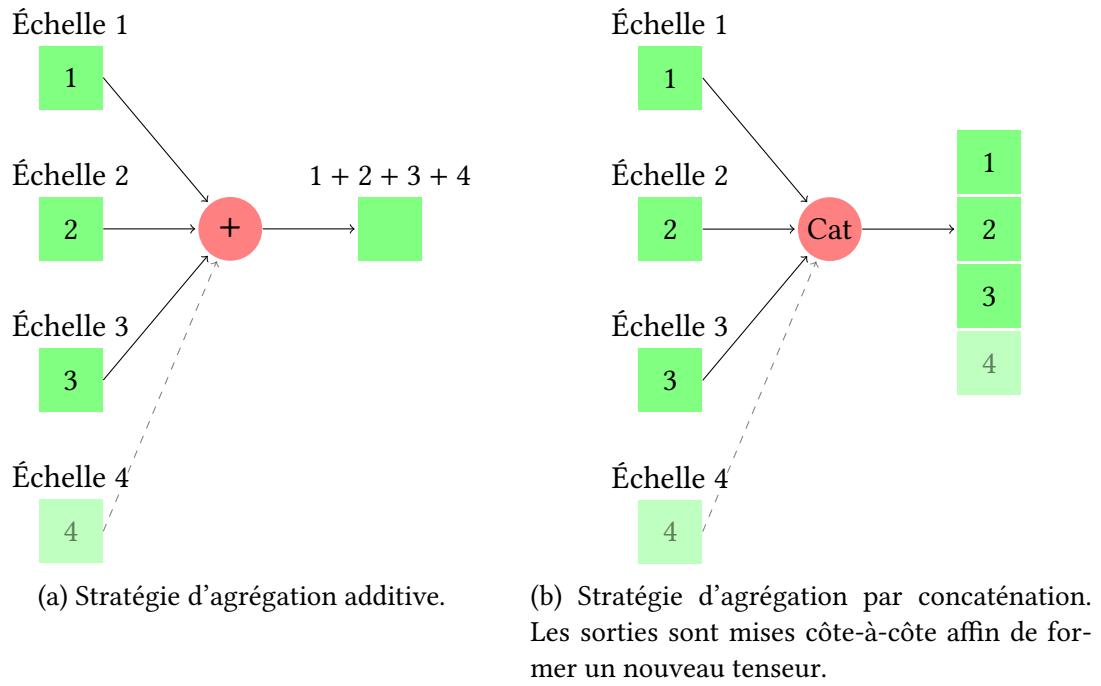


FIGURE 7.3 – Stratégies d'agrégation

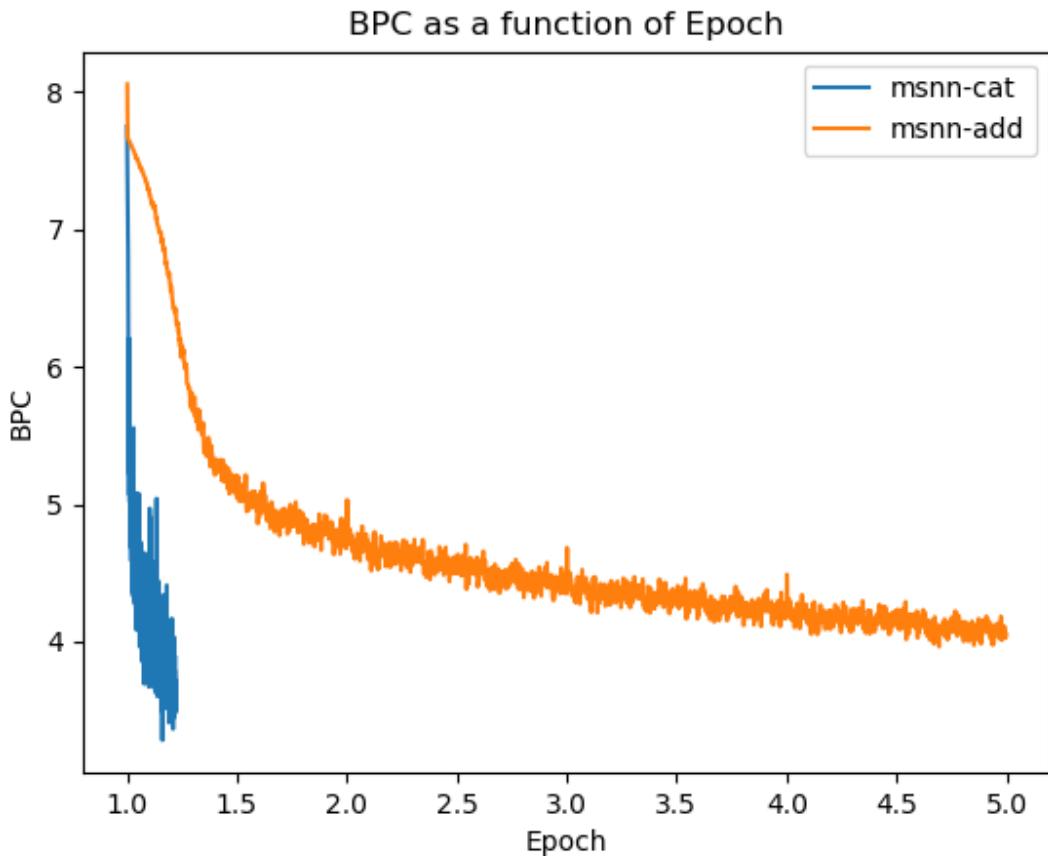


FIGURE 7.4 – Performances comparées des stratégies par concaténation (msnn-cat) et additive (msnn-add). Le temps de calcul alloué est identique. Avec la concaténation on entraîne le modèle sur 1/4 des données, avec l'addition on l'entraîne 5 fois sur l'ensemble des données. Avec la concaténation, on obtient une BPC de 3.5, alors qu'on obtient une BPC de 4 avec l'addition.

7.5.2 Sauvegarde, interruption et reprise de l'entraînement

Une fonctionnalité s'est très vite détachée comme essentielle : la sauvegarde du modèle et la reprise de l'entraînement.

En effet, avec des entraînements très lents et donc longs, il était nécessaire de pouvoir suspendre l'entraînement affin de répartir le temps de calcul sur plusieurs session de plusieurs heures. De plus, les sauvegardes permettent de conserver le modèle une fois entraîné.

Le système implémenté permet d'effectuer cycliquement des sauvegarde du modèle ainsi que de l'état de l'entraînement, permettant ainsi une reprise en l'état de l'entraînement.

Pour la réalisation du système, le principal obstacle à été le malfonctionnement initial des outils fournis par PyTorch. Cela à poussé à la conception d'un système de sauvegarde personnalisé mais malheureusement assez complexe. Cependant, la mise-à-jour majeure de la librairie qui c'est déroulé à point nommé à résolu le problème, et c'est avec les outils de PyTorch que le système de sauvegarde à été implémenté.

Voir l'annexe C.3 pour un rapport contenant plus de détail sur le système de sauvegarde.

7.5.3 Tentatives d'optimisations, fuites mémoires et lenteur de l'entraînement

Les optimisations tentées par la suite ont révélé des fuites mémoires et mis en avant une lenteur excessive de l'apprentissage.

Les optimisations en questions ont été mises en suspens le temps de la résolution de ces deux problèmes. Ces optimisations sont :

- l'utilisation de *batches* simultanés (voir sous-section 7.5.4);
- l'augmentation du nombre de paramètres du modèle (voir sous-section 7.5.5);

Consommation de mémoire et de temps de calcul accrue

Un effet direct de ces optimisations est l'augmentation de la consommation mémoire.

Cette consommation accrue à causé le plantage de plusieurs tests, révélant la présence de fuites mémoires critiques. Un ralentissement progressif de l'entraînement à aussi été découvert pendant l'analyse des plantages. Le plus surprenant a été la corrélation forte découverte entre le temps de calcul et la consommation de mémoire.

Un premier correctif à fourni une amélioration notable mais insuffisante. Il remplaçait le LSTM de chaque couche (voir Figure 7.3.1) par un RNN basique, moins gourmand.

Estimation de la consommation normale du modèle

La première étape, qui est détaillée dans l'annexe C.2, à été d'estimer l'usage normal de la mémoire sans fuite, et d'isoler les paramètres qui ont le plus d'impact sur la consommation mémoire. Cela à confirmé que l'explosion de la consommation n'était pas du à l'architecture en elle même, et qu'il s'agissait bien d'une anomalie.

Résolution des fuites

L'analyse et la résolution des fuites mémoires s'est révélée ardue. Si quelques fuites mineures ont été simple à détecter et réparer, la principale fuite était due à une spécificité non documenté de PyTorch.

En effet, PyTorch utilise la différentiation automatique pour mettre à jour les poids du réseau de neurones artificiels. Pour cela, PyTorch à besoin de connaître la suite d'opérations et l'implication des différents paramètres du modèle et se base sur un « graphe de computation ». C'est la façon dont est gérée ce graphe, couplée aux spécificités de l'architecture GMSNN qui est la cause de la principale fuite mémoire.

L'annexe C.4 contient une des tentatives de résolution du problème.

Conclusion

Le problème de la fuite mémoire à été résolu, et avec lui celui de la lenteur de l'entraînement. On peut déplorer de ne pas avoir analysé plus avant cet étrange lien entre la mémoire et le temps d'entraînement. Cependant, la résolution des fuites mémoires et de la lenteur de l'entraînement était l'objectif principal de cette étape, et l'optimisation du module GMSNN à pu reprendre.

7.5.4 Entrainement par exemples simultanés ; *Batch*

Une fois le problème des fuites mémoires résolu, la première optimisation mise en place est l'utilisation de *batches* parallèles.

batch

Un *batch* (anglais pour lot), est un groupe d'exemples successifs.

Le découpage des données en *batches* permet de répartir l'apprentissage tout au long de l'étude des données. Cela permet d'atteindre de meilleures performances.

Il s'agit d'une optimisation rependue pour l'entraînement de réseaux de neurones. Elle est souvent couplée à un entraînement simultané sur plusieurs *batches*.

Batches parallèles

Un entraînement par *batches* parallèles permet de calculer le résultat de plusieurs exemples simultanément. On calcule ensuite la différence de chaque résultat avec le résultat attendu correspondant. Enfin, on met à jours le modèle en fonction de l'ensemble des exemples. Au final, les calculs des résultat sont parallélisé, et le coût de la mise à jour est mis en commun entre plusieurs exemples.

Cela permet de réduire drastiquement le temps de calcul et d'améliorer la qualité de l'entraînement, au prix d'une plus grande utilisation de la mémoire.

Conflit entre les *batches* parallèles et l'architecture GMSNN

Cependant, le découpage en *batches* pose un problème majeur avec l'architecture GMSNN : elle est basée sur la continuité des exemples fournis, et l'utilisation de *batches* brise la continuité en introduisant un parallélisme.

Une analyse approfondie à permis d'établir une méthode pour contourner et compenser la majorités des aspect du problème.

Voir le rapport de l'annexe C.5 pour les détails de l'analyse des problèmes théoriques de l'utilisation de *batches* avec l'architecture GMSNN.

Voir l'annexe C.2 pour les détails sur l'impact de la taille des *batches* et du nombre de *batches* sur la consommation mémoire.

Voir les annexes C.13, C.16, C.17 et C.18 pour les rapports des tests sur la rotation des *batches*.

Conclusion

Les problèmes liés à l'utilisation de *batches* ont été en majorité résolus ou écartés. Après consultation avec notre maître de stage, nous avons décidé d'utiliser l'entraînement par *batches* malgré les problèmes restants.

7.5.5 Augmentation du nombre de paramètres

Les paramètres, ou poidss, du modèle sont des valeurs qui varient au long de l'entraînement du modèle. On peut dire que ce sont ces valeurs qui « apprennent ».

Le fait d'augmenter ces paramètres augmente la qualité de l'apprentissage, et la précision du modèle appris. Mais cela se fait au coût d'un volume plus important du modèle, et d'une augmentation des calculs nécessaires pour utiliser et entraîner le modèle. Cela se manifeste par un entraînement plus lent et une consommation mémoire plus élevée.

Cependant, grâce aux optimisations mises en place durant la résolution des fuites mémoires (voir sous-section 7.5.3), ces coûts ne sont plus dramatiques.

Il existe plusieurs façon de mettre en place cette optimisation :

- augmenter le nombre de neurones par couches ;
- augmenter le nombre de couches dans le RNNs qui compose chaque « échelle ».

Conclusion

Ces deux pratiques ont été testées, et aucune n'apporte d'amélioration de qualité de l'apprentissage, tout en multipliant le temps d'entraînement. En résumé, ces améliorations apportent des coûts supplémentaires sans aucun bénéfices. Par conséquent, aucune de ces améliorations n'a été conservée.

7.5.6 Entrainement couche par couche

La dernière optimisation mise en place est un nouvel algorithme d'entraînement.

Cet algorithme est une implémentation naïve d'un entraînement couche par couche appliquée à l'architecture GMSNN. Cette algorithme s'apparente aux algorithmes HM [hm].

L'intuition à l'origine de l'algorithme est : il semble que pour apprendre des représentations de haut niveau, le modèle doit en premier lieu apprendre les représentations de bas niveau ; en effet, sans mots, il est difficile de faire des phrases cohérentes.

Cela sous-entend que les échelles les plus proches des données doivent apprendre avant que les échelles supérieures puisse le faire à leur tour. Aussi, il semble inutile d'augmenter la charge de l'algorithme d'entraînement en entraînant des couches qui n'apprennent pas.

Le fonctionnement général de cette algorithme est d'entraîner successivement, une à une, les échelles du modèle en commençant par celle la plus proche des données.

Le fonctionnement détaillé de l'algorithme est disponible dans l'annexe C.7.

Performances

Les performances de l'algorithme sont disponible dans le rapport dans l'annexe C.19.

L'algorithme remplis sa fonction d'alléger la charge calculatoire. En effet, on à une réduction notable du temps nécessaire pour l'entraînement.

De plus, il n'y à aucune variation notable de la qualité de l'entraînement.

Justement, comme seule une échelle apprend, on pourrait s'attendre à une baisse des performances.

Comme le modèle muni d'une seule échelle apprend aussi bien que le modèle multi-échelles, cela remet en question l'utilité de l'architecture GMSNN et de ses échelles multiples.

Conclusion

7.5.7 Conclusion

Cette partie du projet GMSNN à permis d'améliorer notamment les performances du modèle, tout en réduisant drastiquement le coût d'entraînement.

De plus, l'algorithme présenté sous-section 7.5.6 à démontré une faiblesse majeure de l'architecture GMSNN.

On peut aussi noter la mise-à-jour majeure de PyTorch, qui en plus de résoudre certains dysfonctionnements à nécessité le remaniement d'une partie de la base de code.

7.6 Production des exemples et découverte du problème d'encodage

Une fois le modèle fonctionnel, une partie importante de la compréhension et de l'évaluation du modèle est la production d'exemple.

Nous avons retardé cette étape principalement à cause des problèmes de mémoire.

Le principe de cette étape est d'utiliser notre Modèle de la Langue pour produire du langage, afin d'avoir une idée plus concrète qu'un score de la performance du modèle.

7.6.1 Exemples

Voici quelques exemples produits par le modèle. La version du modèle choisie est celle avec le meilleur score, parmi celle enregistrées.

Age du modèle : 465 époques.

Score BPC du modèle : 1.839.

Pour comparaison, le modèle état de l'art avait un score BPC de 1.255 en 50 époques.

Pour rappel, les données sont issues d'une version filtrée de Wikipédia en anglais.

```
1 YeoMMDF| Ph#elementat [[ Damous ]]  
    thatureoftenusevoirbeexpounderstatesandanumberofhisworkformembersthan  
    novelwasmethecebylementorfromthelastPreenancenoldWarInstartedbythe  
    Philosophy ' ' theTayita (   
    amsmethouspeopleamingshelebelobesinthesatietheuniversalistscientis  
    educationof [[ Lakingforts ]].
```

Fragment de code 7.1 – Exemple 1 : une suite de caractères à priori incompréhensibles.

```
1 +EDrFuergCases , areinlesssuchasthesthealterplains .  
2 * In [[ Stefapes ]]  
3 * [[ AcademyAwards===  
4  
5 ANASA) asLASCIIRunder , andas .  
    MatthebusipenclearsandpresidenthaveaqueluelsifthesearchfromAwarerLievol  
    ofany30020 .  
6  
7 It ' '[[ Anim ]]  
8 * [[ UnitedStates | raphicsiteDirection ]]  
9  
10 Theplant-gainheditsextructuretoaseethewarinsteast" ; Oneofthe [   
    ectlywouldnotbytheIntegrationscapianland ]( ora ' '[[ schology ]])  
11 | published :
```

Fragment de code 7.2 – Exemple 2 : des termes balisé comme dans le corpus d'origine, les crochets ouverts sont refermés.

```
1 60447-toNewHarry }}  
2 Thenmainst . Rand ' s intereststhe " ; toinpassingtheEarth ' ' s ( ' '[[ par ]]).  
3
```

```

4 : ' ' ' maimals , anackreloquedoutwidthofgrawwithluteframedapprovingtoundernverby [ [
hebesination ]] of< / smalkan ,
instablishedacondorttodevelopedframesbeforestatedwinkingaroundsinrational
hicarefartoredonaftercanbeagainstthatgroupswouldnear ,
notwhatwasthatisstillastructionCenter , toDagnythat
5
6 On[[éAft ele of Airej ]]
7 [[cy : Alaska ]]
8 [[no : Arni – Anchorage ]]

```

Fragment de code 7.3 – Exemple 3 : des termes balisé comme dans l'exemple 2, et une autre suite de caractères.

7.6.2 Manque d'espaces

Parmi ces exemples, on remarque immédiatement le manque d'espace.

La source de ce phénomène n'est autre qu'un problème dans le corpus source. La version pré-traitée de ce corpus ne contenait aucun espace, donc le modèle à appris une langue dans laquelle l'espace n'existe pas.

C'est un problème majeur, qui à probablement eu un impact élevé sur les performances du modèle. En effet, en anglais comme dans beaucoup de langues occidentales, l'espace est un élément fondamental dans la structuration du langage écrit. Le modèle à ainsi apprit une langue moins structurée, donc plus difficile à apprendre, que l'anglais.

7.6.3 Quelques éléments qui ressortent

Cependant, si on regarde plus en détail les exemples produits, des structures apparaissent.

Si on prend *[[UnitedStates | raphicsiteDirection]] de l'exemple 2 (Code 7.2, ligne 8) ou [[cy:Alaska]] et [[no:Arni–Anchorage]] de l'exemple 2 (Code 7.2, lignes 7 et 8), on a des doubles crochets, qui sont correctement ouverts puis fermés. On remarque aussi la présence de séparateurs (: et |).

Ces structures sont similaires aux annotations présentes dans le code Wikipédia.

Enfin, malgré l'absence d'espaces, on discerne de nombreux mots :

- la suite de mots statesandanumberofhisworkformembersothannovelwasmethe qui ne contient en fait que des mots bien formés en anglais : states and a number of his work for member so than novel was me the (Code 7.1, ligne 1);
- de même pour la suite de mots oldWarInstartedbythePhilosophy : old War In started by the Philosophy (Code 7.1, ligne 1);
- on trouve aussi des noms propres comme le nom de pays : UnitedStates (Code 7.2, ligne 8) et Alaska (Code 7.3, ligne 7).

7.7 Analyse des résultats et arrêt du projet

7.7.1 Analyse des résultats

Avec notre maître de stage, nous avons étudié attentivement les résultats des dernières optimisations sur les performances du modèle (voir annexe C.19). Le résultat le plus dérangeant étant l'absence d'apprentissage des couches supérieures.

À partir des connaissances de la littérature possédées par notre maître de stage et de ce résultat, nous avons conclu que 90% de l'information nécessaire est apprise par la première échelle du modèle. Les autres échelles ne font qu'améliorer ce résultat, et sont peu utiles tant que la première échelle n'est pas complètement entraînée.

À cela s'ajoute la corruption de l'entraînement par la disparition des espaces, qui nécessiterait non seulement de remanier une partie du code tenue pour acquise, mais aussi de refaire la plupart des tests.

7.7.2 Conclusions de l'analyse

La conclusion à laquelle nous sommes arrivés est qu'il aurait fallu recommencer le développement avec un système de gestion des données maîtrisé.

La première étape aurait été de développer un modèle simple, avec une seule échelle, et de le pousser au maximum de ses capacités. Seulement à ce moment là nous aurions pu l'augmenter d'autres échelles.

Il aurait aussi été intéressant de revoir l'architecture pour utiliser un modèle sans récurrences.

Cette conclusion impliquait de recommencer le projet, ou à défaut de le remanier en grande partie.

7.7.3 Arrêt du projet et début du projet suivant

Au moment de cette analyse, notre maître de stage revenait d'une réunion décisive sur le projet PAPUD.

Celle-ci avait permis de définir les objectifs du projet ITEA3-PAPUD, cas d'utilisation BULL (voir chapitre 10).

La nécessité de recommencer le projet GMSNN, couplée à l'opportunité de mettre les conclusions du projet et les compétences acquises en pratique dans un projet à grande échelle, nous ont mené à interrompre projet GMSNN pour consacrer la fin du stage au projet PAPUD.

8 Conclusions sur le projet GMSNN

8.1 Retour sur le travail effectué

Ce projet nous à permis d'implémenter une architecture innovante de réseau de neurones artificiels, à partir du squelette d'un modèle état de l'art. Nous avons pus élaborer un prototype suivant les concepts clés de l'architecture proposée, avant de l'améliorer et de l'optimiser.

Pour cela nous avons étudié un domaine technique dans lequel nous avions peu de connaissances ; nous avons manipulé une librairie qui nous était inconnue ; nous avons géré des tests durant de plusieurs heures à plusieurs jours sur des machines distantes ; nous avons, enfin, affronté un des obstacles les plus importants dans le développement de réseau de neurones artificiels, le problème de l'optimisation.

Bien que le l'architecture GMSNN n'ai pas atteint les performances espérées, le modèle produit est robuste, rapide, et peu volumineux. De plus, l'algorithme présenté sous-section 7.5.6 à démontré une faiblesse majeure de l'architecture GMSNN. Enfin, les problèmes rencontrés dans ce projet ont permit de tirer des conclusions très utiles pour de prochains projets :

- les RNN sont très lents à entraîner ;
- la maîtrise du pré-traitement est fondamentale pour obtenir des bons résultats ;
- pour utiliser une architecture multi-échelle comme celle proposée, il vaut mieux entraîner un modèle simple en premier lieu.

C'est d'un commun accord avec notre maître de stage que nous avons décidé de basculer sur le projet PAPUD.

En conclusion, le projet à abouti sur le rejet de l'architecture proposée. Néanmoins, ce résultat à permit de cerner les principaux écueils de la réalisation d'un Modèle de la Langue multi-échelle, et à ainsi permit un meilleur déroulement du projet suivant.

8.2 Apport personnel du projet

La réalisation de ce projet nous à permis d'approfondir largement nos connaissances en apprentissage profond, et de nous habituer aux problématique de la création et de l'utilisation de réseaux de neurones.

Deuxième partie

Projet PAPUD

9 Présentation du projet ITEA3-PAPUD, cas d'utilisation BULL

La seconde partie du stage s'intègre dans le projet PAPUDprojet ITEA (*Information Technology for European Advancement*)-PAPUD (*Profiling and Analysis Platform Using Deep Learning*), en particulier dans le cas d'utilisation BULL. Nous verrons en détail les objectifs du projet dans la ?? (page ??).

Le projet ITEA3-PAPUD, cas d'utilisation BULL est un projet de l'initiative ITEA3 du réseau EUREKA.

9.1 Collaborateurs

Les personnes avec lesquelles nous avons collaboré durant ce projet sont notre maître de stage M. Christophe Cerisara, ainsi que deux autres chercheurs de l'équipe SYNALP, Mme. Nadia Bel-lalem et M. Samuel Cruz-Lara. Une quatrième chercheuse, Mme. Christine Fay-Varnier, nous à rejoint vers la fin du stage.

9.2 EUREKA et ITEA3

« EUREKA est une initiative européenne, intergouvernementale, destinée à renforcer la compétitivité de l'industrie européenne. » D'après Wikipedia [[wiki_eureka](#)].

ITEA3 est la troisième itération d'un programme du réseau EUREKA nommé ITEA (*Information Technology for European Advancement*). ITEA est un programme de recherche, développement et innovation basé sur un partenariat public / privé, et fonctionnant par appels de projet. Ces appels à projets se concentrent sur des problématiques des technologies de l'information et de la communication, et ce dans une perspective industrielle.

ITEA3 implique plus de 40 pays, ainsi que de nombreuses entreprises.

9.3 Projet PAPUD et cas d'utilisation BULL

C'est lors de la troisième vague d'appels à projets d'ITEA3 que le projet PAPUD à été accepté.

L'objectif du projet PAPUD (*Profiling and Analysis Platform Using Deep Learning*) est l'élaboration d'une série d'outils basés sur les techniques de l'apprentissage profond. La plateforme ainsi produite a pour objectif l'analyse des volumes de données devenus trop grands pour être gérés de façon traditionnelle. Ainsi, le projet PAPUD s'inscrit dans la dynamique d'ITEA3.

Nom complet du projet	16037 PAPUD
Période de réalisation	Janvier 2018 - Décembre 2020 (3 ans)
Appel à projet	ITEA 3 Call 3
Partenaires	16
Coûts estimés	10 927 000 €
Volume de travail estimé (en personne.année)	151,88
Pays participants	Belgique, Espagne, France, Roumanie, Turquie

TABLE 9.1 – Informations générales sur le projet PAPUD, d'après le site de ITEA3 [[about_papud](#)]

9.4 BULL

Présentation de l'entreprise

BULL est une entreprise française spécialisée dans la sécurité informatique et la gestion des gros volumes de données informatiques.

L'entreprise a été rachetée en 2014 par le groupe ATOS.

Secteurs d'activité

D'après le site d'ATOS [[bull_produits](#)], les activités principales de la filiale BULL sont :

- le matériel informatique et logiciel professionnel de haute sécurité ;
- le matériel informatique et logiciel pour l'Armée et la Défense, y compris du matériel de navigation maritime et terrestre ;
- les serveurs de calcul et de stockage, les *data-centers* (infrastructures spécialisées regroupant de nombreux serveurs) et les solutions nuagiques (*cloud*) ;
- les solutions de calcul haute performance (les « supercalculateurs ») ;
- les systèmes intégrés, à savoir du matériel informatique spécifique intégré à un produit, comme par exemple l'ordinateur de bord intégré dans une voiture.

Globalement, BULL concentre ses activités sur le matériel informatique et les logiciels de pointe en matière de sécurité et de fiabilité. Les gammes de produits BULL s'adressent principalement à des grosses entreprises et aux états.

Pour en savoir plus

Des informations plus détaillées sur le projet PAPUD sont disponible sur la page web du projet [**\[about_papud\]**](#).

10 Projet PAPUD

10.1 Contexte

Nous avons vu que parmi les secteurs d'activité de BULL, les serveurs et autres systèmes de traitent de gros volumes de données sont très présents.

Ces outils tombent rarement en panne, mais quand ils le font cela occasionne des pertes très importantes pour l'entreprise.

Il serait donc très intéressant de mettre au point un système de prédition des panes, afin de pouvoir les éviter.

Les données disponibles pour remplir cette tache sont des fichiers de journaux systèmes (décris en détail dans le chapitre 11) de très grande taille. Ils contiennent de nombreuses informations sur les évènements se déroulant dans les outils.

10.2 Solution

D'après les documents de travail officiels (en particulier le fichier README.md du dépôt de code officiel du projet).

Le cas d'utilisation BULL du projet PAPUD est dédié à répondre à cette problématique, en fournissant un système détectant les anomalies (signes de pannes) dans les journaux systèmes.

Pour cela, il a été décidé de modéliser le comportement normal (sans panne) de ces journaux.

Ces journaux sont composés de lignes de textes en anglais. Il est donc possible de produire un Modèle de la Langue capable de prédire la prochaine ligne.

Le plan général des opérations est divisé en 2 parties :

1. on suppose que la structure en dépendances entre les lignes est simplissime : une ligne dépend uniquement de la ligne précédente ; on cherche donc à établir un Modèle de la Langue capable de modéliser au mieux cette dépendance ;
2. une fois le modèle simple établi, on abandonne le postulat précédent, et on cherche à établir à partir du modèle créé un modèle capable de modéliser des dépendances à la fois plus complexes et sur plus d'une ligne au par avant.

Pour ce qui est du modèle simple, il a été décidé de ne pas utiliser de RNN, bien trop lent pour la quantité de données à traiter. À la place, un réseau de neurones artificiels basique, intégralement connecté (comme celui présenté dans la Figure 2.3), sera utilisé.

On peut noter que les conclusions du projet GMSNN ont été appliquées, autant pour le déroulement du projet que pour le type de modèle à utiliser.

10.3 projet PAPUD

La tâche qui nous a été confiée est la réalisation du modèle simple.

Plus exactement, étant donné qu'il était évident que la durée restante du stage serait insuffisante pour réaliser et pousser au maximum le modèle simple, nos objectifs étaient la réalisation d'un prototype du modèle, et de mettre en place les outils nécessaires à l'entraînement. Ceux-ci sont principalement les outils de gestion et de prétraitement des données, les outils d'évaluation des performances du modèle, et l'algorithme d'entraînement du modèle.

La description des caractéristiques du modèle est disponible dans le chapitre 12.

Par la suite, nous désignerons ce projet par « projet PAPUD ».

10.4 Organisation du travail

Contrairement au projet GMSNN, d'autres collaborateurs participaient à ce projet (voir section 9.1). Étant, avec notre maître de stage, les seuls parmi les collaborateurs habitués à manipuler des réseaux de neurones, nous avons travaillé individuellement durant ce projet.

Le projet GMSNN s'étant bien déroulé, une organisation similaire a été mise en place afin de tenir ces autres membres du projet informés de l'avancement et des conclusions de notre travail. C'est-à-dire que des rapports fréquents et des réunions hebdomadaires durant lesquelles nous présentions notre progression ont été mis en place.

Les rapports de ce projet sont également disponibles en annexe. Le rapport d'une des réunions est disponible à l'annexe D.6.

Le code et les rapports sont stockés sur les serveurs Gitlab de l'Inria, avec le système de gestion de version Git. Les collaborateurs du projet ont accès à l'ensemble de ces données, qui ont été mises à jour tout au long du projet.

10.4.1 Organisation initiale du travail

Ce projet PAPUD s'est déroulé sur la base de cycles de développement. C'est durant les réunions hebdomadaires que les prochains objectifs étaient décidés.

En effet, contrairement au projet GMSNN pour lequel il a été simple de définir des périodes réservées aux grandes étapes du projet, ce projet PAPUD s'est déroulé dans un temps très restreint.

Cependant, il a été possible de définir les priorités suivantes :

1. la réalisation d'un prototype fonctionnel;
2. la mise en place d'un algorithme d'entraînement basique;
3. la mise en place de moyens d'évaluer le modèle et l'obtention de premières performances ;
4. le prétraitement des données et la préparation de la gestion des très gros volumes de données à venir ;
5. le temps restant est dédié à l'amélioration des performances.

Une extension de la durée du stage de 1 semaine a été décidée, de façon à augmenter le temps dédié au travail sur le projet. Cela c'est fait en considérant les disponibilités de notre maître de stage, des autres collaborateurs ainsi que les nôtres.

La durée totale du projet a donc été de 5 semaines.

10.4.2 Déroulement réel du projet

Tous les objectifs nécessaires ont été remplis, et deux améliorations notables des performances ont été mises en place.

La répartition réelle du travail du projet est représentée dans la Figure 10.1.

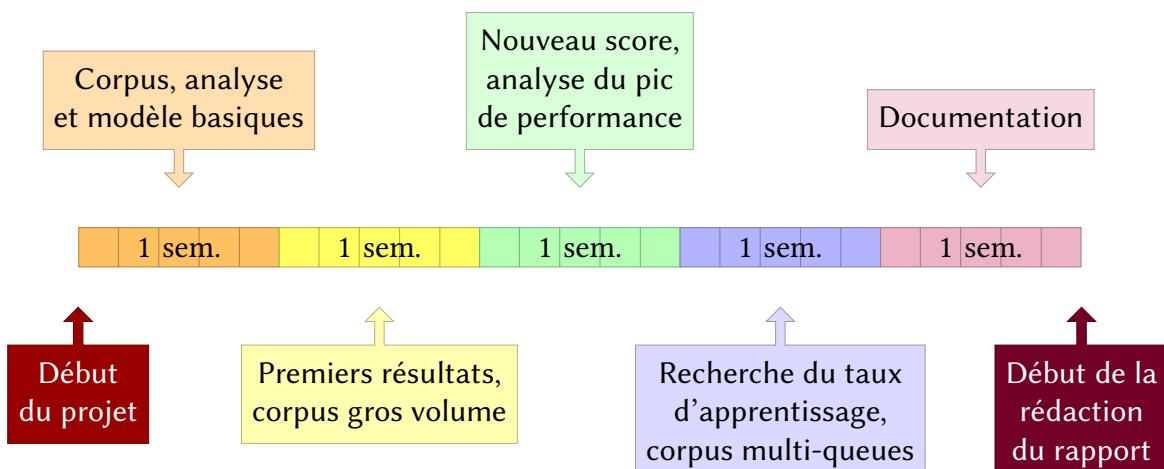


FIGURE 10.1 – Répartition du travail

10.5 Outils

Les outils choisis sont Python et PyTorch, pour trois principales raisons :

- la première partie du projet s'est déroulée avec ces outils, nous étions donc habitués à leur utilisation et à leurs subtilités ;
- la base de code accumulée jusqu'à présent reposait sur ces outils, et pouvait être réutilisée sans trop d'efforts ;
- il était possible de basculer plus tard vers une autre librairie, PyTorch possédant une fonctionnalité permettant la conversion d'un modèle vers les autres principales librairies disponibles pour Python et quelques autres langages.

11 Données disponibles

Les données disponibles sont des fichiers de journaux systèmes.

Ce sont des lignes de texte en anglais extrêmement structuré, qui donnent des informations sur les programmes en cours d'exécution sur l'outil, et les évènements qui se déroulent. Par exemple, des messages d'information, comme « le programme A a tenté de faire B », ou « l'utilisateur C a changé son mot de passe ».

Ces messages sont précédées d'informations comme le moment d'écriture du message et le programme d'où il provient.

De plus, ces journaux systèmes contiennent un nombre très grand de message, et la quantité de données disponible pour le projet dépasse les 400 GiB de texte brut. En comparaison, les données utilisées pour le projet précédent pesaient moins de 10 MiB, soit 40 000 fois moins.

Un échantillon de 9 MiB des données disponibles a été utilisé pendant le projet. Cela correspond à 70 131 lignes de journaux système.

11.1 Extrait des données d'entraînement

Les données sont confidentielles utilisées sont confidentielles. Ainsi, l'extrait présenté ici a été largement modifié. Entre autre, la date et le *timestamp* (nombre correspondant à la date) ont été remplacées par la date de début du stage et le *timestamp* correspondant, et l'utilisateur a été renommé « OOOOOO ».

```
1 1524463200 2018 Apr 23 08:00:00 OOOOOO authpriv info access granted for
   user root (uid=0)
```

Fragment de code 11.1 – Exemple d'une ligne extraite des journaux systèmes. On voit à gauche : le *timestamp*, la date, l'heure, l'utilisateur (OOOOOO), le processus (authpriv), le type de message (info), et le message.

11.2 Prétraitement des données

Le prétraitement des données est composé de :

- la suppression du *timestamp*, de la date, de l'heure, et de l'utilisateur ;
- le découpage du texte restant à une certaine longueur ;
- le remplissage des lignes n'atteignant pas cette longueur par un caractère spécifique nommé caractère de remplissage ;

- la transformation de tous les caractères en nombres, à l'aide d'un dictionnaire ; les caractères apparaissant peu fréquemment ou n'apparaissant pas dans le dictionnaire sont tous remplacé par un caractère spécial nommé « caractère inconnu ».

11.2.1 Traitement des codes hexadécimaux

Par la suite, le remplacement de tout code hexadécimal comme 0x005f par un caractère spécifique du dictionnaire a été ajouté (voir section 13.6). De cette façon, 0x005f, 0xffff , 0x0007 sont remplacé par le caractère <hexX4>, tandis que 0000005f et 89abcdef sont remplacé par <hex8>. Cela permet de s'abstraire de la valeur du code sans perdre l'information liée à sa présence.

12 Modèle à réaliser

Le modèle à réaliser est conçu pour être le plus rapide possible.

Le modèle est un Modèle de la Langue au niveau du caractère, qui prend une ligne et qui prédit la ligne suivante.

Ainsi, l'entrée du modèle est une ligne de taille fixe de nombres correspondant à des caractères.

La sortie est un tableau à 2 dimensions, qui contient pour chaque caractère prédit une distribution de probabilité sur les caractères connus (répertoriés dans le dictionnaire décrit section 11.2).

12.1 Utilisation d'un réseau de neurones artificiels basique

Comme présenté dans le chapitre 10, un réseau de neurones artificiels basique, intégralement connecté (comme celui présenté dans la Figure 2.3), a été choisi pour le modèle.

Étant le réseau de neurones artificiels le plus simple, il est extrêmement rapide à entraîner.

12.2 Réduction de la taille de l'entrée

Comme écrit plus haut, l'entrée du modèle est une ligne de caractères. Pour chacun d'entre eux, comme dans le modèle GMSNN, est transformé en un tenseur par un module d' *embedding*.

Grâce à une opération appelée « *max-pooling* », qui correspond à prendre pour chaque case du tenseur final le

12.3 Lien avec l'état de l'art

Il a été récemment montré que les Modèle de la Langue basés sur une séquence « *embedding - max-pooling* - réseau de neurones artificiels intégralement connecté » ont des performances

13 Réalisation

13.1 Définition du modèle

La toute première étape du projet à été de définir l'architecture à utiliser (décrite chapitre 12). C'est sous la forme de réunions que cette étape s'est déroulée.

Tout d'abord, nous avons élaboré le modèle avec notre maître de stage. Par la suite, nous avons présenté nos conclusions aux autres membres du projet, qui les ont approuvées.

13.2 Implémentation du modèle et transfert des outils de base

Une fois le modèle défini, l'étape suivante à été l'implémentation d'une version de base, ainsi que des outils nécessaire pour les utiliser.

Ces outils sont :

- un premier outil de prétraitement de manipulation des données ;
- un algorithme d'entraînement sur ces données, récupéré du projet précédent ;
- un outil d'analyse et de traçage des performances du modèle, qui est une version améliorée, plus fluide et plus puissante, des outils de visualisation du projet précédent (voir section 7.4) ;
- un système de sauvegarde et d'interruption de l'entraînement similaire à celui du projet précédent (voir sous-section 7.5.2)

13.2.1 Premiers résultats

Les premiers résultats du modèle sont encourageants, autant sur la rapidité du modèle que sur ses performance. Ils sont disponibles dans l'annexe D.3.

13.3 Adaptation du système de gestion de corpus au volume de données

L'étape suivant l'implémentation du modèle fonctionnel à été l'adaptation du système de gestion de corpus aux grands volumes de données à traiter dans le futur.

Pour cela, un premier prototype à été réalisé. Il utilise des mécanismes optimisés du langage Python pour la lecture de fichier. L'objectif de ce prototype est de maintenir uniquement une fiable quantité de données en mémoire.

Le principe de fonctionnement est simple :

1. on charge en mémoire une portion des données ;
2. on applique le prétraitement sur ces données ;
3. on transfert les données traitées à l'algorythme d'entraînement ;
4. on recommence de l'étape 1 avec la portion suivante des données.

L'intégration de cet outil s'est faite en parallèle de la suite du projet.

13.4 Entrainement par batch

La première optimisation du modèle à été l'intégration de l'algorithme d'entraînement par *batches* (voir ???).

Il à été nécessaire de définir le nombre de *batches* optimal en terme de temps de calcul et de performance. Les détails concernant ce choix sont disponibles dans l'annexe D.4.

Cette optimisation à permis d'accélérer considérablement l'entraînement du modèle.

13.5 Changement de métrique pour évaluer la qualité du modèle

Cette partie du projet n'est pas dédie à une optimisation, mais à la rectification d'une métrique inadaptée.

Jusqu'à cette étape, les performance étaient évaluées à partir du score utilisé pour mettre à jour le modèle. Ce score est difficile à utiliser pour se faire une idée réelle des performance du modèle.

Il à donc été remplacé par une métrique nommée « précision », qui est le taux de caractères correctement prédits (le bon caractère au bon endroit).

13.5.1 Précision de base

Ce changement de métrique à été l'occasion de déterminer ce que l'on appelle la précision de base (*baseline accuracy* en anglais).

Cette valeur représente la précision d'un modèle qui répondrait toujours la même chose. Elle permet de déterminer si le modèle produit apprend réellement de l'information, et dans quelle mesure.

Le détail de ces résultats est disponible dans les rapports des annexes D.6 et D.2.

13.6 Analyse d'une étrange variation de performance dans l'apprentissage

La seconde optimisation apportée au modèle était dédiée à réduire les sources d'erreur dans les données.

Dans les courbes d'apprentissage du modèle basique, d'étranges baisses de performances sont visibles, toujours sur la même portion des données.

Il a été décidé de consacrer du temps à l'analyse de ce phénomène.

L'étude du phénomène a révélé une forte présence de codes hexadécimaux (comme 0x005f ou 4A6D005F) dans le fragment des données mis en cause, codes qui semblent être la cause de la baisse de performance.

Il a donc été décidé, comme décrit sous-section 11.2.1, d'intégrer au prétraitement des données la gestion de ces codes. Cette intégration a cependant dû attendre la fin de la réalisation de la nouvelle version du gestionnaire de données (décrite ??).

Le déroulement de l'analyse et les conclusions détaillées sont disponibles dans l'annexe D.5.

13.7 Optimisation du taux d'apprentissage

La troisième optimisation mise en place pour le modèle et le réglage d'un des paramètres de l'algorithme d'apprentissage, est une valeur nommée « taux d'apprentissage ».

Ce paramètre permet de déterminer la vitesse d'apprentissage du modèle. Cependant, un mauvais réglage entraîne des conséquences catastrophiques sur le modèle, comme un apprentissage extrêmement lent, voir une divergence de l'apprentissage.

Il est donc nécessaire de correctement choisir ce paramètre. La procédure classique en apprentissage profond repose sur l'essai-erreur : on entraîne le modèle avec différentes valeurs pour le taux d'apprentissage, et on choisit celles qui ont donné le meilleur résultat pour les entraînements à venir.

Un module permettant la détermination du taux d'apprentissage idéal a donc été implémenté, et utilisé.

La nouvelle valeur définie pour le a permis d'augmenter largement la vitesse de convergence du modèle.

Le détail du processus de choix et des résultats est disponible dans les annexes D.7 et D.8.

13.8 Mise en place d'un système de gestion de corpus plus puissant

La dernière optimisation mise en place est la transformation du système de corpus en une version plus puissante.

13.8.1 Fichiers multiples

L'étude des données disponibles a révélé une structure en nombreux fichiers successifs. Ainsi, à la demande de notre maître de stage, nous avons donné au nouvel outil la capacité à utiliser plusieurs fichiers comme source de données continue. Cette fonctionnalité a été implémentée de façon optimisée en temps de calcul et en espace optimisé.

13.8.2 Prétraitement modulaire

D'autre part, durant le déroulement du projet nous avons remarqué que l'ajout d'une étape de prétraitement était ardue avec l'ancien outil. En gardant cela à l'esprit, nous avons développé un système de prétraitement modulaire, dans lequel chaque étape du traitement est séparée et interchangeable. Cette architecture permet d'ajouter ou de retirer à volonté des étapes au traitement. C'est d'ailleurs lors de l'implémentation de ces modules que le prétraitement mentionnée ?? et ?? à été intégré.

13.8.3 Processus multiples

Une autre amélioration, tirant profit de l'aspect modulaire du traitement, est l'utilisation de multiples processus en parallèle. Chacun d'entre eux est dédié à une étape du prétraitement. Cela permet de répartir la charge de travail sur les différents processus, à la façon d'un travail à la chaîne. Ainsi, l'outil est beaucoup plus rapide que la version précédente.

13.8.4 Performances

Le nouvel outil est largement plus rapide et efficace que son prédécesseur. De plus, il augmente la facilité d'utilisation et d'entretien.

La description des performances et du fonctionnement en détail de l'outil sont disponibles dans le rapport de l'annexe D.10.

14 Conclusions sur le projet PAPUD

Le projet PAPUD s'est déroulé sans accroc, et l'intégralité des objectifs ont été remplis.

On peut compter 4 optimisations principales, outre le prétraitement des codes hexadécimaux et le choix de la taille des *batches* :

- la réalisation d'un outil d'optimisation du taux d'apprentissage ;
- la réalisation d'un outil multi-fichiers multi-processus de gestion du corpus.

On peut noter l'attention particulière portée à la documentation du code. En effet, le travail effectué n'est que la première pierre du cas d'utilisation BULL du projet PAPUD. Il est donc normal que nous ayons laissé une base de code propre et intégralement documentée pour notre successeur sur le projet.

15 Rétrospective sur le stage

15.1 Bilan sur

15.2 Bilan professionnel

15.3 Bilan personnel

16 Discussion et perspectives

16.1 Poursuite du projet PAPUD

17 Rétrospective sur le stage

17.1 Bilan sur le travail effectué

17.2 Bilan professionnel

17.3 Bilan personnel

17.4 Poursuite du projet PAPUD

Annexes

Table des annexes

A Listes des tables, des figures et des fragments de code	67
B Glossaire, acronymes et noms d'entités	69
C Rapports d'avancement du projet GMSNN	75
D Rapports d'avancement du projet PAPUD	147
E Copie de la convention de stage	175
F Copie de l'avenant à la convention de stage	181

A Listes des tables, des figures et des fragments de code

Liste des tableaux

Liste des illustrations

Liste des fragments de code

B Glossaire, acronymes et noms d'entités

Glossaire

Apprentissage automatique

Défini sous-section 2.2.1 (page 5). p. 5, 6, 16

Apprentissage profond

Défini sous-section 2.2.2 (page 6). p. 3, 5, 6, 7, 8, 15, 17, 21, 39, 42, 55

batch

p. 32, 33, 34

Bogue

Un Bogue (*bug* en anglais) est « un défaut de conception d'un programme informatique à l'origine d'un dysfonctionnement. »D'après Wikipédia [bug] p. 25

Cloud

p. 43

Data-centers

p. 43

Différentiation automatique

p. 8, 33

embedding

Un *embedding* est un tenseur particulier produit par un module éponyme. Il est la représentation d'un caractère ou d'un mot. p. 26, 27, 51

État caché

p. 27

état de l'art

p. 16, 25, 27, 36, 39

Gitlab Flavoured Markdown

Le Gitlab Flavoured Markdown est une variante du Markdown supportant des fonctionnalités particulières telles que l'intégration d'images et de cases à cocher. p. 75, 147

GMSNN

Défini ?? (page ??). p. 16, 21, 22, 25, 26, 27, 33, 34, 35, 39, 51

GPU

Un processeur graphique (*Graphical Processing Unit* en anglais, GPU) est un composant d'ordinateur spécialisé, qui montre d'excellentes performances dans les calculs impliquant des matrices (ex. : images) p. 18

Grammaires formelles

p. 14

GRU

Défini sous-section 2.3.5 (page 9). p. 9, 15

LSTM

Défini sous-section 2.3.5 (page 9). p. 9, 15, 25, 27, 32

Markdown

« Markdown est un langage de balisage léger [dont le] but est d'offrir une syntaxe facile à lire et à écrire. Un document balisé par Markdown peut être lu en l'état sans donner l'impression d'avoir été balisé ou formaté par des instructions particulières. » D'après Wikipédia¹. p. 75

Matrice

p. 69

Modèle

Défini sous-section 2.3.2 (page 6). p. 6, 16, 18, 21, 25, 26, 27, 30, 31, 32, 33, 34, 35, 37, 38

Modèle de la Langue

Un Modèle de la Langue (*Language Model* en anglais) est une « distribution de probabilité sur une séquence de mots [ou de caractères] », utilisé pour estimer la probabilité d'apparition du prochain mot ou caractère. Autrement dit, c'est une représentation servant à prédire le prochain mot à partir des mots précédents. D'après Wikipédia². p. 3, 14, 15, 21, 25, 36, 39, 45, 51

Module GMSNN

Ce module est un RNN. C'est l'objet principal du projet GMSNN p. 27, 33

Non-linéarité

Défini Figure 2.3.4 (page 7). p. 7

Paramètre

p. 32, 33, 34

Poids

Défini ?? (page ??). p. 33, 34

Réseau de neurones

Défini sous-section 2.3.1 (page 6). p. 3, 4, 5, 6, 7, 8, 9, 15, 21, 29, 33, 39, 45, 51

RNN

Défini sous-section 2.3.5 (page 9). p. 9, 15, 21, 25, 26, 27, 29, 32, 34, 39, 45

Sémantique computationnelle

p. 14

Tenseur

Un tenseur en apprentissage profond est un type de matrice particulier adapté à la technique de la différentiation automatique p. 26, 30, 51

1. [wiki_md](#).

2. [wiki_lm](#).

Traitement de la parole

p. 14

Traitement préalable des données

Défini sous-section 2.3.2 (page 7). p. 6

Traitement Automatique des Langues

Défini sous-section 2.2.3 (page 6). p. 3, 5, 6, 13, 14

Acronymes

GMSNN

p. *Glossaire* : GMSNN

GPU

p. *Glossaire* : GPU

GRU

p. *Glossaire* : GRU

LSTM

p. *Glossaire* : LSTM

Modèle de la Langue

p. *Glossaire* : Modèle de la Langue

RNN

p. *Glossaire* : RNN

Entités, projets et sigles

AREQ

Assemblée des Responsables des Équipes p. 12

ATOS

p. 43

BULL

p. 42, 43, 45

CNRS

Centre National de la Recherche Scientifique p. 12, 13

EUREKA

« EUREKA est une initiative européenne, intergouvernementale, destinée à renforcer la compétitivité de l'industrie européenne. » D'après Wikipedia³. p. 42, 73

Gitlab

Gitlab p. 75

Grid5000

p. 18, 25

INRIA

Institut National de Recherche en Informatique et en Automatique p. 12

ITEA3

troisième instance d'ITEA (*Information Technology for European Advancement*), une initiative de recherche, développement et innovation du réseau EUREKA. p. voir aussi EUREKA, 1, 38, 42

LORIA

Laboratoire Lorrain d'Informatique et ses Applications p. 12, 14

PAPUD

Profiling and Analysis Platform Using Deep Learning p. 1, 38, 42

projet GMSNN

projet basé sur une proposition innovante d'architecture de réseau de neurones, faite par M. Christophe Cerisara p. 4, 16, 27, 35, 38, 46, 75

projet PAPUD

projet ITEA3-PAPUD, cas d'utilisation BULL p. 1, 3, 4, 17, 38, 39, 42, 43, 45, 46, 57, 147

PyTorch

Librairie Python dédiée à l'apprentissage profond. p. 18, 25, 32, 33, 35

SYNALP

SYNALP (*SYmbolic and statistical NATural Language Processing*) est une équipe de recherche du département 4 du LORIA p. 4, 12, 13, 14, 42

Université de Lorraine

Université de Lorraine p. 12, 13

3. [wiki_eureka](#).

C Rapports d'avancement du projet GMSNN

C.1 Informations sur les rapports contenus dans la présente annexe

Les sections suivantes contiennent les rapports intermédiaires fournis à notre maître de stage au cours du projet GMSNN.

C.1.1 Format d'origine des rapports

Le langage Markdown, plus spécifiquement dans le dialecte nommé Gitlab Flavoured Markdown, fournit une syntaxe facile à lire et à écrire. Il permet la rédaction de documents agrémentés entre autres d'images, de formules, de tableaux et de fragments de codes. Enfin, l'affichage du Gitlab Flavoured Markdown est supporté par Gitlab.

Ces particularités en font un langage de premier choix pour l'écriture de rapports destinés à être lus au format informatique directement sur Gitlab.

C.1.2 Transcription des rapports

L'intégration des rapports intermédiaires dans ce rapport à nécessité l'adaptation du contenu en Gitlab Flavoured Markdown au format papier.

Certains éléments n'ont pas pu être transcrit tels-quels, en particulier les liens, et les tableaux et images de grande taille.

C.1.3 Contenu et langue des rapports

Le contenu des rapports n'a été ni modifié ni corrigé, et est livré en anglais tel qu'écrit à l'origine.

L'anglais a été choisi comme langue de rédaction des rapports pour maintenir la cohérence avec le code, écrit et documenté en anglais lui aussi, et avec la littérature, principalement rédigée en anglais. Ce choix évite aussi d'alourdir le contenu déjà complexe des documents avec des traductions maladroites de termes techniques.

C.2 Étude des problèmes de mémoire

Analysis of memory usage

Analysis report

by E. Marquer, 2018/05/23, Synalp and Université de Lorraine

C.2.1 Abstract

Experimental results shows (using nvidia-smi) an increasing memory usage for 4 layers, from an already enormous 3GB to more than 6GB, causing an out-of-memory error.

The objective of the following computations is to estimate the memory consumption of the model, to confirm the hypothesis of a memory leak, and verify that the model should not overflow memory.

C.2.2 Formulas

Tensor and Variable size estimation

Byte size of a tensor is close to 6 times the products of all of its dimensions. Byte size of a variable is similar to that of the corresponding tensor.

```
1 import torch, pickle
2
3 # Object to mesure
4 o = torch.autograd.Variable(torch.ones(100, 100, 100))
5 o = torch.ones(100, 100, 100)
6
7 len(pickle.dumps(o, 0)) / (100 * 100 * 100)
8 #result = 6
```

Computations

$$\begin{aligned}
total &= hidden_states + msnn_weights + emb_weights + out_weights \\
&= detach_interval * (growth_factor + 1) * layers * 6 * (hidden_size * batch_size * sequence_length) \\
&\quad + (((layer - 1) * 8 * hidden_size * (hidden_size + 1)) \\
&\quad + 4 * hidden_size * (hidden_size + emb_size + 2)) * 6 \\
&\quad + (nwords * (emb_size + 1)) * 6 \\
&\quad + (nwords * (layers * hidden_size)) * 6 \\
\\
&= 6 * (detach_interval * (growth_factor + 1) * layers * (hidden_size * batch_size * sequence_length)) \\
&\quad + ((layers - 1) * 8 * hidden_size * (hidden_size + 1)) \\
&\quad + 4 * hidden_size * (hidden_size + emb_size + 2) \\
&\quad + (nwords * (emb_size + 1)) \\
&\quad + (nwords * (layers * hidden_size)))
\end{aligned} \tag{C.1}$$

$$\begin{aligned}
hidden_states &= total_history * 6 * dim \\
&= detach_interval * (growth_factor + 1) * layers * 6 * dim \\
&= detach_interval * (growth_factor + 1) * layers * 6 * \\
&\quad (hidden_size * batch_size * sequence_length)
\end{aligned} \tag{C.2}$$

$$\begin{aligned}
msnn_layer_weights &= weights_ih + weights_hh + bias_ih + bias_hh \\
&= 4 * hidden_size * input_size + 4 * hidden_size * hidden_size \\
&\quad + 4 * hidden_size + 4 * hidden_size \\
&= 4 * hidden_size * (hidden_size + input_size + 2) \\
&= \begin{cases} 8 * hidden_size * (hidden_size + 1) & \text{for all layers except the first one} \\ 4 * hidden_size * (hidden_size + emb_size + 2) & \text{for the first layer} \end{cases}
\end{aligned} \tag{C.3}$$

$$\begin{aligned}
msnn_weights &= msnn_layer_weights * layers \\
&= ((layers - 1) * 8 * hidden_size * (hidden_size + 1)) \\
&\quad + 4 * hidden_size * (hidden_size + emb_size + 2)
\end{aligned} \tag{C.4}$$

$$\begin{aligned}
emb_weights &= (bias + weights) * 6 \\
&= (nwords * emb_size + emb_size) * 6 \\
&= (nwords * (emb_size + 1)) * 6
\end{aligned} \tag{C.5}$$

$$out_weights = (nwords * (layers * hidden_size)) * 6 \tag{C.6}$$

Estimate with basic set of parameters

```
1 detach_interval = 50
2 growth_factor = 5
3 layers = 7
4 hidden_size = 1840 / 4
5 batch_size = 2
6 sequence_length = 100
7 emb_size = 400
8 nwords = 205
9
10 total = 6 * (detach_interval * (growth_factor + 1) * layers * hidden_size *
11     batch_size * sequence_length + ((layers - 1) * 8 * hidden_size * (
12         hidden_size + 1)) + 4 * hidden_size * (hidden_size + emb_size + 2) + (
13         nwords * (emb_size + 1)) + (nwords * (layers * hidden_size)))
14
15 " {}GB {}MB {}kB {}B".format(int(total%(1024**4) / 1024**3), int(total
16     %(1024**3) / 1024**2), int(total%(1024**2) / 1024), int(total%1024))
17 # result: '1GB 741MB 614kB 9B'
```

detach_interval	growth_factor	layers	hidden_size	batch_size	sequence_length	emb_size	nwords	total
50	5	7	1840 / 4	2	200 / batch_size	400	205	1GB 153MB 68kB 6B
100	5	7	1840 / 4	2	200 / batch_size	400	205	2GB 234MB 579kB 262B
200	5	7	1840 / 4	2	200 / batch_size	400	205	4GB 397MB 577kB 774B
50	10	7	1840 / 4	2	200 / batch_size	400	205	2GB 50MB 323kB 390B
50	5	6	1840 / 4	2	200 / batch_size	400	205	1008MB 912kB 414B
50	5	5	1840 / 4	2	200 / batch_size	400	205	840MB 732kB 822B
50	5	4	1840 / 4	2	200 / batch_size	400	205	672MB 553kB 206B
50	5	3	1840 / 4	2	200 / batch_size	400	205	504MB 373kB 614B
50	5	2	1840 / 4	2	200 / batch_size	400	205	336MB 193kB 1022B
50	5	1	1840 / 4	2	200 / batch_size	400	205	168MB 14kB 406B
50	5	7	1840 / 2	2	200 / batch_size	400	205	2GB 431MB 598kB 94B
50	5	7	1840 / 8	2	200 / batch_size	400	205	573MB 28kB 890B
50	5	7	1840 / 4	1	200 / batch_size	400	205	1GB 153MB 68kB 6B
50	5	7	1840 / 4	2	200	400	205	2GB 234MB 579kB 262B
50	5	7	1840 / 4	2	200 / batch_size	200	205	1GB 150MB 743kB 534B
50	5	7	1840 / 4	2	200 / batch_size	800	205	1GB 157MB 764kB 998B
50	5	7	1840 / 4	2	200 / batch_size	400	500	1GB 159MB 182kB 984B

Most impactful factors (memory-wise)

- detach_interval : detach_interval * 2 = memory * 2
- growth_factor : growth_factor * 2 = memory * 2
- hidden_size : hidden_size * 2 = memory * 2
- batch_size * sequence_length (number of examples by sequence) : batch_size*sequence_length * 2 = memory * 2
- layers : layers * 2 = layers * 2

The others factors considered (emb_size and nwords) have almost no impact on memory. It confirms that the most memoryphage element is the MSNN. Also, to keep a stable memory usage, batch_size * sequence_length ratio must be kept constant; increasing batch_size while lowering sequence_length can increase processing speed whithout impacting memory usage.

Impact of layer increase

We can notice that the impacts of layers is most noticeable during the first phases of training, during the creation of the first layer. Later on, the creation frequency of new layers is small, and the change is minimal. For example, from 6 to 7 layer, we need ‘12500‘ sequences to pass, and the increase of memory of about ‘7/6‘; from 7 to 8 layer, we need ‘62500‘ sequences to pass, and the increase of memory of about ‘8/7‘; from 8 to 9 layer, we need ‘312500‘ sequences to pass, and the increase of memory of about ‘9/8‘; and so on.

Partial derivatives :

```
1 d = detach_interval
2 g = growth_factor
3 l = layers
4 h = hidden_size
5 b = batch_size
6 s = sequence_length
7 e = emb_size
8 n = nwords
9
10 complete formula:
11 6 *
12   d * (g + 1) * l * h * b * s +
13   ((l - 1) * 8 * h * (h + 1)) +
14   4 * h * (h + e + 2) +
15   (n * (e + 1)) +
16   (n * (l * h))
17 )
18
19 simplified formula:
20 6*(8*l*h^2 + 1*d*g*h*b*s + 1*d*h*b*s + 8*l*h + l*n*h + 4*e*h + e*n + n - 4*h
21 ^2)
22
23 partial derivate on given variable:
24 d: (6(8*l*h^2 - 4*h^2 + 1*d*g*h*b*s + 1*d*h*b*s + 8*l*h + 4*e*h + l*n*h + e*n
25 + n))
26 g: 6*l*d*h*b*s
27 l: 6*(d*h*b*s*(g+1) + 8*h*(h+1) + nh)
28 h: 6*(l*d*g*b*s + l*d*b*s + l*n + 16*l*h + 8*l + 4*e - 8*h)
29 b: 6*l*d*h*s*(g+1)
30 s: 6*l*d*h*b*(g+1)
31 e: 0
32 n: 6*(l*h+e+1)
```

C.3 Sauvegarde du modèle

Analysis and implementation of job save and restart

Implementation report

by E. Marquer, 2018/05/24
Synalp and Université de Lorraine

C.3.1 Abstract

As the jobs are taking longer than an hour per epoch, it has become necessary to keep an image of the model, the training parameters and the current state of the model to be able to interrupt training when needed.

Multiple choices are available :

- an easy but heavy serialisation (pickle) of the whole system;
- a “full” save of the model and parameters, excluding everything that can be recomputed easily;
- a “partial” save of the model, removing a part of the less important information.

C.3.2 End solution : saving using pytorch utilities

The saving solution currently implemented is the `torch.save(trainer, file)` and `trainer = torch.load(file)` and the alternatie version `trainer = torch.load(file, map_location=lambda storage, loc: storage)` allowing the loading of a CUDA model out of CUDA.

This solution posed a number of problems when it was first tested (internal non-parameter attributes where not correctly saved and loaded), that is why it was not considered at first.

But a more recent try resulted in a perfect result, replacing the need of a custom serialisation system.

C.3.3 Full serialisation

This kind of serialisation is done thanks the pickle package, on the whole Trainer object.

```
1 import pickle
2
3 def load(filename: str) -> object:
4     with open(filename, 'rb') as f:
5         return pickle.load(f)
6
7 def save(filename: str, trainer: object) -> None:
8     with open(filename, 'wb') as f:
9         pickle.dump(trainer, f)
```

C.3.4 Full save

Elements to save

- Corpus file name
- cuda_on
- batch_size
- Trainer
 - Model (see specific points)
 - self.epoch current epoch
 - self.batch and self.i current position in training epoch
 - self.start_time needs a little work : storing the elapsed time, and when loading, remove elapsed time from load time
 - self.log_interval
 - self.save_interval
 - self.bptt
 - self.nwords
 - self.repackage_interval
 - self.repackage_strategy
 - self.reset_growth
 - self.reset_hidden
 - self.epochs
 - self.save_folder
- Model (MSNN)
 - self.training
 - input layer (embeddings) using torch.save(self.emb, f) and torch.load(f)
 - MSNN (see specific points)
 - output layer (linear)
- MSNN
 - Final
 - self.input_size
 - self.hidden_size
 - self.growth_factor
 - self.batch_size
 - self.cuda_on
 - self.layer_id
 - self.max_detach
 - self.repackage_strategy
 - self.max_layers
 - Next MSNN
 - self.detach_count
 - self.rnn
 - self.hidden
 - self.seq_count
 - self.detach_count
 - self.transmitted_output
 - self.transmitted_hidden

Elements to forget and recreate

- Trainer
 - Corpus
 - Corpus file name is needed
 - Optimizer : resign learning rate, weight_decay and model's parameters (create after model is loaded) using `torch.optim.SGD(model.parameters(), lr=args.lr, weight_decay=args.weight_decay)`
 - Criterion : `criterion = torch.nn.CrossEntropyLoss()`
 - self.layers using `self.model.msnn.get_layers()` (create after model is loaded)
 - self.train_data using `batchify(corpus.train, batch_size, cuda_on)`
 - self.val_data using `batchify(corpus.valid, batch_size, cuda_on)`
 - self.plotdata using {"Epoch": None, "Layer": None, "Frac": None} and `self.init_plotter()` (perhaps with a new file, otherwise in append mode without cleaning the file)
 - self.msnn_backup, should be empty if the new backup system (using the with statement) is implemented
- Model
 - None
- MSNN
 - self.tensors

C.3.5 Partial save

The elements to keep are the same as with the Full save, except : - hidden states and transmitted output are to be detached

C.3.6 Other things that need to be added to interrupt and resume job

Methods to interrupt and resume epoch loop and training loop

Interruption strategy Interruption can be done by :

- saving the state at specific timestep ;
- saving automatically when shutting down.

The second option isn't really realistic, as an interruption wouldn't allow a big enough margin to save the model. Even though, it could be added to allow manual interruption at certain timesteps (smaller than the ones implemented in the first option).

As the first option is the only viable one, multiple timesteps are possible, from the shortest to the longest :

- after each operation (example : after forward pass, and after backward pass, and after optimization ...);
- after each sequence ;
- after n sequences ;
- after each epoch ;
- after n epochs ;
- after the whole training.

Option [1.] is not viable, as it would consume a lot of time relative to each computation for the sole writing process. Option [3.], of which option [2.] is a specific case, allows both a fine granulation with only a minimal loss if anything were to occur, and a reduced burden computation-wise. From option [4.] onward, the save time is negligible, and even if the granulation is mediocre, it offers fine milestones for specific usage (usage of an already trained mode for example).

Currently, at creation time, with no history, the model save file is about 770MB.

What will be implemented is the third option, saving the model after n sequences.

Resume strategy It will be necessary to adapt the training loops a little to allow resuming at any sequence in any epoch.

C.3.7 Other methods implemented

Method using **with** keyword for backup system, during evaluation

```
1 class BackupContextManager:
2     def __init__(self, model: DetRNN):
3         self.model = model
4
5     def __enter__(self):
6         # Backing up training linked data
7         self.msnBackup = self.model.msnBackup()
8
9     def __exit__(self, type, value, traceback):
10        # Restoring training linked data
11        self.model.msnRestore(self.msnBackup)
12
13
14 with BackupContextManager(model):
15     """Do computations"""
```

C.4 Solution tentative pour les fuites mémoires

Analysis of a source of the memory leak problem : the history management system

Analysis report

by E. Marquer, 2018/05/28, Synalp and Université de Lorraine

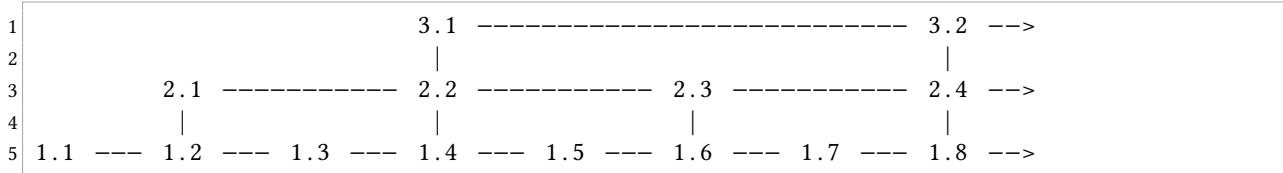
C.4.1 Abstract

A big memory leak is present in the model. One of the identified cause is a malfunction in the history management system.

C.4.2 Problem

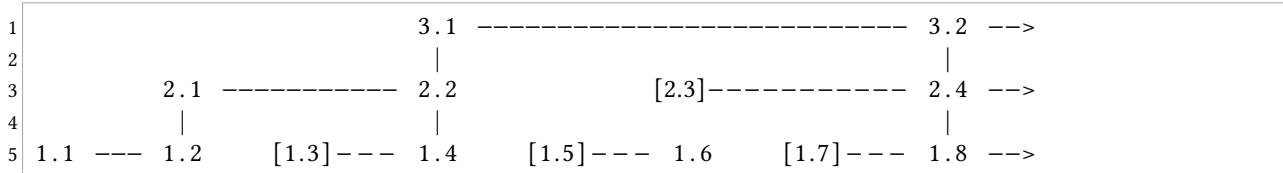
It seems simply detaching a hidden state is not enough :

With a graph :

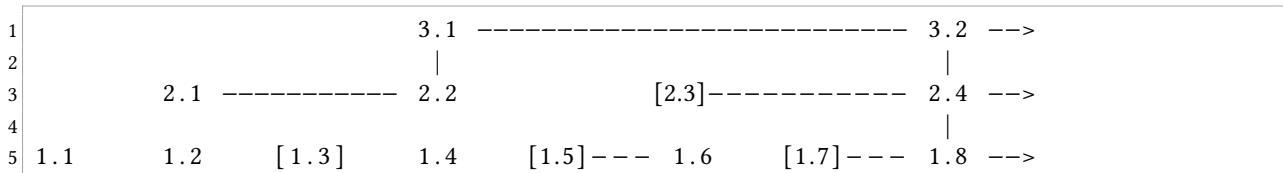


where all nodes are hidden states, and each layer is a line, with a transmission rate of 1 transmission every 2 hidden state.

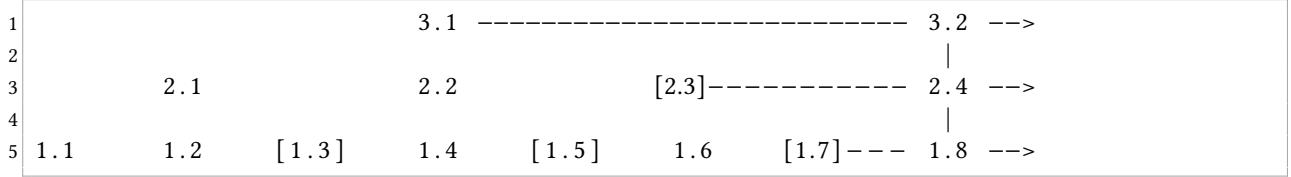
When detaching every 2 hidden, with [i,j] a detached node, the graph becomes :



But it should be :



Or with a more aggressive strategy :



C.4.3 Solution

To solve the problem, keeping track of all history is necessary, and with the way PyTorch works, it won't be a problem to keep references to history before deleting them

C.5 Problèmes théoriques liés à l'entraînement batch par batch

Analysis of batch training in msnn

Analysis report

by E. Marquer, 2018/05/29, Synalp and Université de Lorraine

C.5.1 Abstract

A common method to improve training time of a RNN is the batch based training, but MSNN are highly dependent on past history and continuity. This training strategy is based on passing simultaneously multiple inputs to the network. Training with 3 batches is equivalent to a parallel training over 3 corpora composed of respectively every 1st batch, every 2nd batch, and every 3rd batch. It necessitates the splitting of the corpus, and doing so breaks the continuity between the different parts. As such, it would be difficult to use the batch based training for MSNN.

C.5.2 Bacthifying strategies

There are multiple batchifying strategies, here explained with the Alphabet as a corpus.

MSNN is currently trained with the BPTT (and Truncated-BPTT) algorithm. By passing multiple sequences, there are two possible batchifying :

- batchifying across each sequence of the corpus ;
- batchifying across the corpus then sequence of the corpus.

No batchifying

Table 1

Batch\Timestep	1	2	3	4	5	6	7	8	9	10	11	...	24	25	26
1	A	B	C	D	E	F	G	H	I	J	K	...	X	Y	Z

BPTT sequence-wide batchifying

Example with a BPTT sequence length of 3 (first inputs of the sequence are in bold) :

Table 2

Batch\Timestep	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	A	B	C	G	H	I	M	N	O	S	T	U	Y	Z
2	D	E	F	J	K	L	P	Q	R	V	W	X		

Corpus-wide batchifying

Batch\Timestep	1	2	3	4	5	6	7	8	9	10	11	12	13
1	A	B	C	D	E	F	G	H	I	J	K	L	M
2	N	O	P	Q	R	S	T	U	V	W	X	Y	Z

Other batchifying strategies

Other batchifying strategies exists, mostly by spreading the corpus along the batch dimension and not the timestep dimension.

One such strategy would be :

Batch\Timestep	1	2	3	4	5	6	7	8	9	10	11	12	13
1	A	C	E	G	I	K	M	O	Q	S	U	W	Y
2	B	D	F	H	J	L	N	P	R	T	V	X	Z

C.5.3 Analysis of the different strategies

Spreading the corpus along the batch dimension

The worst possible strategies seem to be C.5.2spreading the corpus along the batch dimension : each input of the corpus is a succession of characters separated by a gap of the length of the batch dimension, so :

- the input is always incomplete, as the different batches do not interact with each other;
- the batches are full of discontinuity;
- the network is not reusable for any number of batches and input.

Sequence batchifying

Then there is the BPTT sequence-wide batchifying. In each sequence, the batches are internally coherent, but between sequences, the batches are discontinued (C|G, F|J, ...). Moreover, if the first layers' input are internally coherent, upper layers are subjected as similar input as presented in C.5.2, and the problems of the corresponding strategy is back.

Corpus batchifying

The last and best strategy is the Corpus-wide batchifying. Even if this one need the pre-processing of the corpus, it offers a lot of advantages :

- the input is not discontinued, even in the last layer;

- the network is usable for any number of batches, as such strategy is equivalent to a parallel training over distinct corpora, each composed of a part of the original corpus;
- the only layers really affected by the remaining discontinuity are the last ones, as over multiple epochs, they would only get the same part of the corpus and would not be able to extract info from the whole corpus :

Table 5

Batch\Epoch	1	2	3	4	...
1	Part 1	Part 1	Part 1	Part 1	...
2	Part 2	Part 2	Part 2	Part 2	...
3	Part 3	Part 3	Part 3	Part 3	...

This last discontinuity can be solved by a rotation of the batches across multiple epochs :

Table 6

Batch\Epoch	1	2	3	4	...
1	Part 1	Part 2	Part 3	Part 1	...
2	Part 2	Part 3	Part 1	Part 2	...
3	Part 3	Part 1	Part 2	Part 3	...

But that solution leaves the problem of the hidden state, which contains important information, and exists in multiple different exemplary, one for each batch. Another consequence is the need of n^* epochs, for n batches, to accumulate equivalent history.

C.5.4 Conclusion

Corpus-wide batchifying seems to be the best of all batchifying strategies, but it still presents multiple downsides :

- the existence of multiple parallel “memory”;
- the need of n^* epochs, for n batches, to accumulate equivalent “memory” for each batch compared to non-batchified training.

C.6 Tentative de réduction du temps de calcul par utilisation de l'algorithme d'entraînement « *Truncated BPTT* »

Analysis and implementation of improved Truncated-BPTT training algorithm

Implementation report

by E. Marquer, 2018/05/29
Synalp and Université de Lorraine

C.6.1 Abstract

Last update (implementation of explicit history to solve memory leaks problem) solved memory problems, leaving the time consumption problem (hundreds of hours for a single epoch).

As the most time-consuming process is the backpropagation, the most evident way to reduce time consumption is to improve the training strategy.

One of the possible optimization is the Truncated Backpropagation Through Time (Truncated-BPTT or TBPTT).

C.6.2 Notations

The following notations are from an introduction article on BPTT [**bptt-intro**] :

- **TBPTT(n,n)** : Updates are performed at the end of the sequence across all timesteps in the sequence (e.g. classical BPTT).
- **TBPTT(1,n)** : timesteps are processed one at a time followed by an update that covers all timesteps seen so far (e.g. classical TBPTT by Williams and Peng).
- **TBPTT(k1,1)** : The network likely does not have enough temporal context to learn, relying heavily on internal state and inputs.
- **TBPTT(k1,k2)**, where $k_1 < k_2 < n$: Multiple updates are performed per sequence which can accelerate training.
- **TBPTT(k1,k2)**, where $k_1 = k_2$: A common configuration where a fixed number of timesteps are used for both forward and backward-pass timesteps (e.g. 10s to 100s).

The base implementation of the model, using the (`sequence_length`, `batch_size`, `values`) model for the inputs (and outputs), is already an implementation of the **TBPTT(n,n)** algorithm.

What would improve a lot time efficiency is to implement a **TBPTT(k1,k2)** algorithm.

C.6.3 Algorithm

The algorithm can decompose into 4 steps : 1. Present a sequence of k_1 timesteps of input and output pairs to the network. 2. Compute loss across the k_2 last timesteps. 3. Backpropagate loss 4. Update weights

C.6.4 Pseudo-python code

Old algorithm

```
1 for sequence in sequences:
2     # 1. Present a sequence of *n* timesteps of input to the network.
3     output, hidden = model.forward(sequence.input, hidden)
4
5     # 2. Compute loss across the *n* timesteps.
6     loss = criterion(output, sequence.targets)
7
8     # 3. Backpropagate loss
9     loss.backward()
10
11    # 4. Update weights
12    optimizer.step()
```

New algorithm

```
1 for sequence in sequences:
2     # 1. Present a sequence of *k1* timesteps of input to the network.
3     output, hidden = model.forward(sequence.input, hidden)
4
5     # 2. Compute loss across the *k2* last timesteps.
6     loss = criterion(output[:-k2], sequence.targets[:-k2])
7
8     # 3. Backpropagate loss
9     loss.backward()
10
11    # 4. Update weights
12    optimizer.step()
```

C.7 Entrainement couche par couche

Layer by layer training

Analysis report

by E. Marquer, 2018/06/25, Synalp and Université de Lorraine

C.7.1 Abstract

Multiple advanced training algorithms use layer “freezing”, meaning that the layer will not be trained.

As it would be interesting to use those algorithms to get the most out of the current architecture, a layer “freezing” will be implemented. To do so, a dummy algorithm will be implemented. This algorithm is a naive layer by layer training.

C.7.2 Layer by layer training

This training is used to see if convergence can be sped up, if performance can be improved, and if the layered architecture is of any use.

Principle

The general principle is to train each layer individually, and to fine-tune them together frequently.

An iterative presentation would be :

1. Create a layer
2. Train the layer alone
3. Train all the layers together
4. Restart from [1.]

Another way to explain this algorithm is that it is a variation of the IM algorithm, were storing the output of the training of a layer is replaced by recomputing those results. It removes the drawback of the increase in memory usage, while speeding up training(time-wise at least, convergence-wise at best).

Example for a 4 layered MSNN

1. We train for n epochs the first layer. It is expected to learn a maximum of things of a low level of abstraction and time dependency.
2. Then, we freeze this layer, and start training a second one over n epochs. It is expected to learn a higher level of abstraction and time dependency from the representations in the first layer's hidden state.
3. We train those two layers together to fine-tune them.
4. We then add a third layer, freeze the first two layers, and train the third layer of information extracted from the second layer.
5. We train the three layers together.
6. We add a new layer and train it alone.
7. We train the four layers together.
8. We train all the layers until the ends of epochs.

Speeding up convergence and improving performance

Training each layer individually requires less computation during backpropagation. Moreover, by pushing each layer to learn the maximum of information it can learn, we can expect each layer to specialize at their scale.

A more detailed way to understand the intuition behind that is that a layer closer to the data has to learn very basic features. Step by step, every layer is constrained to its respective scale (by its memory span due to the architecture). Each layer has a minimal set of knowledge to learn before benefiting to the whole network. Even if a part of this learning is shared over the layers, at least over the first epochs there would be nothing to gain but noise by training all the layers together.

Globally, by skipping the “noisy” part of the training, a reduction of training time (the real computation time, not the number of epochs needed) is to be expected. A bonus effect would be a small acceleration of convergence, as training would be less noisy.

Use of the layered architecture

By training the model layer by layer, we can expect to see if training a single layer with equivalent parameters would have the same effect (if the individual training time is big enough).

C.8 Premier test du modèle réimplémenté

Test run of `detrnn.py`

Test report

by E. Marquer, 2018/04/26, Synalp and Université de Lorraine

C.8.1 Abstract

Test to see time needed, with GPU and without, to run the basic model of DetRNN.

C.8.2 Paradigm

With branch *reimplement*, allocated time 30 min. Test run of *detrnn.py* with full log output, with and without *cuda*. It does not matter if learning ends, as test is only to get time statistics.

Node

grimani-2, with GPU

C.8.3 Results

About 3 to 4 batch-sets are computed in 5 minutes without *cuda*. About 1 batch-set is computed in 1 s with *cuda*.

Details :

- without GPU : [detrnn2018_4_26-16h39.log](#)
- with GPU : [detrnn2018_4_26-16h41.log](#)

C.8.4 Potential ameliorations & next steps

Next step is to test the state-of-the-art version, but probably only with *cuda*.

C.9 Premier entraînement complet du modèle réimplémenté

Test run of `detrnn.py`

Test report

by E. Marquer, 2018/04/27 Synalp and Université de Lorraine

C.9.1 Abstract

Test to do a complete run, with GPU, of the basic model of DetRNN.

C.9.2 Paradigm

With branch *reimplement*, allocated time 24h, not interactive

Test run of *detrnn.py* with info log output, with *cuda*, for 4 epochs.

With curve auto-plotting, and plot data backup in case of interruption.

Node

OAR_JOB_ID=1554682 with GPU

Job start time : 2018-04-27 14 :11 :00

Estimated job stop time : 2018-04-28 14 :11 :00

Command used : bash oarsub -q production -p "GPU <> 'NO'" -l "nodes=1,walltime =24:00:00" ~/awd-lstm-lm/rundet.sh

C.9.3 Results

Total run time for 4 epochs : 5h33

The most rapid progress was during first epoch, with a maximal decrease of loss of 3/epoch, then the decrease of loss became a constant 0.25/epoch.

Plot

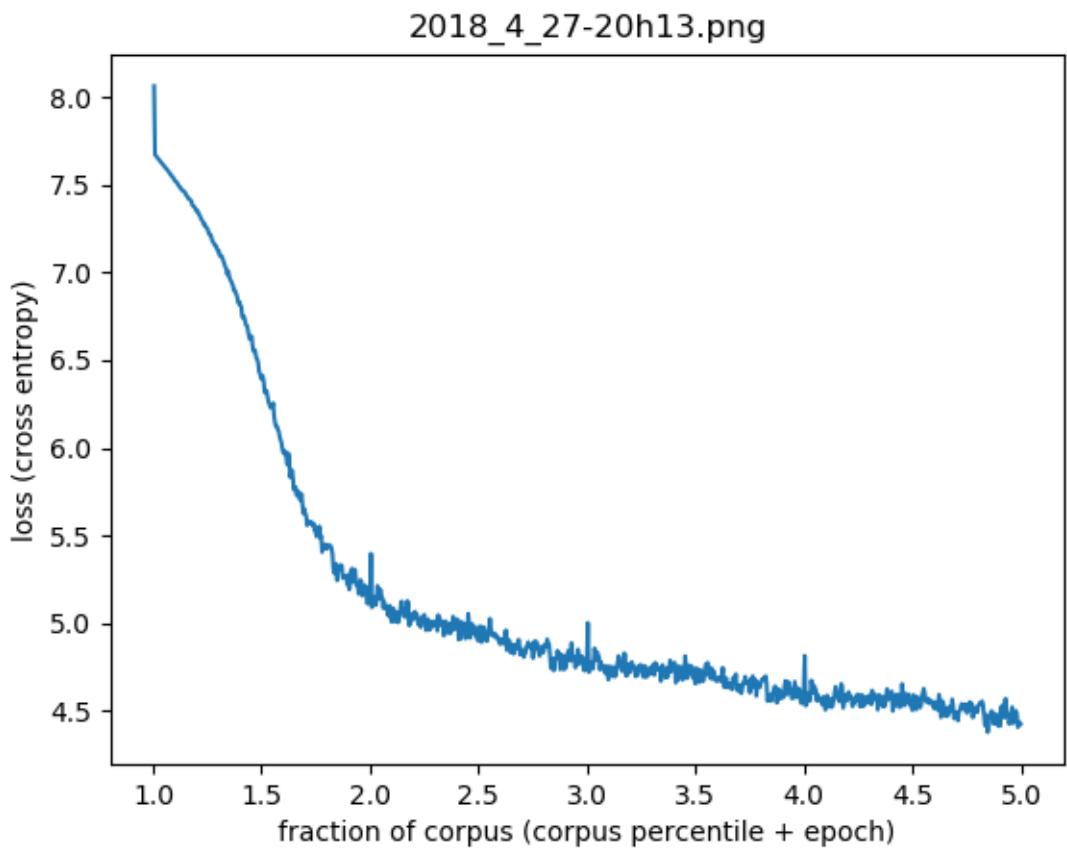


FIGURE C.1 – plot

Details

- log [detrnn2018_4_27-14h11.log](#)
- plot [2018_4_27-16h42.png](#)

C.9.4 Potential ameliorations & next steps

Next step is to test with more epochs, or test the growing model.

C.10 Premier entraînement à 50 époques du modèle réimplémenté

Test run of *detrnn.py*

Test report

by E. Marquer, 2018/04/30, Synalp and Université de Lorraine

C.10.1 Abstract

Test to do a complete run on 50 epoch, with GPU, of the basic model of DetRNN.

C.10.2 Paradigm

This test run of *detrnn.py*, with INFO level log output, loss by percentile and vbpc by epoch, will be executed with *cuda*, for 50 epochs.

The test is done with branch *reimplement*, an allocated time of 76h, not interactive

Run time was estimated for 50 epochs according to the results for 4 epochs (see [2018-04-27_test_run_detrnn.md](#)) :

$$(5\text{h } 33\text{min} / 4 \text{ epoch}) * 50 \text{ epoch} = 4162.5\text{min} = 69\text{h } 22\text{min } 30\text{s}$$

With a security margin of 10h, partially due to reduced batchsize, run time is 80h.

/ Had to reduce batchsize down to 40 because of memory errors */*

Node

OAR_JOB_ID=155659 with GPU

Job start time : 2018-04-30 12 :02 :08

Estimated job stop time : 2018-05-03 16 :02 :08

Command used : bash oarsub -q production -p "GPU <> 'NO'" -l "nodes=1,walltime =80:00:00" ~/awd-lstm-lm/rundet.sh

C.10.3 Results

Total run time for 50 epochs : with real stop time of 2018-05-02 17 :16 :56, the total run time of the training is approximately 53h (2days 5h), corresponding to a little more than an hour per epoch.

BPC-wise, the DetRNN hardly goes under 2.7 even after 50 epoch, with a change of 0.5 BPC in the last 30 epochs.

We can postulate that even after 200 epoch, the DetRNN will not have a BPC under 2.

Plot

BPC/fraction of corpus BPS per fraction of the corpus (an interval of 1 correspond a complete corpus, or an epoch).

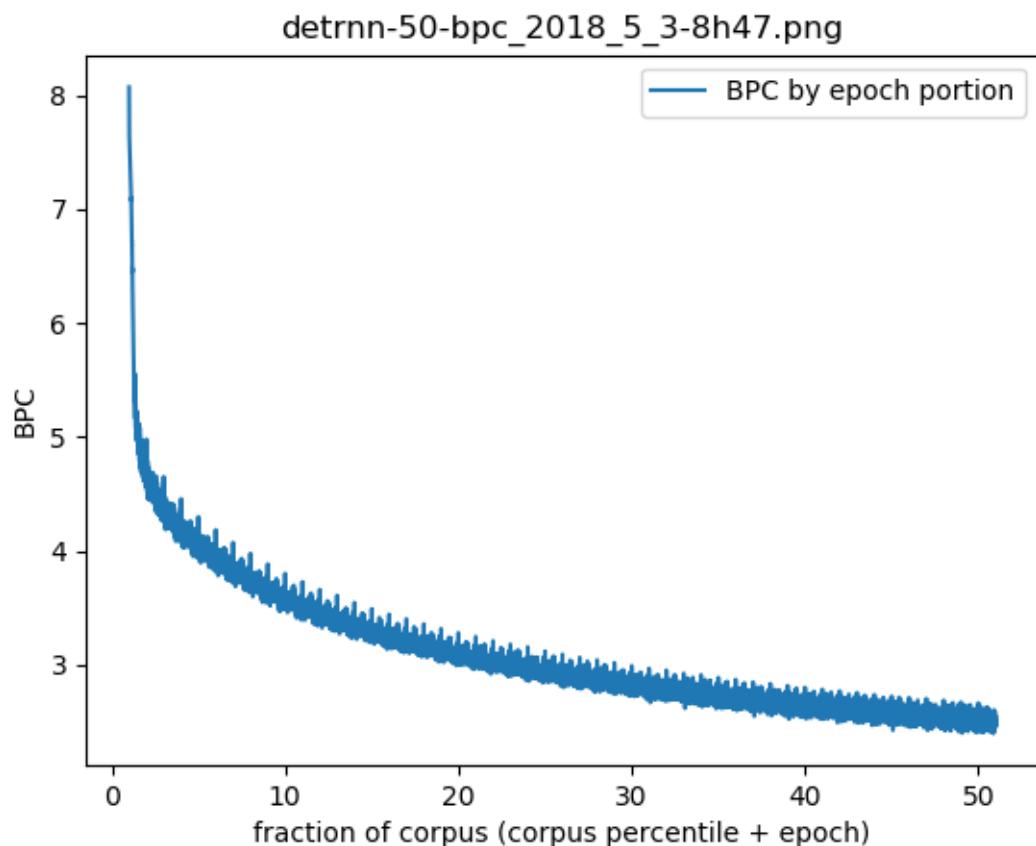


FIGURE C.2 – BPC

ValBPC/epoch Mean BPC over the epoch, at the end of each epoch.

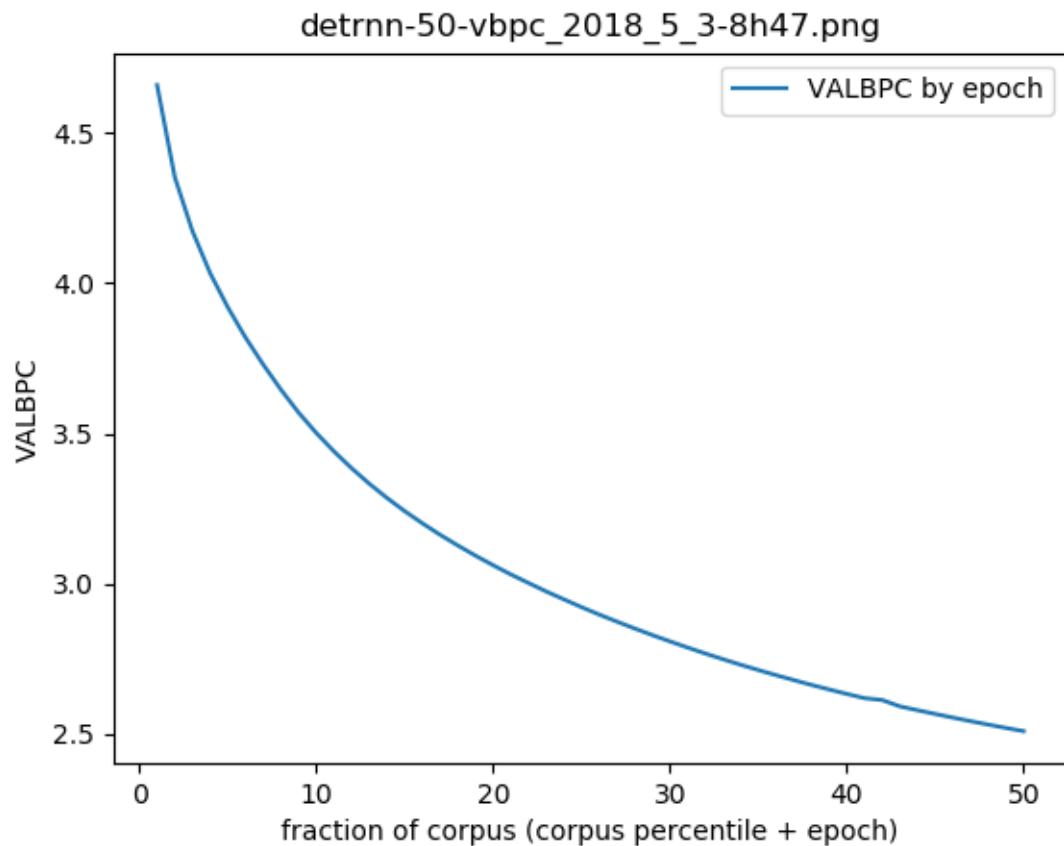


FIGURE C.3 – ValBPC

Loss Loss per fraction of the corpus (an interval of 1 correspond a complete corpus, or an epoch).

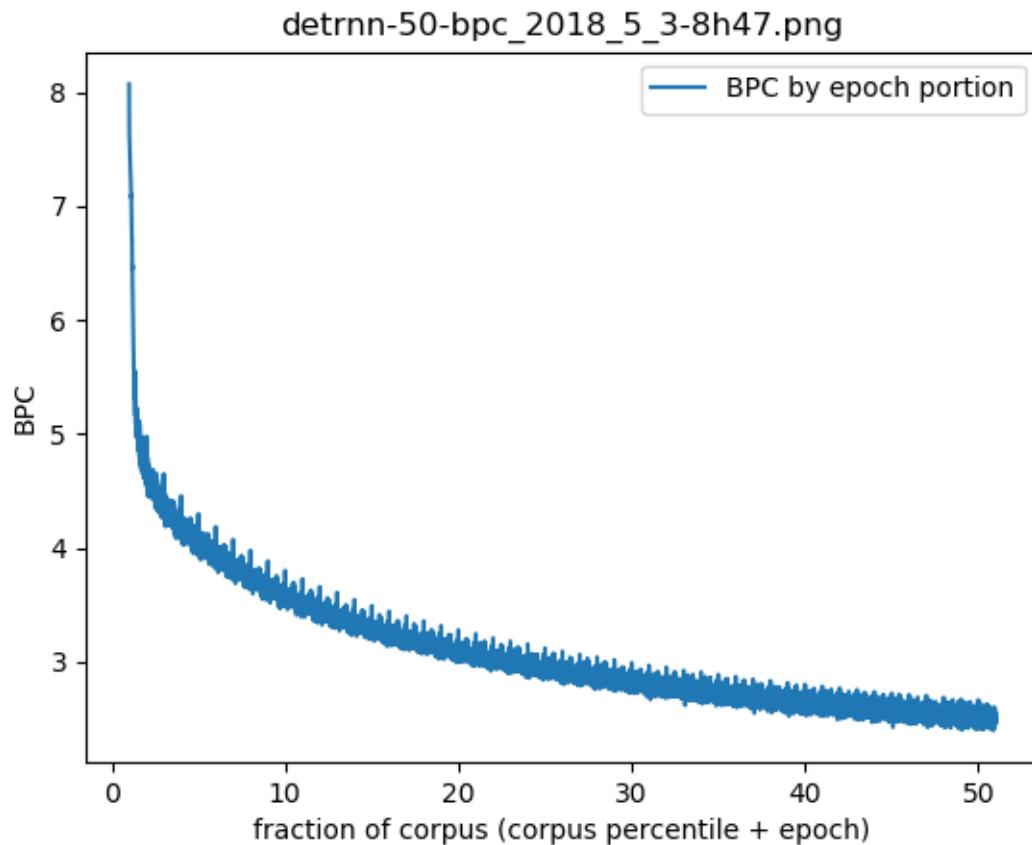


FIGURE C.4 – Loss

Logs

The log is available at https://gitlab.inria.fr/emarquer/awd-lstm-lm/blob/reimplement/logs/detrnn-50_2018_4_30-12h2.log.

C.10.4 Potential ameliorations & next steps

Next step is to test the growing model.

C.11 Premier test du modèle multi-échelles

Test run of `detrnn.py`

Test report

by E. Marquer, 2018/05/03, Synalp and Université de Lorraine

C.11.1 Abstract

First performance test of the Test to do a run on 4 epochs, with GPU, of the basic model of MSNN, with additive output strategy.

C.11.2 Paradigm

This test run of `detrnn.py`, with DEBUG level log output, loss per percentile and vbpc per epoch, that will be executed with `cuda`, for 4 epochs.

The test is done with branch `growing`, an allocated time of 4h, not interactive.

Run time was estimated for 4 epochs according to a debug results for 0.2 epochs :

1 (10 min / 0.2 epoch) * 4 epoch = 200 min = 3h 20min

With a security margin of 40min, run time is 4h.

/!\ Had to reduce batchsize down to 40 and halve hidden size because of memoryerrors /!

Hyperparameters

Hyperparameter	Value
nhidden	920
embedsize	400
bptt	200
batch_size	40
eval_batch_size	32
lr	0.001
wdecay	1.2e-06
cuda_on	True
log_interval	20
nepochs	4
max_seqs	15

Node

OAR_JOB_ID=1558426 with GPU

Job start time : 2018-05-04 14:43:40

Estimated job stop time : 2018-05-04 18:43:40

Command used :

```
1 oarsub -q production -p "GPU <> 'NO'" -l "nodes=1,walltime=04:00:00" ~/alt-
repo/awd-lstm-lm/rundet.sh
```

Status verification loop :

```
1 let x=0; while [ "true" ]; do echo "$x" $(oarstat -s -j 1558141); let ++x;
sleep 120; done
```

C.11.3 Results

Total run time for 4 epochs : with real stop time of 2018-05-03 17:57:26, the total run time of the training is approximately 3h 13min, corresponding to the predicted 50 min per epoch.

Comparative analysis

With half the number of hidden parameters, and a yet unknown number of layers, the basic MSNN has very similar results than the classical DetRNN.

However, when analysing closely the learning speed, the MSNN seems to be starting with a slower BPC decrease than the DetRNN, and it also seem to be faster later on. Those variations are probably due to the number of hidden layers.

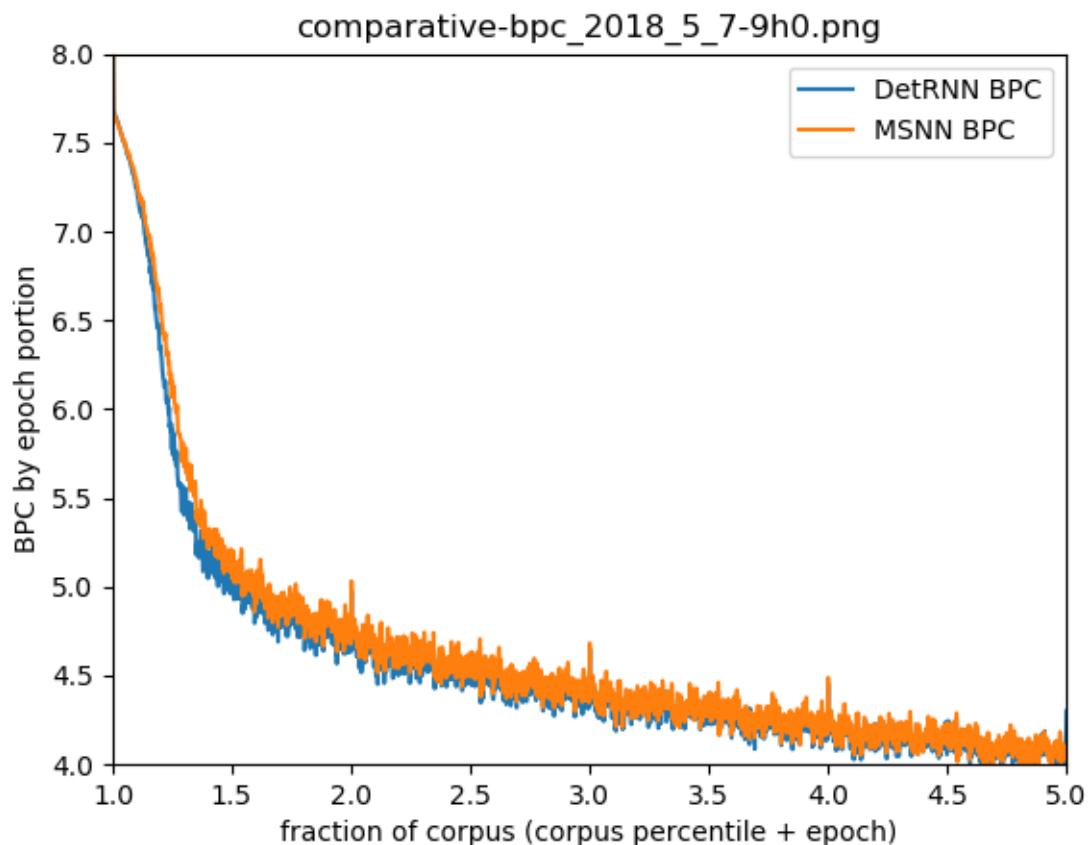


FIGURE C.5 – Comparative BPC

Plot

BPC/fraction of corpus BPS per fraction of the corpus (an interval of 1 correspond a complete corpus, or an epoch).

ValBPC/epoch Mean BPC over the epoch, at the end of each epoch.

Loss Loss per fraction of the corpus (an interval of 1 correspond a complete corpus, or an epoch).

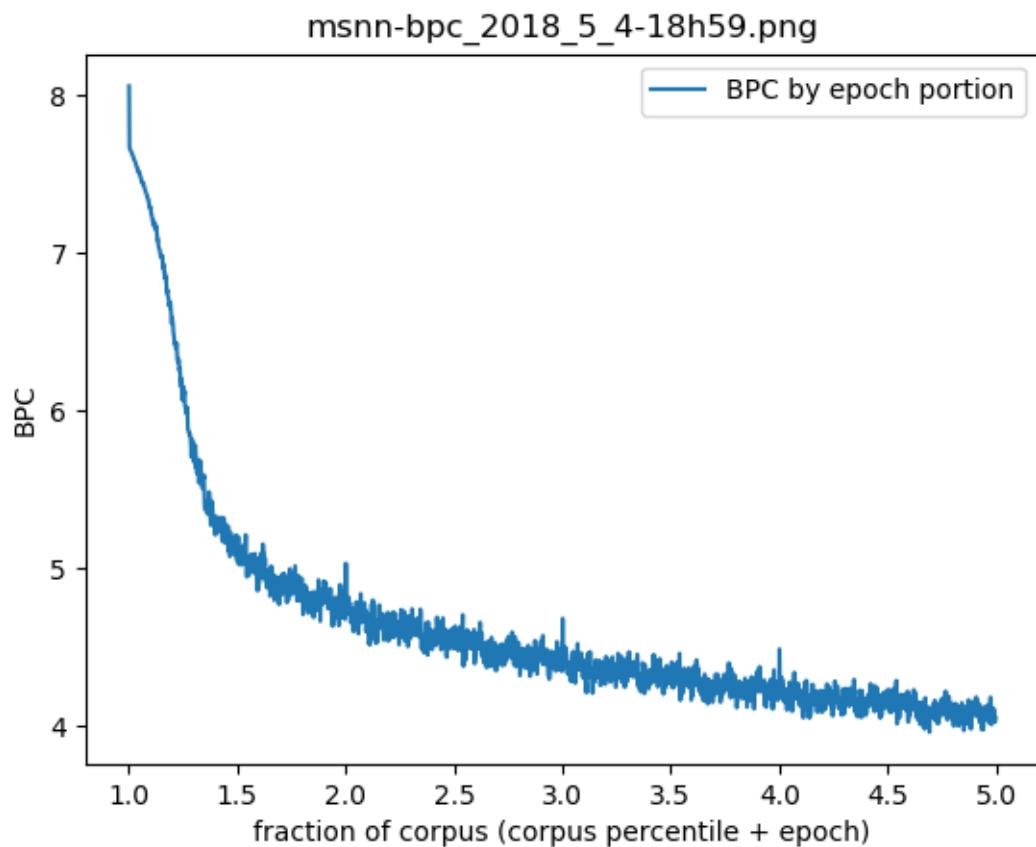


FIGURE C.6 – BPC

Logs

Reduced log is available at [msnn-base/msnn_2018_5_4-14h43.log](#).

C.11.4 Potential ameliorations & next steps

A necessary amelioration is to add a way to track the number of layers.

As of now, the upper hidden layers participates in the output only when updated. It is necessary to make them participate at every step.

The test process is not well defined : what to do when the eval batch is discontinued from the training batch ? what if it is in the same corpus, but not directly adjacent ? A possible yet hazardous solution would be to evaluate a “distance” between the training and evaluation batches, and reset the hidden states depending on that distance (a higher distance would reset a higher number of layers).

Lastly, as the number of values to remember is increasing (bpc, loss, layer number, ...) it would be interesting to improve the .plotdata system.

Next step is to test the recurrently defined growing model.

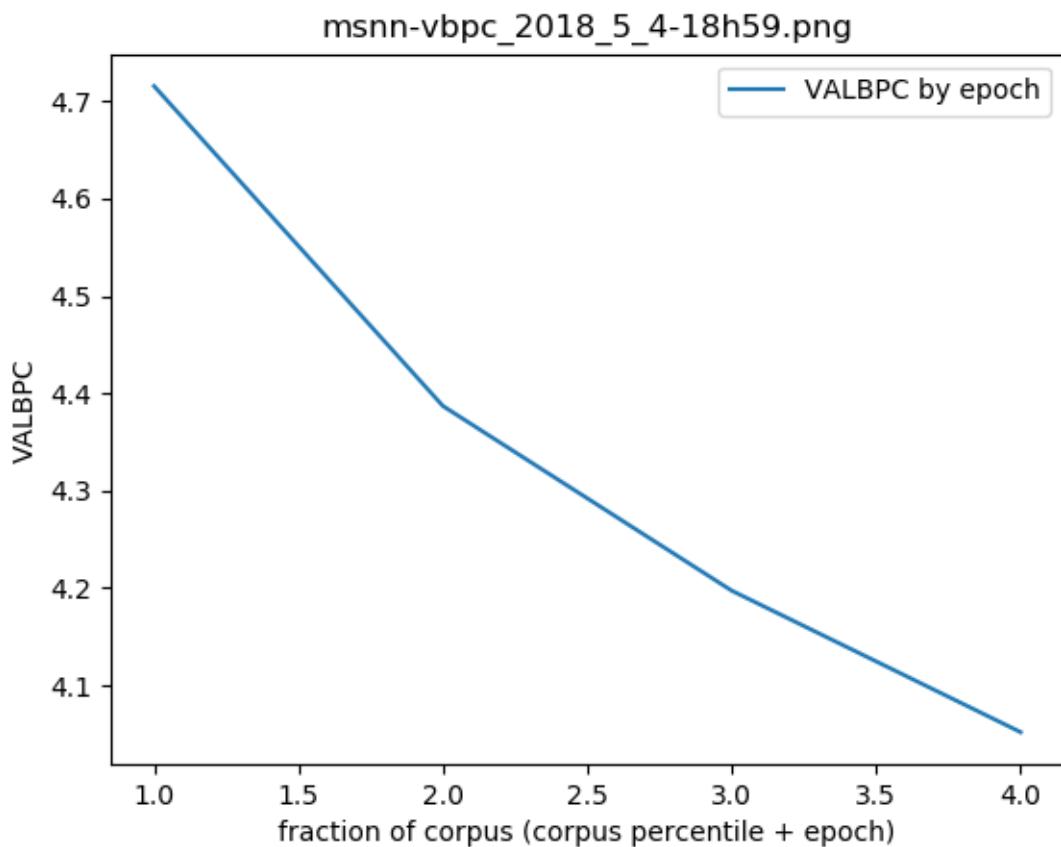


FIGURE C.7 – ValBPC

C.12 Comparaison des stratégies de fusion des résultats des différentes couches

Test run of `detmsnn.py`

Test report

by E. Marquer, 2018/05/16, Synalp and Université de Lorraine

C.12.1 Abstract

Performance test of the ‘cat’ (concatenated) output strategy compared to the ‘add’ strategy. Test to do a run on 4 epochs, with GPU, of the basic model of MSNN, with concatenated output strategy.

C.12.2 Paradigm

This test run of `detmsnn.py`, with INFO level log output, loss per percentile and vbpc per epoch, is executed with *cuda*, for 4 epochs.

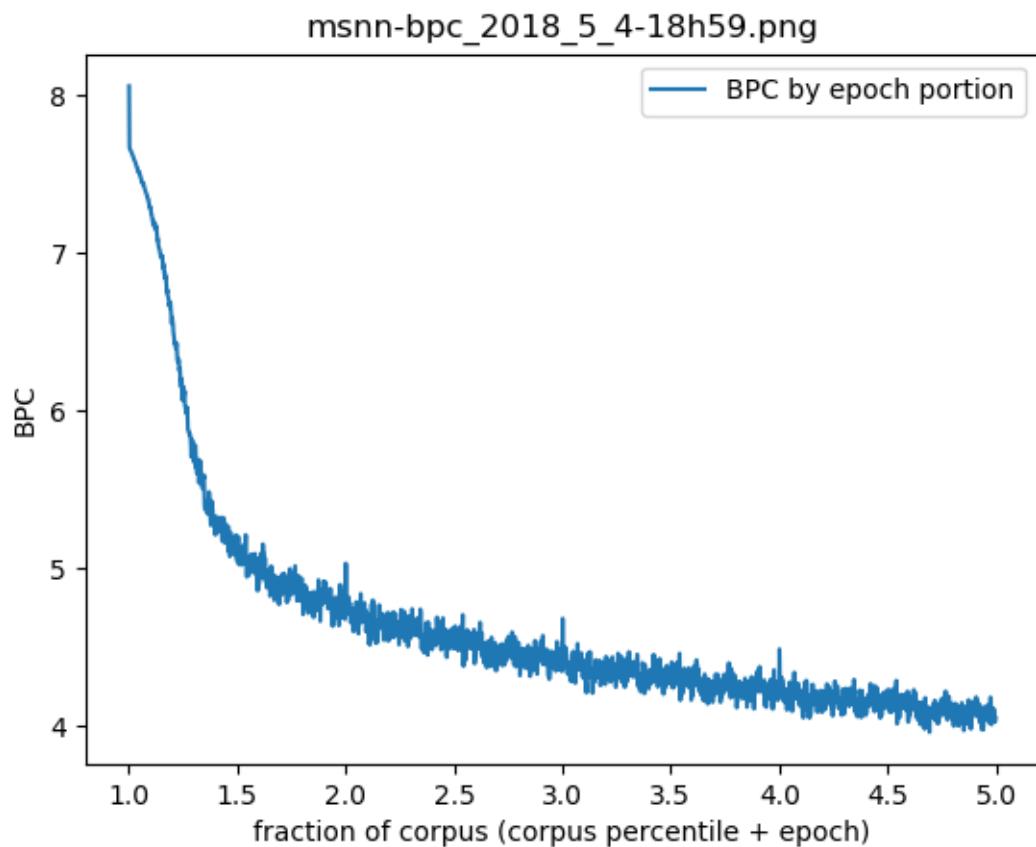


FIGURE C.8 – Loss

The test is done with branch **growing**, an allocated time of 24h, not interactive

/!\ Had to reduce evaluation corpus size down to 1/1000, to reduce computation time while keeping a big enough corpus to compute BPC /!

Hyperparameters

Hyperparameter	Value
nhidden	920
embedsize	400
bptt	200
batch_size	1
eval_batch_size	1
lr	0.001
wdecay	1.2e-06
cuda_on	True
log_interval	100
nepochs	4
max_seqs	5

Node

OAR_JOB_ID=1563805 with GPU grimani-1

Job start time : 2018-05-16 08 :53 :20

Estimated job stop time : 2018-05-17 08 :53 :20

Command used :

```
1 oarsub -q production -p "GPU <> 'NO'" -l "nodes=1,walltime=24:00:00" "bash runmsnn.sh"
```

Status verification loop :

```
1 let x=0; while [ "true" ]; do echo "$x" $(oarstat -s -j 1563805); let ++x; sleep 120; done
```

C.12.3 Results

Total run time for 4 epochs : with real stop time of 08 :53 :28, the total run time of the training is approximately 24h, with only 22% of one epoch done, and a final Validation BPC of 3.57.

Estimated run time for a full epoch : 24h / 22% \approx 109h/epoch. This corresponds to 436h for a 4 epoch run, and this is critical.

The ‘cat’ strategy is way more efficient in corpus consumption, even if it is dramatically slower than the additive strategy.

Comparative analysis

The comparative plot shows that with the ‘cat’ strategy, BPC diminution is way faster than with the additive strategy. Computation-time wise, it is obvious that the ‘cat’ strategy is slower, with ? ?h/epoch, than the ‘add’ strategy, with 50min/epoch. This difference is too large to be due to the device alone.

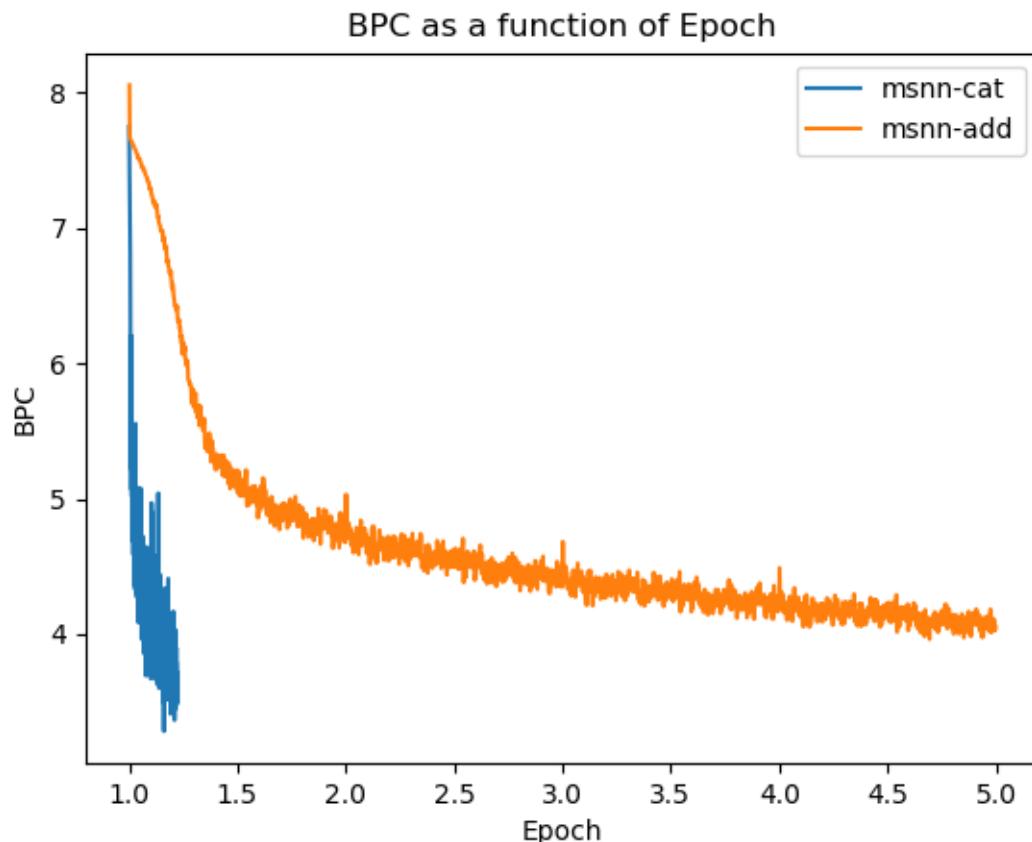


FIGURE C.9 – Comparative BPC

Plot

BPC/fraction of corpus BPC : BPC per fraction of the corpus (an interval of 1 correspond a complete corpus, or an epoch).

Validation BPC : BPC per fraction of the corpus, on the validation corpus.

Layers : Number of layers per fraction of the corpus.

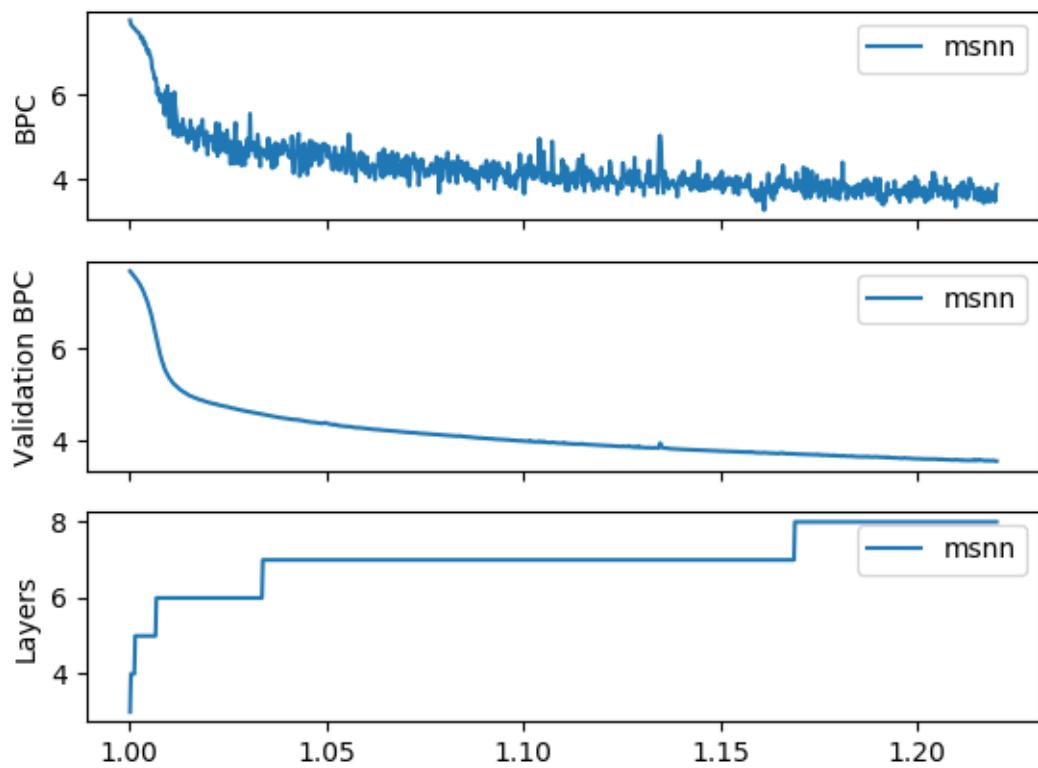


FIGURE C.10 – Comparative BPC

C.12.4 Potential ameliorations & next steps

Next step is to try to reduce run time.

C.13 Test des effets du changement de taille des paquets (batch)

Test run of detmsnn.py

Test report

by E. Marquer, 2018/05/16, Synalp and Université de Lorraine

C.13.1 Abstract

Performance test of batch size variation without deeper . Test to do a run on 4 epochs, with GPU, of the basic model of MSNN, with concatenated output strategy.

C.13.2 Paradigm

This test run of *detmsnn.py*, with INFO level log output, loss per percentile and vbpc per epoch, is executed with *cuda*, for 4 epochs.

The test is done with branch *growing*, an allocated time of 24h, not interactive

**/!\ Had to reduce evaluation corpus size down to 1/1000, to reduce computation time while keeping a big enough corpus to compute BPC /!\
!**

Hyperparameters

Hyperparameter	Value
nhidden	920
embedsize	400
bptt	200
batch_size	16
lr	0.001
wdecay	1.2e-06
cuda_on	True
log_interval	100
save_interval	100
nepochs	4
max_seqs	5

Node

OAR_JOB_ID=1567202 with GPU grimani-1

Planned job start time : 2018-05-19 02:41:08 Job start time : 2018-05-19 02:41:08

Estimated job stop time : 2018-05-22 14:41:08

Command used :

```
1 oarsub -q production -p "GPU <> 'NO' " -l "nodes=1,walltime=84:00:00" "bash  
runmsnn.sh"
```

Status verification loop :

```
1 let x=0; while [ "true" ]; do echo "$x" $(oarstat -s -j 1567202); let ++x;  
sleep 120; done
```

C.13.3 Results

Total run time for 4 epochs : with real stop time of ?, the total run time of the training is approximately ?h, with only, and a final Validation BPC of 3.57.

Comparative analysis

NO PLOT HERE

Plot

BPC/fraction of corpus BPC : BPC per fraction of the corpus (an interval of 1 correspond a complete corpus, or an epoch).

Validation BPC : BPC per fraction of the corpus, on the validation corpus.

Layers : Number of layers per fraction of the corpus.

NO PLOT HERE

C.13.4 Potential ameliorations & next steps

Next step is to continue run time reduction.

C.14 Entraînement sur le corpus complet avec beaucoup de temps alloué

Long-run of RNN-MSNN

Test report

by E. Marquer, 2018/05/29, Synalp and Université de Lorraine

C.14.1 Abstract

The run was done on the reduced enwik8 corpus.

The test is composed of 4 successive runs :

- 1 run of 2h on grimani-4;
- 2 runs of 12h, both on grimani-1;
- 1 run of 50h on grele-11;

End causes are as follow :

- Run 1 : out of time (2h);
- Run 2 : out of time (12h);
- Run 3 : end of epoch crash (7h30);
- Run 4 : end of epoch crash (19h15);

Mean time for an epoch is about 19h 15min (on the reduced version of the corpus). Two epochs were completed.

C.14.2 Results

Each run crashed between epochs, so a bit of patching had to be made on top of fixing the bug.

Memory

Both RAM and video RAM are still subject to a constant leak in memory. But even if it does not show on the plots (scale is too small), logs confirm that there is no leak during validation.

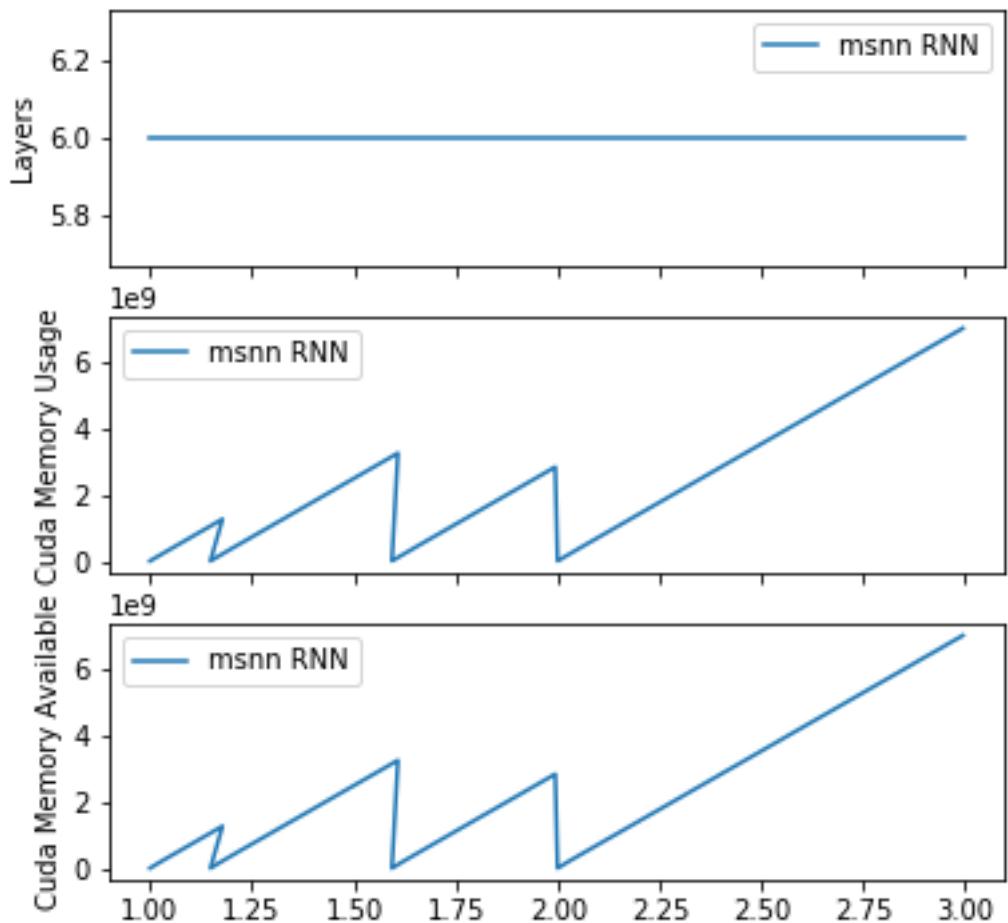


FIGURE C.11 – Memory usage

```
ida3/envs/pytorch/bin/python msnr_starter.py --save-folder logs/long-run_2018-07-06/ --cuda-on --resume-model logs/long-run_2018-07-06/models/.fullmodel
```

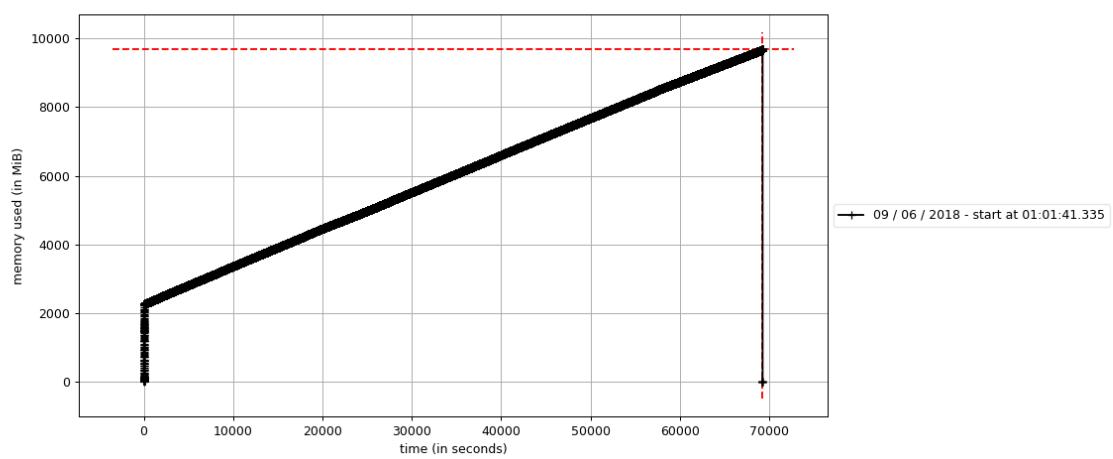


FIGURE C.12 – RAM third run

An other noticeable property is that “Run Time”, corresponding to the time to train over \log_{interval} sequences, is mostly proportional to CUDA memory usage. The source of the cuda memory leak is probably the same as what makes training slower.

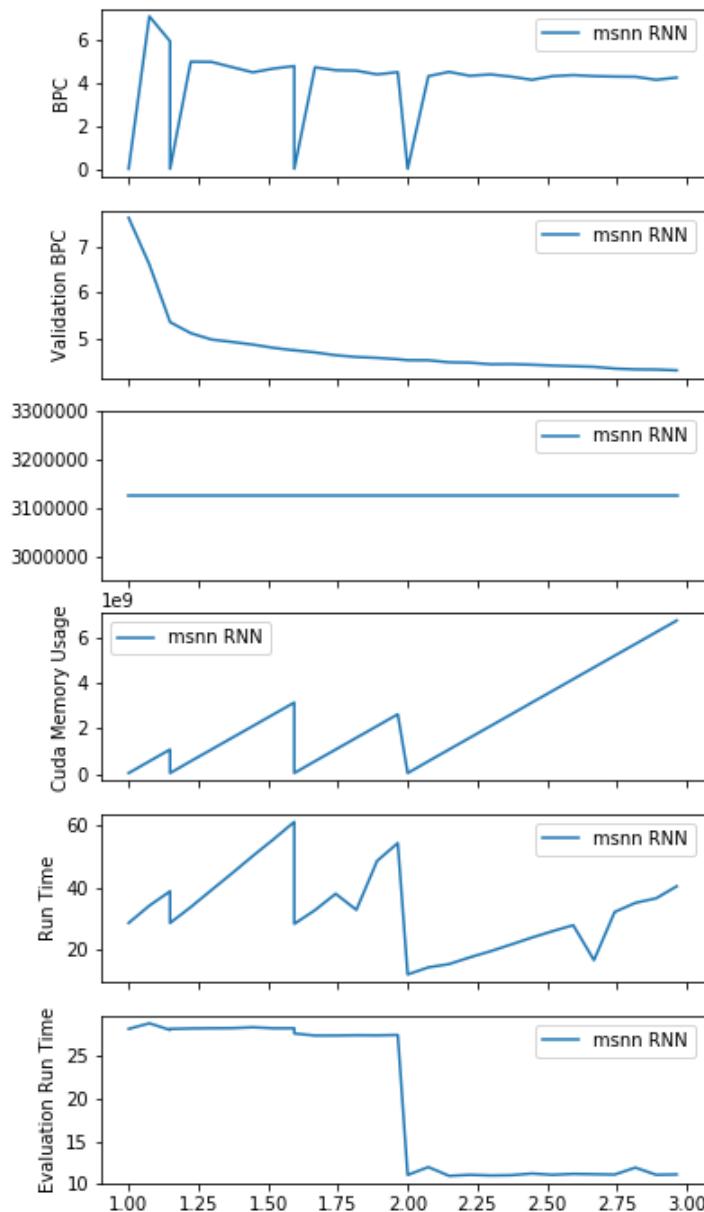


FIGURE C.13 – Memory and computation time

BPC / Validation BPC

BPC and Validation BPC

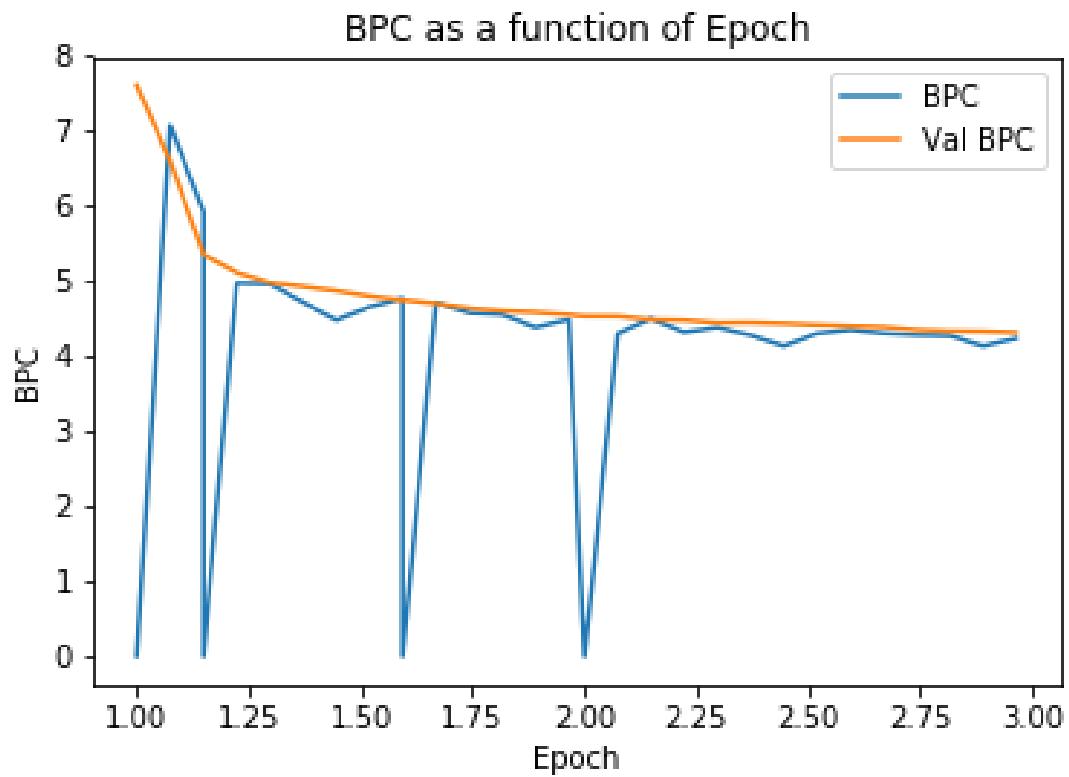


FIGURE C.14 – BPC

Job restart

When resuming a job, CUDA memory is entirely freed. Same thing can be said about RAM.

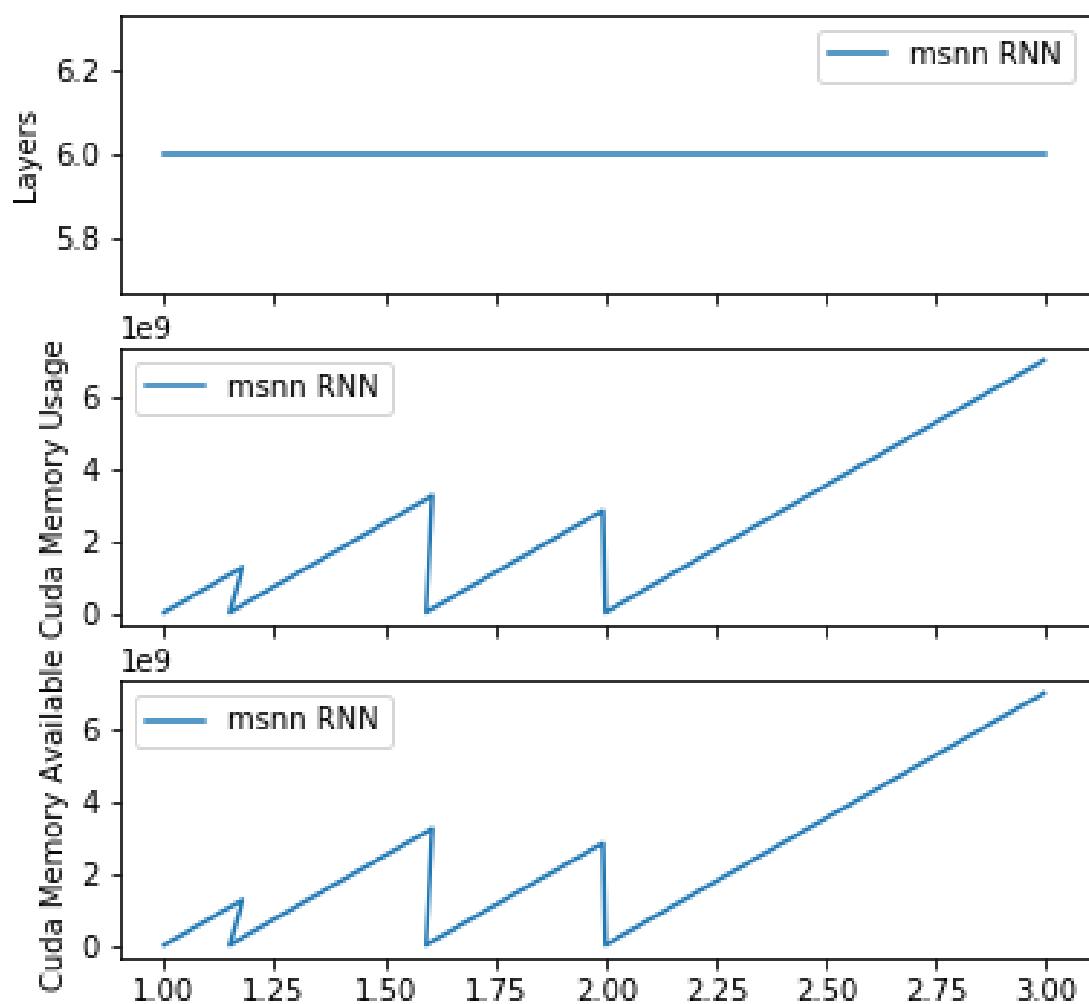


FIGURE C.15 – Memory usage

```
/home/emarquer/miniconda3/envs/pytorch/bin/python msnn_starter.py --save-folder logs/long-run_2018-07-06/ --cuda-on
```

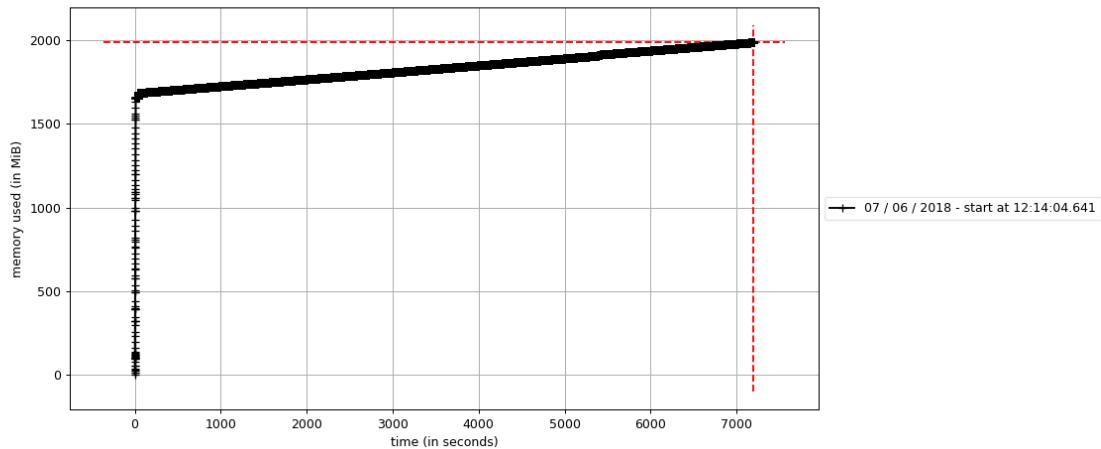


FIGURE C.16 – RAM first run

```
1da3/envs/pytorch/bin/python msnn_starter.py --save-folder logs/long-run_2018-07-06/ --cuda-on --resume-model logs/long-run_2018-07-06/models/.fullmodel
```

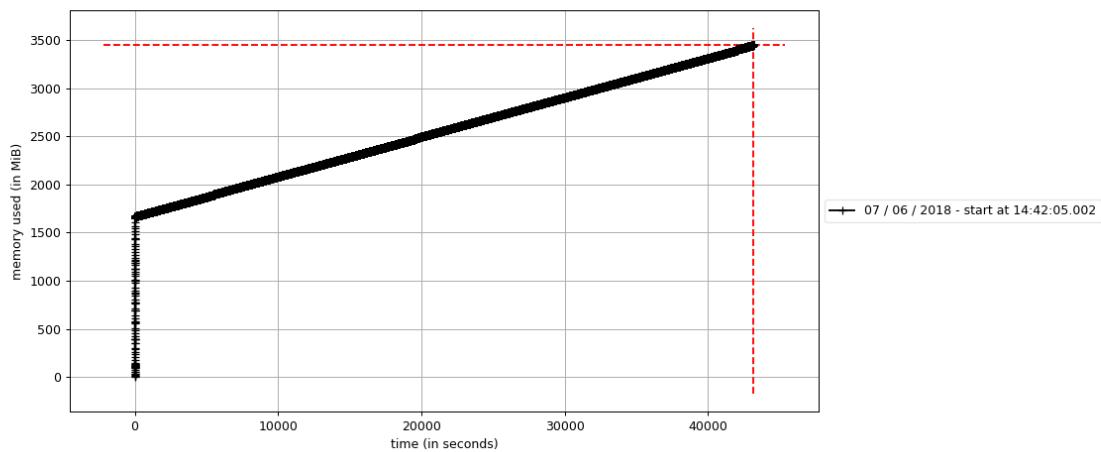


FIGURE C.17 – RAM second run

```
1da3/envs/pytorch/bin/python msnn_starter.py --save-folder logs/long-run_2018-07-06/ --cuda-on --resume-model logs/long-run_2018-07-06/models/.fullmodel
```

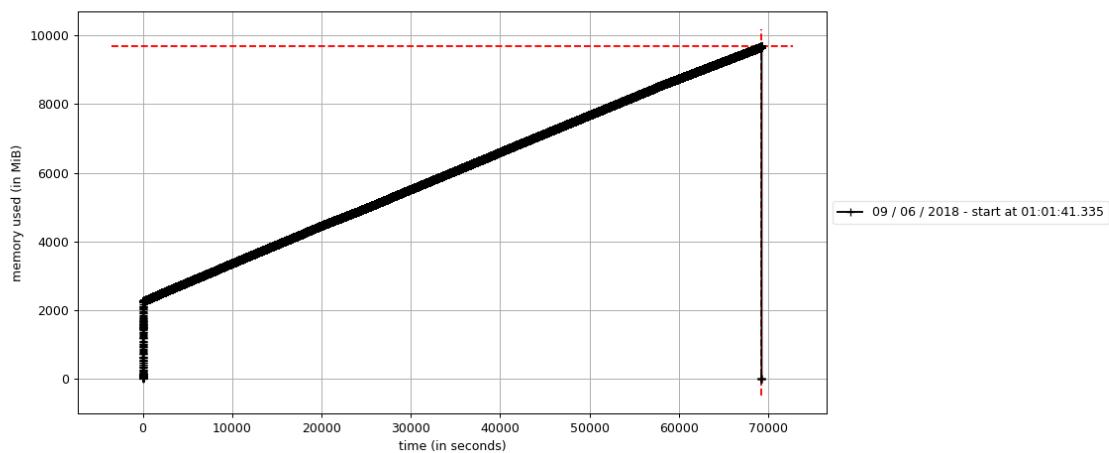


FIGURE C.18 – RAM third run

Memory is freed each time the job is restarted, meaning either a part of the necessary is discarded, or unnecessary data is kept in memory. As in CUDAles tests a memory maximum was reached, CUDA seems to be the source of the leak (data copies not removed, ...).

C.14.3 Next steps

Debug end of epoch bug. Try to patch memory leak. Continue training.

C.15 Changement de stratégie de gestion d'historique

Change history

Analysis report

by E. Marquer, 2018/06/12, Synalp and Université de Lorraine

C.15.1 Abstract

As computation graph has been confirmed to exist/be considered only in the last history, keeping an explicit history is no longer necessary. Worst, it may hinder garbage collection, by keeping references to Tensors that are completely useless.

A new model of history has been made to keep only a reference to the last hidden state. This new model will be referred as “last”, and the old one keeping an explicit history as “classical”.

C.15.2 Tests

Multiple tests were done on the new “enwik8mini” corpus of 100,000 characters :

- “last” a test run on 2 epochs with 1 batches of “last” history
- “last b2” a test run on 10 epochs with 2 batches of “last” history
- “last b2 10epoch” a test run on 10 epochs with 2 batches of “last” history
- “classical” a test run on 2 epochs with 1 batches of “classical” history

Results (1 epoch)

The results are over 1 epoch, with values from the start of the first and the second epoch. As “last b2 10epoch” and “last b2” have the same set of parameters, they coincides.

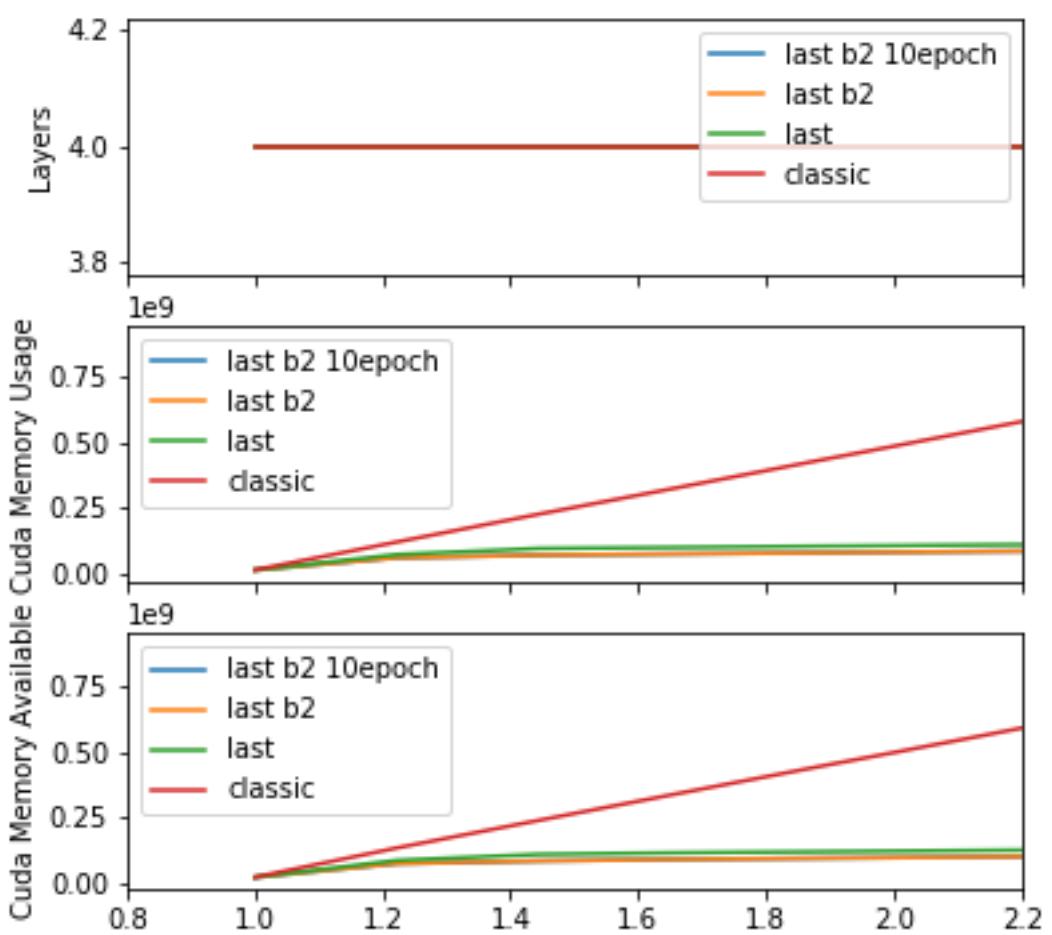


FIGURE C.19 – Memory 1 epoch

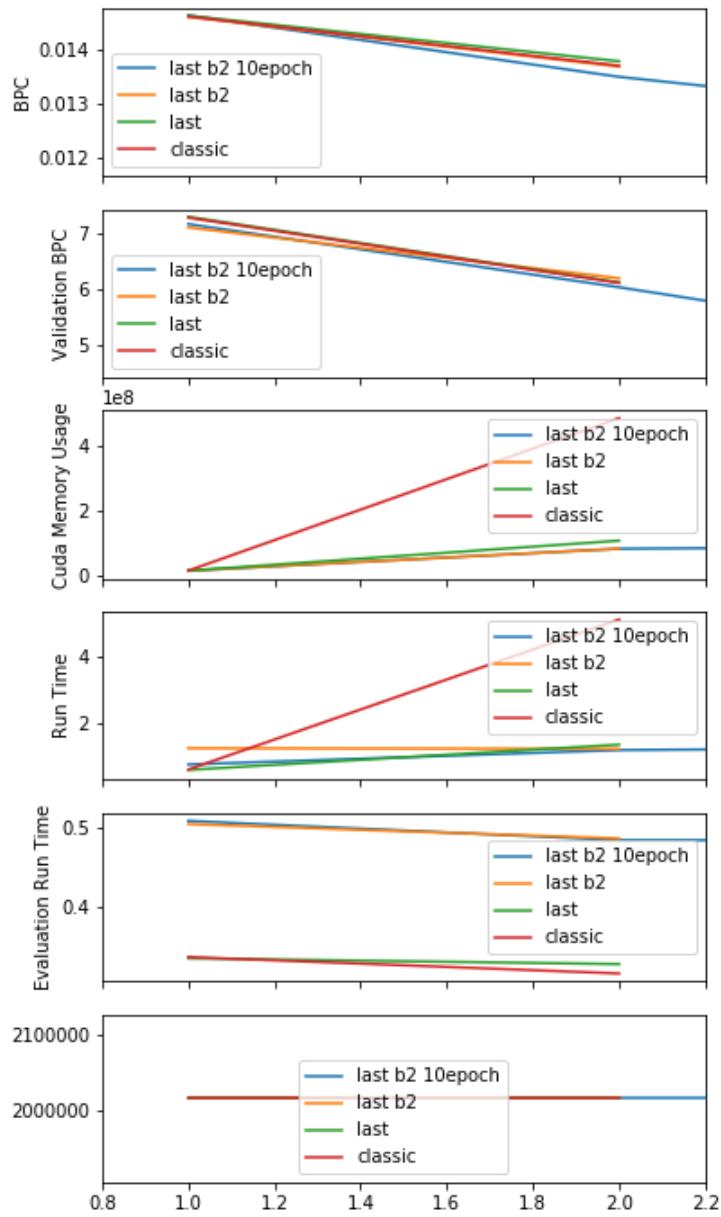


FIGURE C.20 – All info 1 epoch

Results (full)

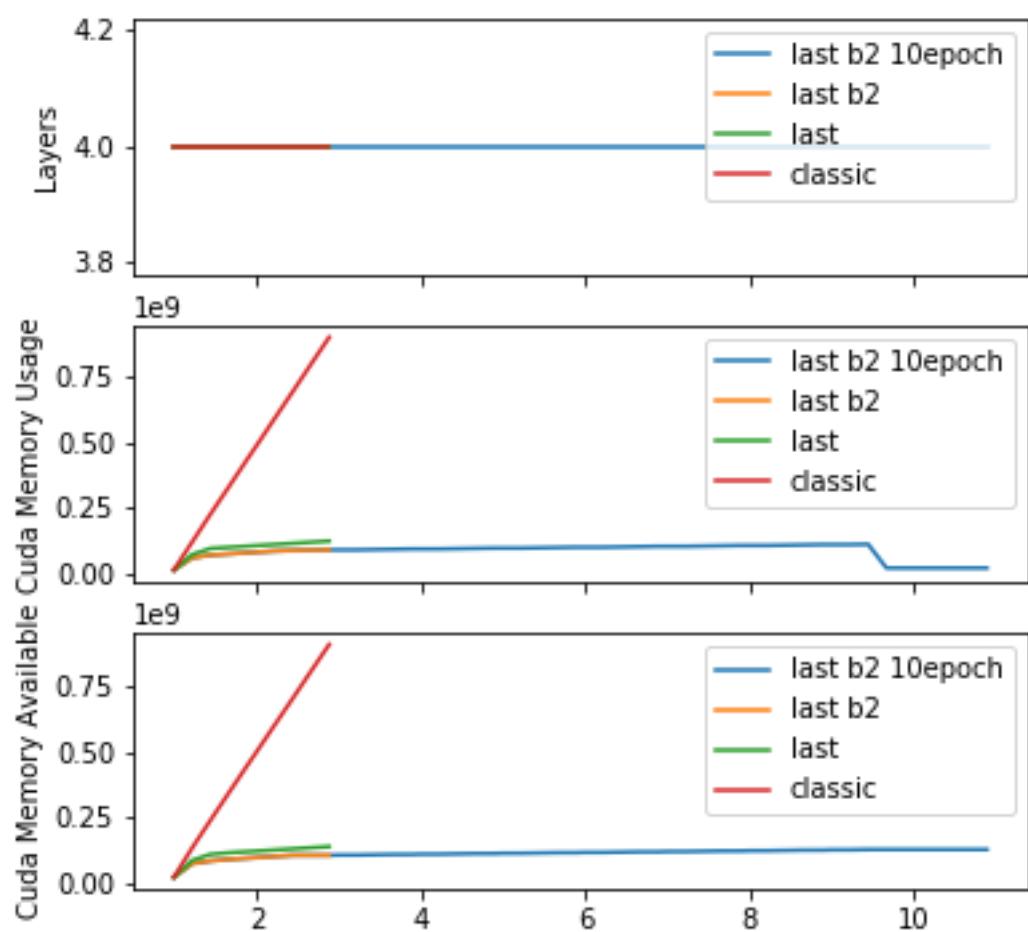


FIGURE C.21 – Memory

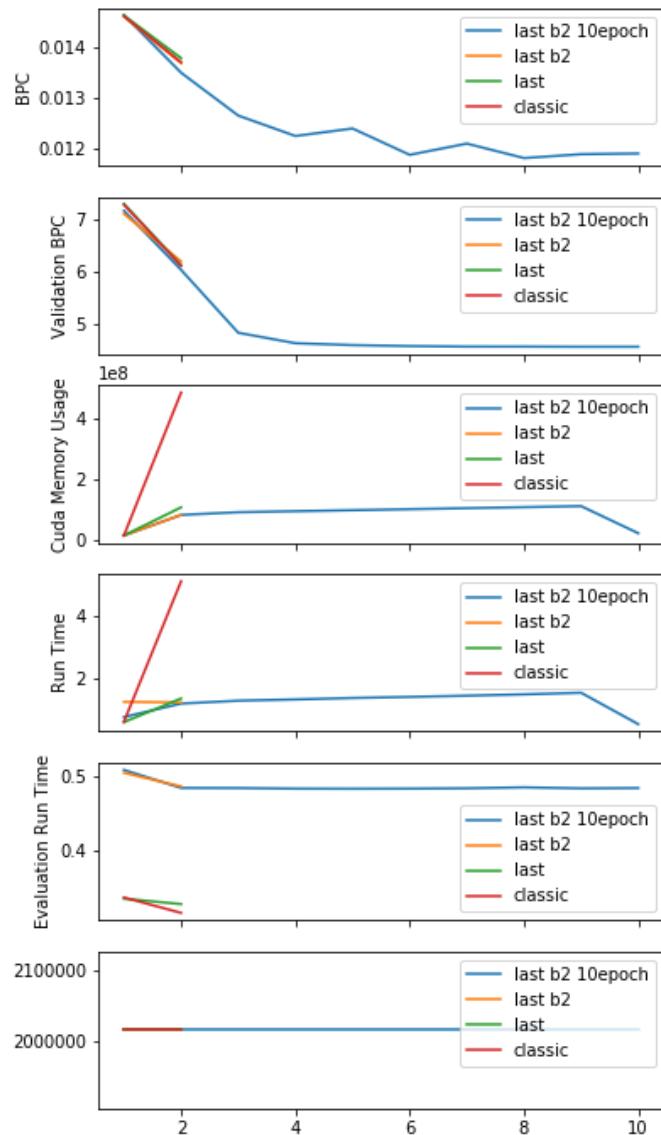


FIGURE C.22 – All info

```
quer/miniconda3/envs/pytorch/bin/python msnn_starter.py --save-folder logs/test4/ --cuda-on --corpus data/enwik8mini --batch-size 2 --epochs 10
```

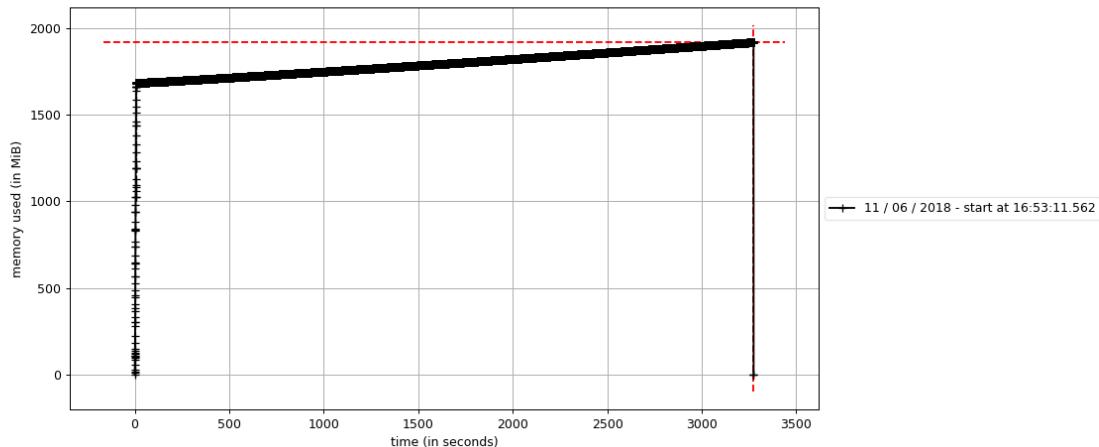


FIGURE C.23 – RAM 10 epoch

Ram consumption for “last b2 10epoch”

C.15.3 Conclusion

Memory leak

The leak in CUDA memory has reduced drastically.

It seems there is still a leak, this leak is not present when CUDA is not used. This leak is present in both RAM and graphical RAM, but the remaining leak in graphical RAM is not a concern anymore, with only a dozen MiB over 1,000,000 characters (10 epochs * 100,000 characters). However, the leak in RAM has not reduced, and even if it is not a problem thanks to the available RAM in the cluster, it would be preferable to identify the source of the leak.

Run (training) time

Additionally, the correlation of CUDA memory usage and training run time is kept, so the small remaining CUDA leak is probably linked to the computation graph. With CUDA memory usage drastically reduced, run time has been reduced too.

Currently, training over 100,000 characters (with 2 batches) only takes 5 minutes.

Performances

No conclusion can be drawn on learning performances, due to the size of the learning corpus. Even though, it is encouraging that even with a small corpus 10 epochs are not enough to have over-fitting (at least with 2 batches).

C.15.4 Next steps

1. long training over the “enwik8reduced” or the “enwik8” corpus;
2. fix (or at least identify) RAM leak

C.16 Test de l'implémentation des *batchs*

Long-run of RNN-MSNN

Test report

by E. Marquer, 2018/06/13 Synalp and Université de Lorraine

C.16.1 Abstract

The test is composed of 2 successive runs :

- id 1582586 : batch-size 1, bptt 200 on grele-4;
- id 1582587 : batch-size 2, bptt 100 on grimani-6.

Run time is about 10h in each case, corresponding to 2h30 for an epoch. In both case corpus batches rotation over epochs was disabled.

Shared parameters

parameter	value
corpus	enwik8reduced
history_strategy	layer-constant-length
max_history	25
bptt	<i>variable</i>
batch_size	<i>variable</i>
epochs	4
lr	1e-3
weight_decay	1.2e-6
epoch	4
valid_len	500,000
log_interval	500
save_interval	500
memory_interval	100
hidden_size	460
embed_size	400
growth_factor	5
rnn_type	RNN
reset_hidden	False

parameter	value
reset_growth	True
cuda_on	True

C.16.2 Results

At the end of each epoch, we see a spike in BPC, due to the first evaluation of the epoch. As corpus rotation is disabled, it is not surprising that with two batches over-fitting appears.

Moreover, we can note that run time is constant and memory usage tend to a constant value (1.6 GiB), with no difference between 1 and 2 batches. Note : the product bptt * batch_size is equal for 1 and 2 batches, so this result is the one predicted by the equations.

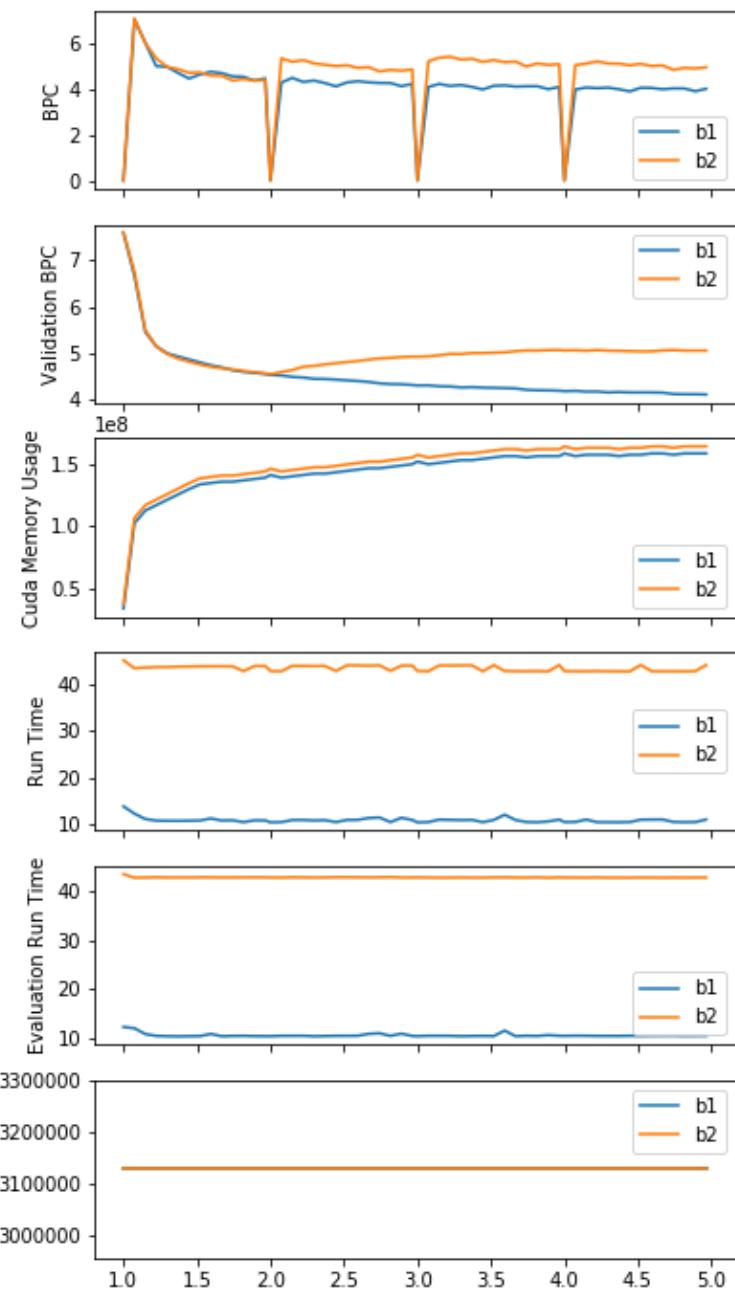


FIGURE C.24 – RAM third run

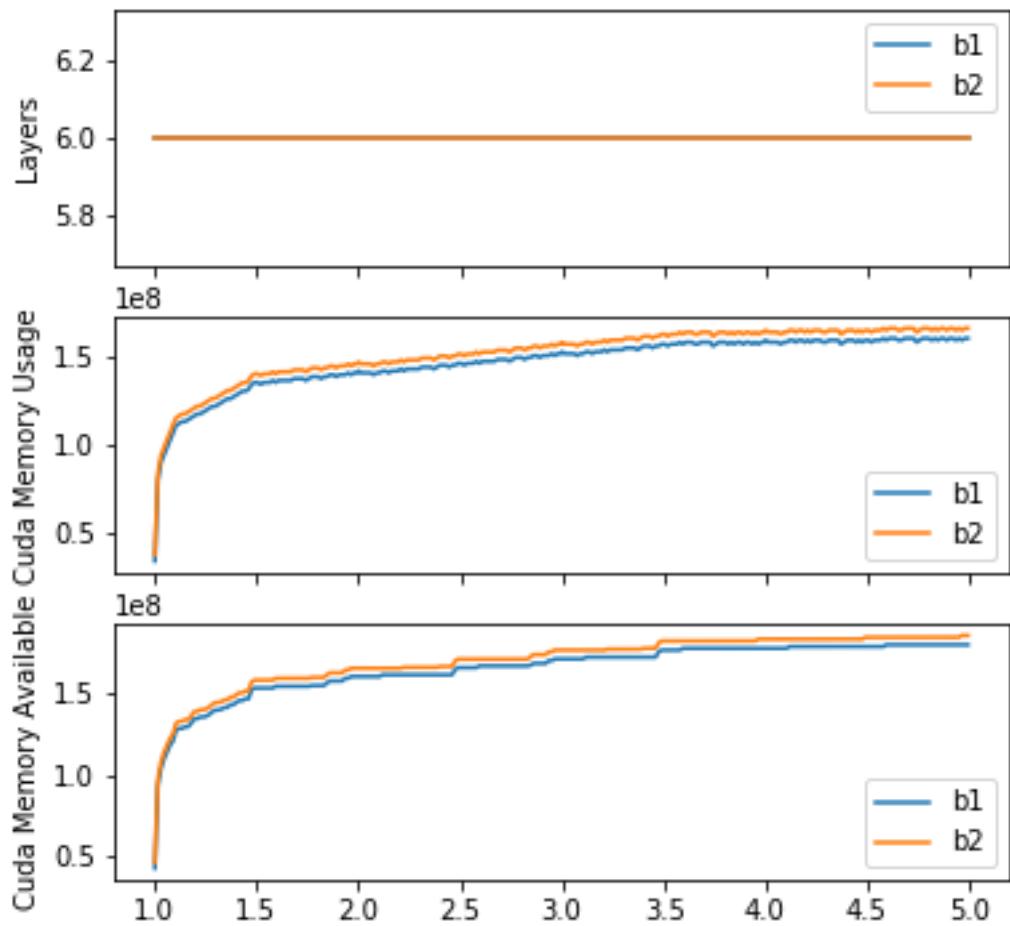


FIGURE C.25 – Memory usage

C.16.3 Next steps

- batches :
 1. Implement corpus rotation
 2. See if corpus rotation solves over-fitting
 3. Compare run time on comparable machines
- long run :
 1. Continue batch 1 for more epochs
 2. See when over-fitting appears

C.17 Test des performances des *batchs*

Run on the same node of RNN-MSNN with different batch size

Test report

by E. Marquer, 2018/06/15, Synalp and Université de Lorraine

C.17.1 Abstract

The test is composed of 2 runs :

- id 1583339 : batch-size 1, bptt 200 on grele-2
- id 1583336 : batch-size 2, bptt 100 on grele-1

Run time per epoch varies from 28 to 20 min for a single batch and from 20 to 11 min for two batches.

Shared parameters

parameter	value
corpus	enwik8reduced
max_history	25
bptt	<i>variable</i>
batch_size	<i>variable</i>
epochs	4
lr	1e-3
weight_decay	1.2e-6
epochs	4
valid_len	500,000
log_interval	500
save_interval	500
memory_interval	100
hidden_size	460
embed_size	400
growth_factor	5
rnn_type	RNN
reset_hidden	False

parameter	value
reset_growth	True
cuda_on	True

C.17.2 Results

At the end of each epoch, we see a spike in BPC, due to the first evaluation of the epoch (BPC is reinitialised to 0, causing a spike). With 2 batches, BPC is also more stable.

As of now, run time is computed after running validation, so training time is `real_training_time = run_time - validation_time`. But validation time increase with the number of batches as the number of characters seen increase (the whole validation corpus is used for each batch, so the number of characters seen during validation is `validation_corpus_len * batches`). That explains that the evaluation time for a batch is lower than for 2 batches.

By removing evaluation time, we can obtain a reasonable and coherent (with regard to epoch run time) estimation of training time over a sequence.

Epoch run time :

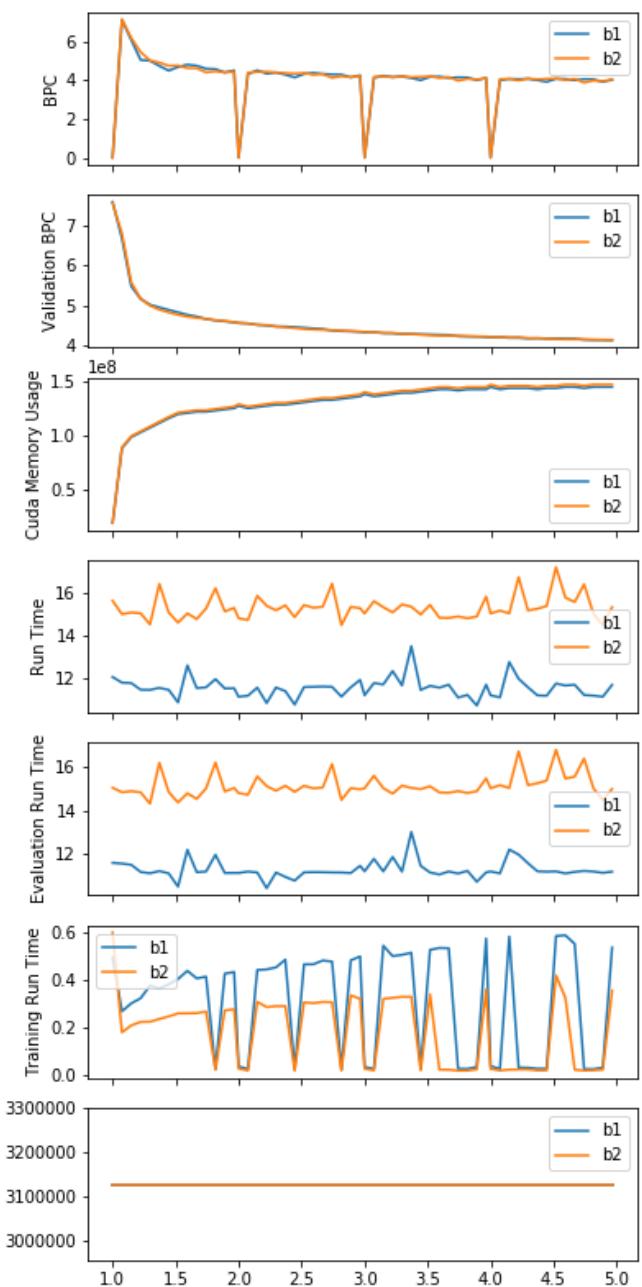
Epoch	Run time b=1	Run time b=2
1	28 min	20 min
2	23 min	17 min
3	22 min	14 min
4	20 min	11 min

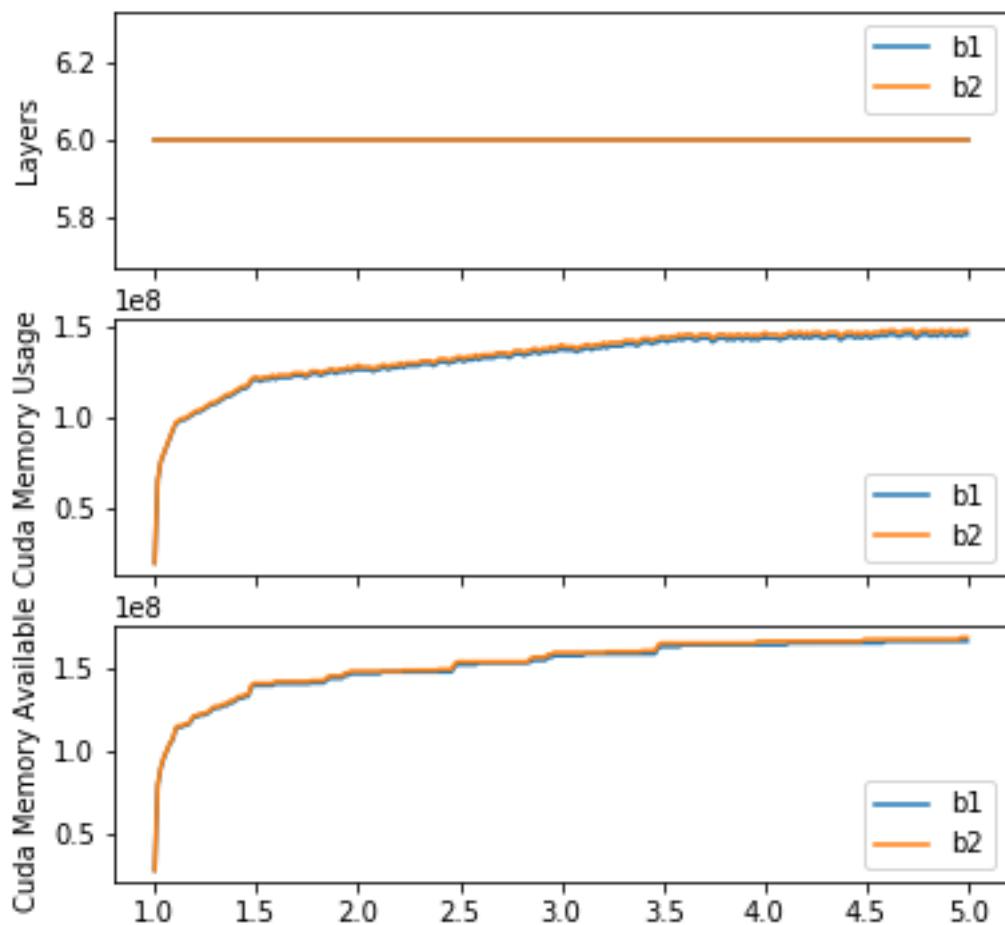
There is a notable decrease in epoch run time (time necessary to run over an epoch), of about 10 min over the 4 epochs, with both runs.

Possible causes for the decrease of run time : - part of the graph is already computed, and this part is skipped; - corpus data is already loaded in cuda memory.

Both of them are highly unlikely to cause such a decrease. A more precise training time storage may provide an explanation.

Plots





C.17.3 Next steps

- Data
 - Implement more precise time data saving (epoch run time and training run time)
 - Runs (objective : 100 epochs)
 - Continue running batches 1 and 2
 - Try with higher number of batches
 - [Optional] Other experimental branch
1. Implement prepared attention module in every layer
 2. Analyse results
 3. Patch probable memory leaks

C.18 Test des performances des *batchs* sur 50 époques

Run over more than 50 epochs with varied batch size

Test report

by E. Marquer, 2018/06/19, Synalp and Université de Lorraine

C.18.1 Abstract

The test is composed of 4 runs on grele, with :

- bptt 200, batch-size 1
- bptt 100, batch-size 2
- bptt 50, batch-size 4
- bptt 25, batch-size 8

Run has been interrupted by an overflow of disk space due to the detailed logs ; next runs will used reduced logs.

Run time per epoch varies from 30 to 7 min.

Shared parameters

parameter	value
corpus	enwik8reduced
history_strategy	layer-constant-length
max_history	25
bptt	<i>variable</i>
batch_size	<i>variable</i>
epochs	4
lr	1e-3
weight_decay	1.2e-6
epoch	4
valid_len	500,000
log_interval	500
save_interval	500
memory_interval	100
hidden_size	460
embed_size	400

parameter	value
growth_factor	5
rnn_type	RNN
reset_hidden	False
reset_growth	True
cuda_on	True

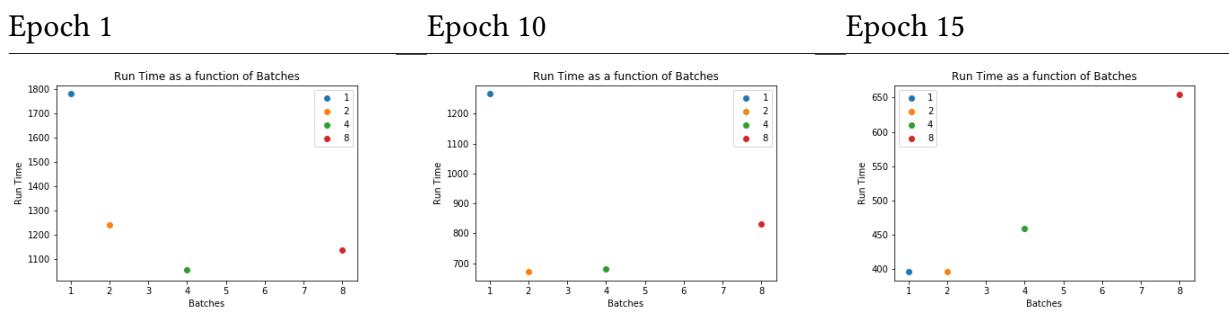
C.18.2 Results

At the end of each epoch, we see a spike in BPC, due to the first evaluation of the epoch (BPC is reinitialised to 0, causing a spike). With any number of batch, while keeping the `bptt * batch_size` ratio, BPC and Validation BPC do not vary.

Even with 200 epochs, with batch size of 1, there is no trace of over-fitting.

Epoch run time :

Epoch	Run time b=1	Run time b=2	Run time b=4	Run time b=8
1	30 min	21 min	18 min	19 min
10	21 min	14 min	14 min	17 min
15	7 min	7 min	8 min	11 min



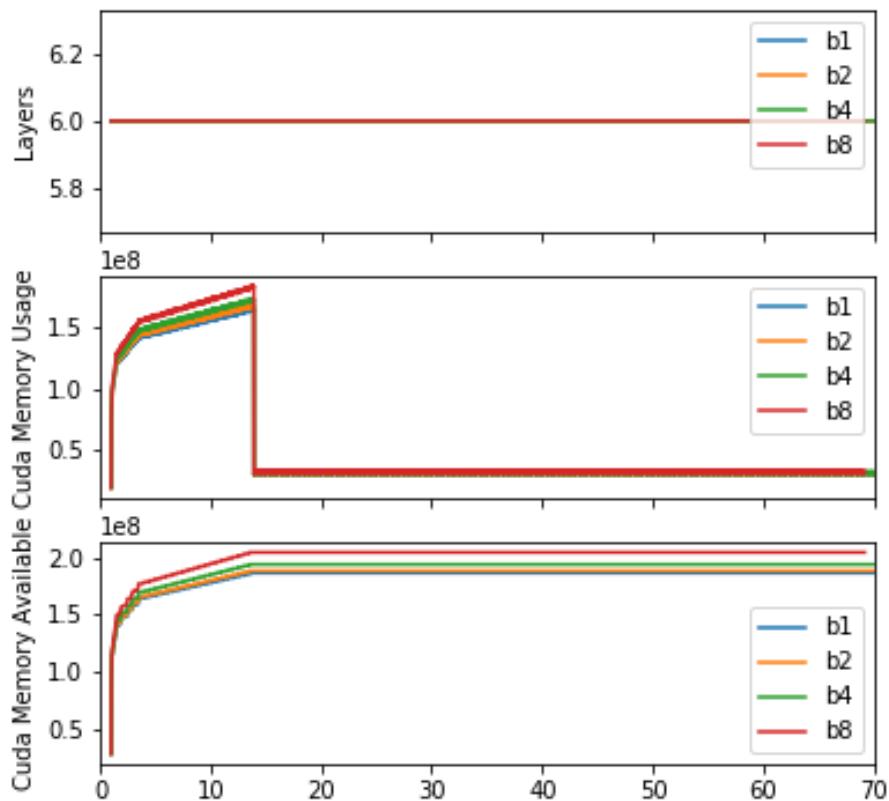
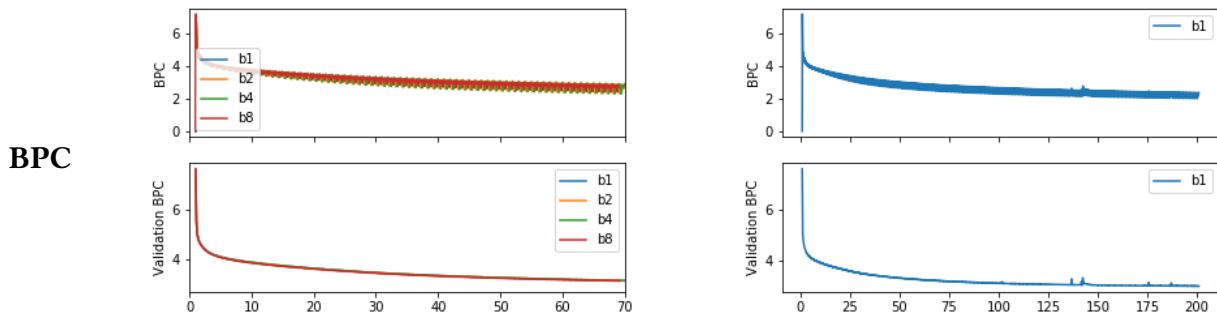
Run time can be split over two set of epochs : before, and after the 15th epoch. Before the 15th epoch, run is faster with more batches, with the exception of batch-size 8, which is slower than batch-size 2 and 4. After the 15th epoch, run is slower with more batches.

To optimize run time, it is necessary to balance run time before 15 epochs. If a high number of batches is planned (50), between 2 and 4 batches are preferable, because the run time after epoch 15 has a lot of impact on global run time; however, if only a few epochs are planned (20), a high number of batches is preferable, as run time before epoch 15 is the most important.

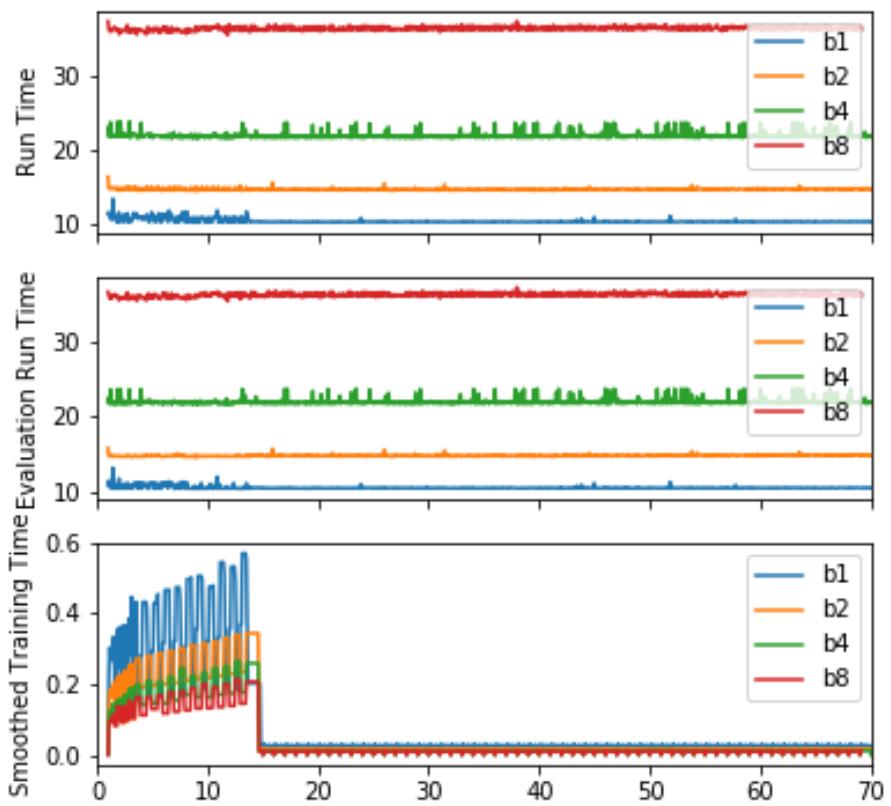
With current corpus, 2, 3 or 4 batches are the most interesting setup.

Decrease in run time is most probably due to the history of the upper layers, that needs multiple epochs to fill.

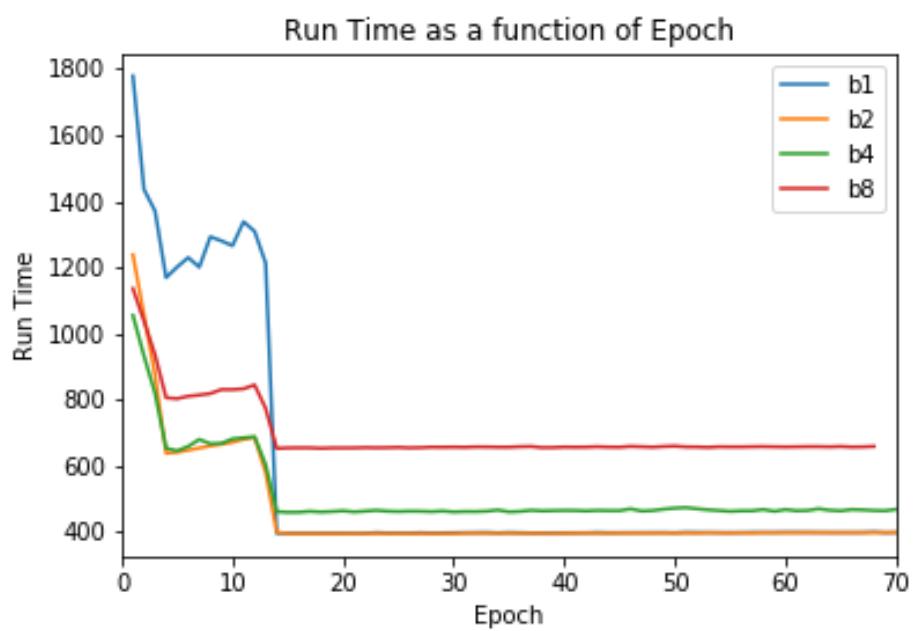
Plots



Memory



Run time (in seconds)



Epoch run time

C.18.3 Next steps

- Independent impact of batch number and sequence length :
 - Run with fixed batch number and varying sequence length
 - Run with fixed sequence length and varying batch number
- Impact of corpus length on optimal number of batch and sequences :
 - Run with varying sequence length and varying batch number over the full length corpus
- Number of parameters :
 - Increase the hidden layer size
- [Future] Transmission rate impact :
 - Compare optimal values of parameters, and BPC reached with varying transmission rate

C.19 Test des performances des différentes améliorations

Run over more than 50 epochs with varied batch size

Test report

by E. Marquer, 2018/06/25, Synalp and Université de Lorraine

C.19.1 Abstract

The test runs are for two parallel experiments :

- test which of 2 and 3 batches are the most interesting
- test the impact of layer by layer training (with an intuitive algorithm developed out of work-time)

The test is composed of 4 runs on grele, with :

- bptt 200/2, batch-size 2
- bptt 200/3, batch-size 3
- bptt 200/2, batch-size 2, layer by layer training,
- bptt 200/2, batch-size 2, layer by layer training, 20 epochs of individual training for each layer, 10 epochs of common fine-tuning for already trained layers (see below the explanation of this algorithm)

Shared parameters

parameter	value
corpus	enwik8reduced
history_strategy	last
max_history	25
lr	1e-3
weight_decay	1.2e-6
epochs	1500
valid_len	500,000
log_interval	500
save_interval	500
memory_interval	100
hidden_size	460
embed_size	400
growth_factor	5
rnn_type	RNN

parameter	value
reset_hidden	False
reset_growth	True
cuda_on	True

Model keys and specificities

When noting is specified, all models have :

- 2 batches, and a sequence length of 200/2
- 1 RNN layer per MSNN layer

Model	Specificity
b2	classical model, comparison basis
b3	3 batches, sequence length of 200/3
s	“scheduled(10,10)” : use layer by layer training; 10 epochs for individual training, 10 epochs for intermediary fine-tuning
s-a20-l10	“scheduled(20,10)” : use layer by layer training; 20 epochs for individual training, 10 epochs for intermediary fine-tuning
l2	2 RNN layer per MSNN layer
l3	3 RNN layer per MSNN layer
s_l3_a	3 RNN layer per MSNN layer, “scheduled(10,10)” (see model “s”), attentive intermediary input

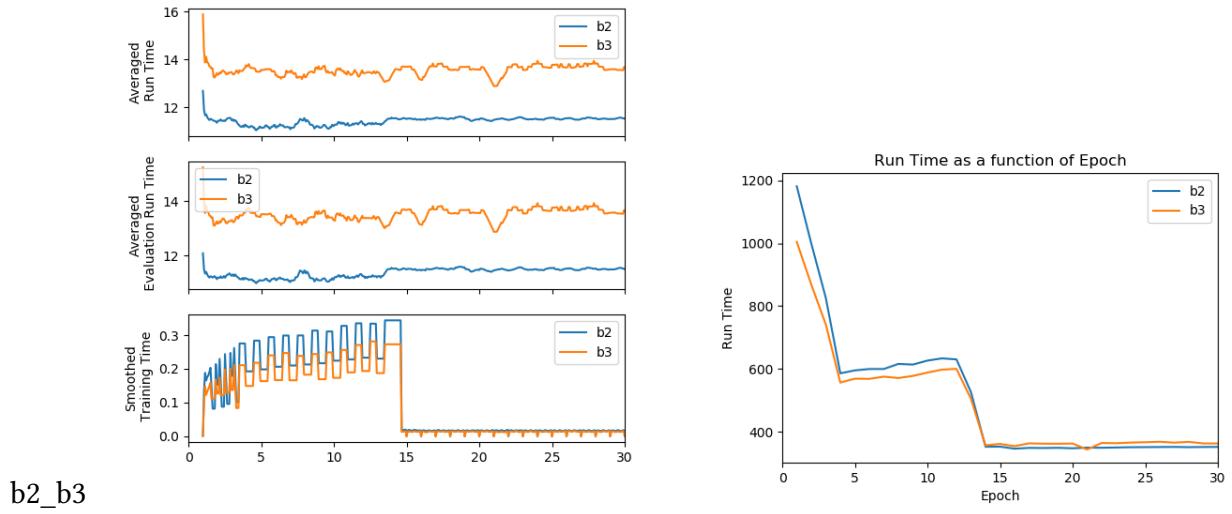
Series

Series	Model	Respective name of models on plots	Objective
b2_b3	b2, b3	“b2”, “b3”	Compare training with 2 and 3 batches
1	b2, l2, l3	“1 RNN layer”, “2 RNN layers”, “3 RNN layers”	See impact of number of RNN layers
lbl	b2, s, s-a20-l10	“Classical”, “Layer by layer (10, 10)”, “Layer by layer (20, 10)”	See impact of layer by layer training
sum	b2, s_l3_a	“Classical”, “3 RNN layers, LbL, attentive”	Check if layer by layer training, attentive model, and multi-RNN-layered architectures are compatibles

C.19.2 Results

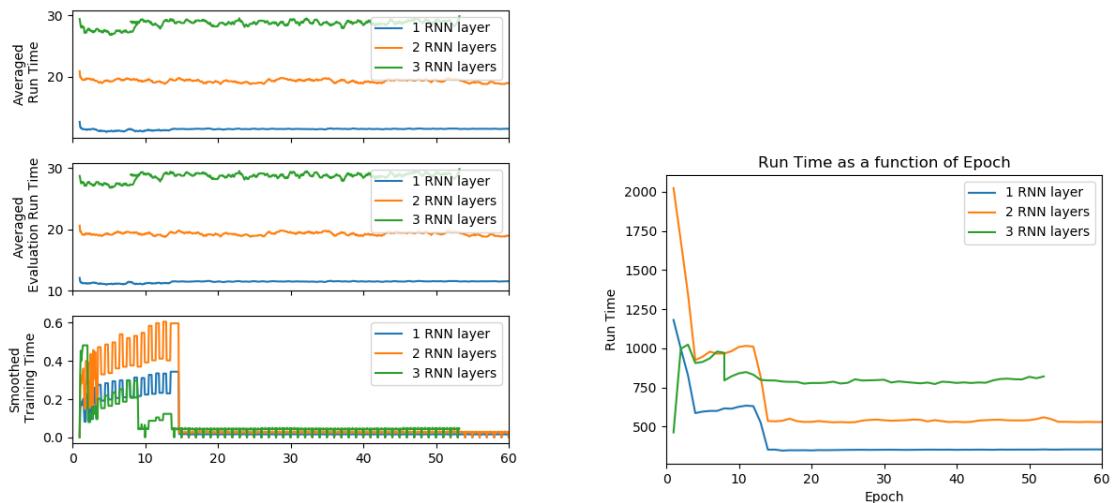
Series Time

Time to run an epoch BPC

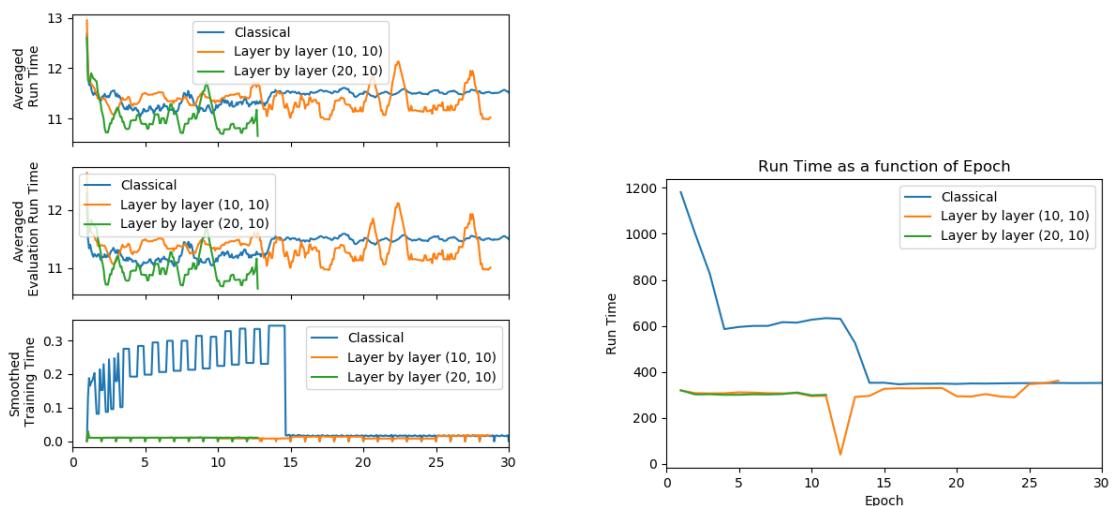


b2_b3

I



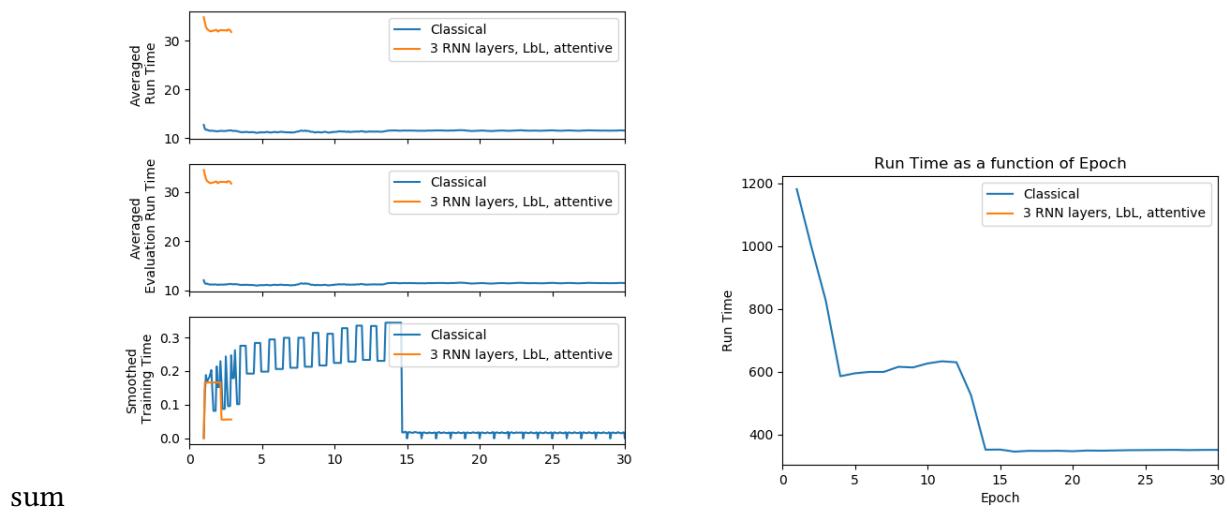
I

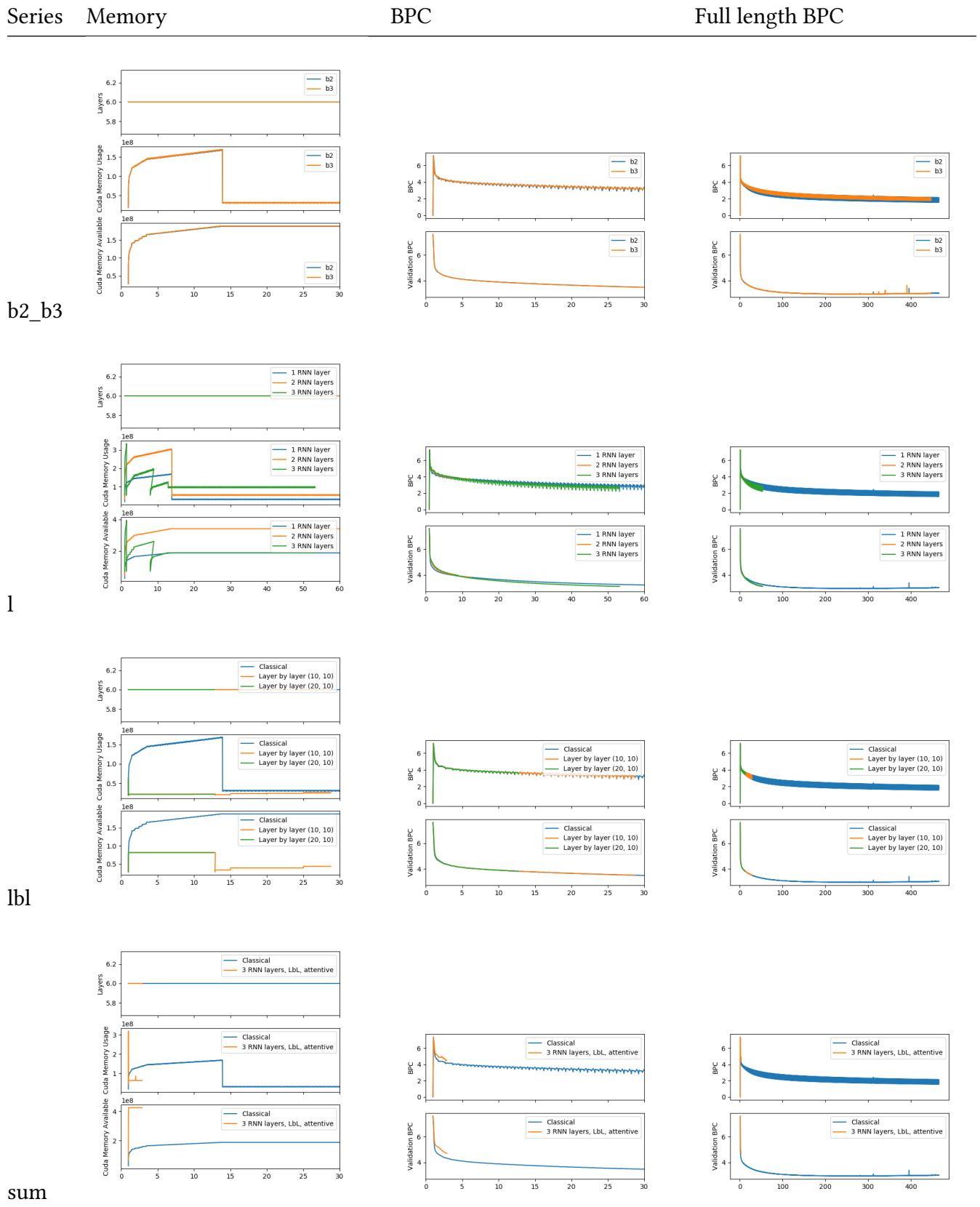


lbl

Series Time

Time to run an epoch BPC





Plots will be referred to with : (Ex : "l :Memory")

Time necessary to reach indicated validation BPC

Model	5 BPC	4 BPC	3 BPC
b2	0H 19M 41S	1H 19M 47S	15H 13M 7S
b3	0H 16M 44S	1H 11M 48S	15H 3M 22S
s	0H 5M 19S	0H 30M 59S	
s-a20-l10	0H 5M 19S	0H 30M 26S	
l2	0H 33M 43S	2H 27M 51S	
l3	0H 7M 41S	2H 13M 2S	
s_l3_a	0H 16M 7S		

Empty fields are where network has not reached BPC yet ~~Crossed out~~ fields are where data has been corrupted (because of an interruption in training)

Epochs necessary to reach indicated validation BPC

Model	5 BPC	4 BPC	3 BPC
b2	0.296	6.000	141.4
b3	0.367	5.954	136.4
s	0.371	6.000	
s-a20-l10	0.371	6.000	
l2	0.593	6.519	
l3	0.816	7.296	
s_l3_a	1.148		

Empty fields are where network has not reached BPC yet

C.19.3 Analysis

Run time and memory on “l” series (l :memory, l :time, and l :time to run an epoch)

Training of “l3” model was interrupted 2 times before the 15th epoch, each time cleaning the memory, inducing training time reduction and epoch time reduction (due to the correlation of memory usage and training time).

Run time and memory on “lbl” series (lbl :memory, lbl :time, and lbl :time to run an epoch)

The specificity of layer by layer training are a lead to the cause of the collapse of memory usage and run time during the 15th epoch.

During layer by layer training, memory usage is kept constant over the first 15 epochs, whereas with classical training, memory usage increases during this period.

Layer by layer training reduces simultaneous graph accumulation over the different epochs by training a layer at a time. It possibly reduces SGD inertia too.

Performance of layer by layer training

No notable variation on BPC. For run time and memory, see above.

Performance of multi-RNN-layered architecture

A slight improvement of BPC can be seen with 3 layers (sadly, BPC data for 2 layers is corrupted). Computation-time is proportional to the number of layers (time data for 3 layers is unusable).

Performance of 3 batches compared to 2 batches

2 and 3 batches have an almost identical performance (time-wise, BPC-wise and memory-wise), except 3 batches seem to get less dispersed BPC.

Results of long training with 2 and 3 batches (b2_b3 :full length bpc)

The plot “b2_b3 :full length bpc” shows that after 200 epochs, validation BPC stagnates. After 300 epochs, validation BPC begins to increase very slowly, those are the first signs of over-fitting. These observations were cross-checked with the raw data.

Note on attention module

A model with the attention module was tested for 5 hours, and did not reach a full epoch. It can be deemed that without a training algorithm like layer by layer training, using attention is not viable.

D Rapports d'avancement du projet PAPUD

D.1 Informations sur les documents contenus dans la présente annexe

Les sections suivantes contiennent les rapports intermédiaires fournis à notre maître de stage et au autres membres du projet au cours du projet PAPUD.

D.1.1 Format d'origine, transcription et contenu des rapports

Pour les mêmes raisons que pour l'annexe précédente (décrise dans la section C.1), les documents présentés dans cette annexe ont été rédigés en anglais, au format Gitlab Flavoured Markdown ; les versions présentées ici sont des transcriptions aussi fidèles que possible de ces documents.

D.2 Informations générales

General information

D.2.1 Corpus

firsttest

Baseline accuracy True baseline accuracy (for char -) : training 0.443 ; valid 0.414 ; test 0.423

char	training	valid	test
-	0.443	0.414	0.423
	0.064	0.067	0.063
a	0.021	0.022	0.023
o	0.015	0.013	0.011
<unk>	0.0	0.0	0.0

D.2.2 Grid-5000

To develop and train the model, we are using Grid-5000 computers clusters. For specific information on Grid-5000, see <https://www.grid5000.fr/>.

GPU-equipped nodes

Due to the properties of neural networks models, it is really efficient to use GPUs to train them.

Here are the main GPU-equipped nodes of Grid-5000 :

Nodes	GPU	Graphical memory	RAM	Production	CUDA
graphique-1	2 x Nvidia Titan Black	2 x 6GB	64GB	X	2880
graphique-2/6	2 x Nvidia GTX 980	2 x 4GB	64GB	X	2048
graphite-1/4	Intel Xeon Phi 7120P	16GB	256GB		?
grele-1/14	2 x Nvidia GTX 1080 Ti	2 x 11GB	128GB	X	3584
grimani-1/6	2 x Nvidia Tesla K40M	2 x 12GB	64GB	X	2880

[NOT TESTED YET] Model conversion

See <https://github.com/ysh329/deep-learning-model-convertor>

D.3 Résultats de l'implémentation basique

Results of the basic implementation of the model

2018/07/09 - SYNALP - Esteban MARQUER

D.3.1 Paradigm

The test is run on a minimal number of epoch (10), with a minimal model.

The training algorithm used is an example by example training.

Model architecture

The model is a line by line predictive model, composed of : - a character embedding layer; - a pooling layer; - a linear layer; - an output layer.

The output of the model is a probability distribution over known characters for every character of the predicted line.

D.3.2 Results

GPU memory usage

As expected from the model architecture, GPU memory usage is constant.

Computation time

Loss and accuracy

The loss used is cross-entropy loss, a character per character negative-log-likelihood loss over the soft-maxed distribution.

Overall, the loss gives a score to the prediction of the model, by comparing the target character and a distribution of probabilities for each character. If the probability for the target character is high and other character low, the model does a good prediction of the character, and the score given is low. The closer the score is to 0, the better it is. The scores of each characters is averaged, producing a global loss over the line.

Accuracy is a percentage. The closer to 100 % the better. As the loss is bound by 0 and +Infinity, and the closer to 0 the better, a correct transformation to accuracy could be : $\exp(-\text{loss})$ for an accuracy between 0 and 1.

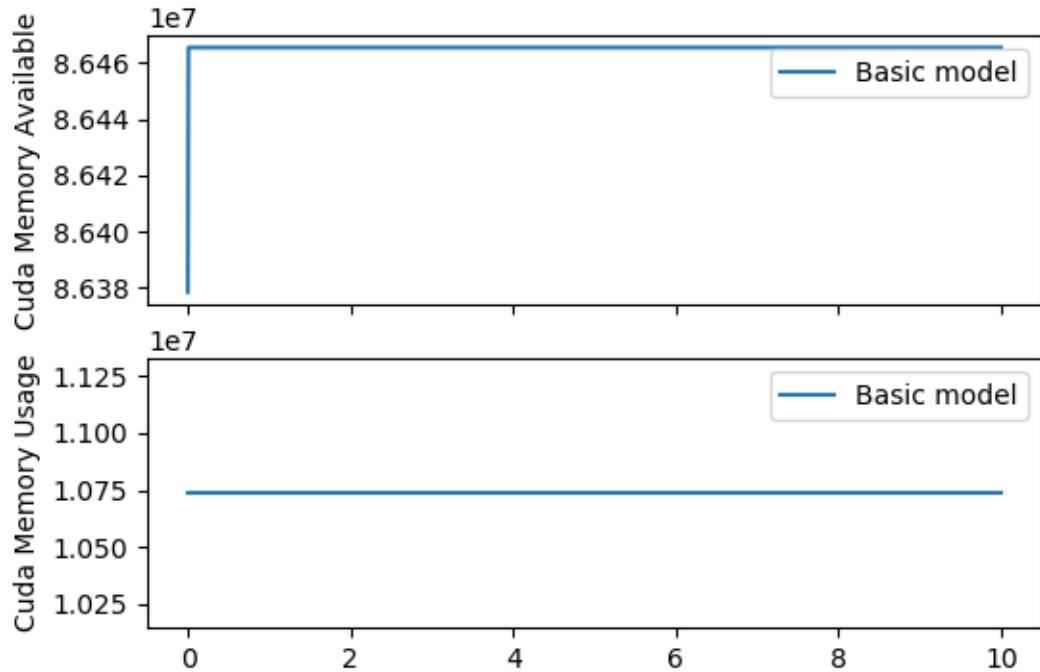


FIGURE D.1 – memory usage

The small spike recurrently appearing in the loss and accuracy is most likely due to a noisy part of the corpus (around the middle of the corpus) causing the model to learn wrongly on those specific examples.

The best precision obtained at the end of 10 epochs is 50%, corresponding to a loss of about 0.7.

D.3.3 Improvements and next steps

Mini-batch

Currently, the models learn one example at a time, meaning it computes the result for a line of input, compares it to the target, and updates weights. A common algorithm is the mini-batch algorithm, computing simultaneously a set of examples, their loss compared to the target, and updates the weights of the model all at once for the whole set of examples.

This algorithm speeds-up training while making the most of the GPU.

Dynamic corpus

While with the current corpus there is no real problem in storing the whole corpus in the memory, the future corpus will be over 400GB of text. It is necessary to replace the current method by a dynamic loading and transformation of the parts of the corpus currently used by the model. An ideal solution would be to read the target data directly from the archive containing the corpus.

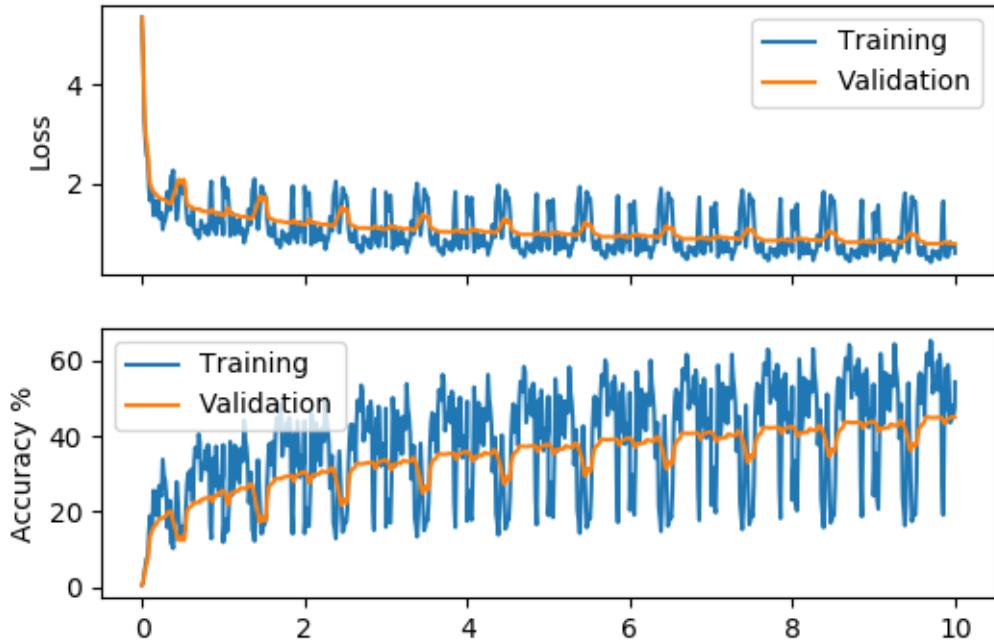


FIGURE D.2 – loss

D.4 Paquets (*batchs*) simultanés

Increasing the number of simultaneous examples

2018/07/10 - SYNALP - Esteban MARQUER

D.4.1 Paradigm

The test is run on a small number of epoch (20), with a minimal model and a new training algorithm.

The training algorithm used is a **mini-batch training**, meaning we compute the output for multiple examples all at once, we compute an averaged loss over those examples, and we update the model.

The potential effects of this algorithm are :

- an increase of GPU memory usage, as computations are done on larger data ;
- a decrease of computation time, with the number of computations reduced ;
- a smother training loss, because it is averaged over multiple examples ;
- avoidance of some local minima.

A second test with random batch-size between 1 and 1000 was done on 50 epoch, to evaluate the effect of the batch size and find an optimum.

D.4.2 Results

GPU memory usage

As more examples are fed to the model, there is a very slight increase in GPU memory usage : $0.013e7$ B, corresponding to 127kiB (this amount is negligible with more than 10GiB available and a current usage of about 10MiB).

Conclusion : increasing the number of simultaneous examples has no substantial downsides memory-wise.

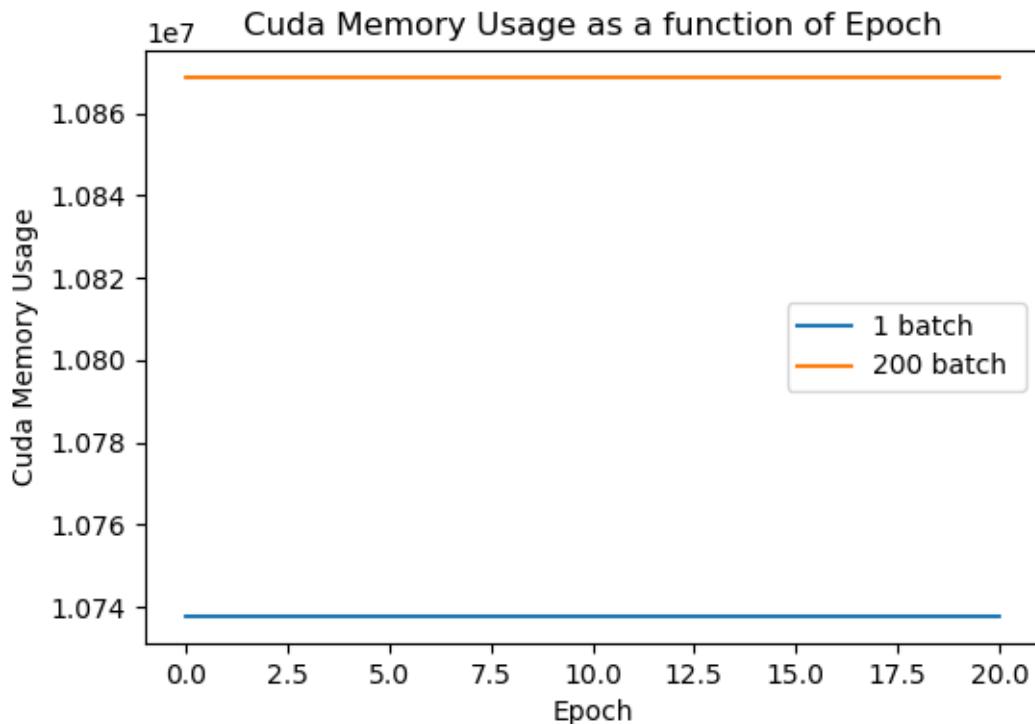
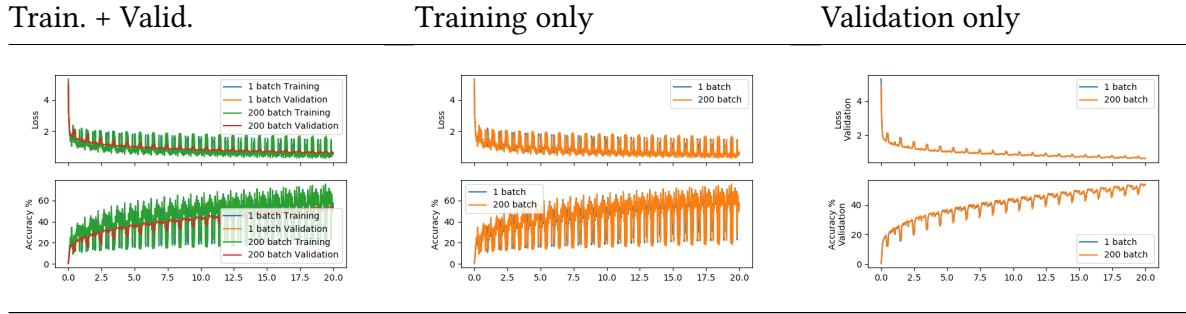


FIGURE D.3 – memory usage

Loss and accuracy

As loss is averaged on multiple examples, it should be smoother. But, probably because the number of simultaneous examples is too small, there is no noticeable change of loss, with the curves superposed.

Conclusion : increasing the number of simultaneous examples has no substantial effect loss-wise.



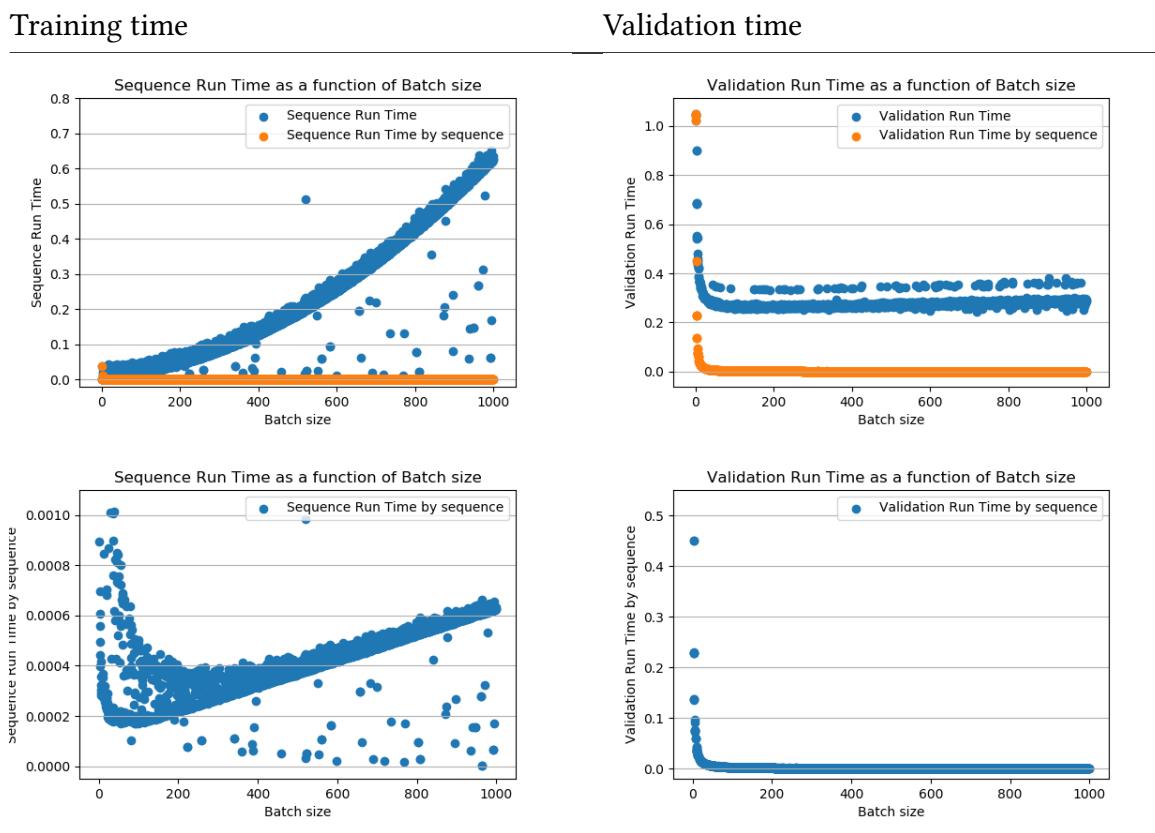
Computation time

The computations done on GPU benefit from grouping similar operations. By computing multiple examples together, we can use this property to speed up training. Moreover, retro-propagation and model updates are less frequent, reducing computational load and training time.

The best-case time allow an improvement from 10ms to less than 2ms per **training** sequence with a batch-size of 50, and the worst-case one allow 4ms per training sequence with a batch-size of 200.

A small gain can be achieved on **validation** time by increasing batch size over 50, but increasing it more has no effect.

Conclusion : increasing the number of simultaneous examples leads to a notable improvement of computation time.



D.4.3 Conclusion

Even if there is no improvement of loss or memory, the gain in computation time is enough to accept this algorithm.

The ideal batch-size (with the current node “grele”) is between 50 and 200. In future works, a batch-size of 200 will be used, as it present the best worst- and best-case time performances.

D.4.4 Improvements and next steps

Dynamic corpus

The dynamic corpus implementation is ready (except small details) and working, only integration is left.

Buffer size The dynamic corpus can use a buffer, and the size of this buffer must be at least the size of the batch. It will be necessary to test which size is optimal. An optimal buffer has the minimal size to make computation time over the buffer size only slightly higher than pre-loading time. It allows training to continue without interruption, while maintaining a low memory usage.

D.5 Analyse du pic de performance

Performance spike analysis

2018/07/18 - SYNALP - Esteban MARQUER

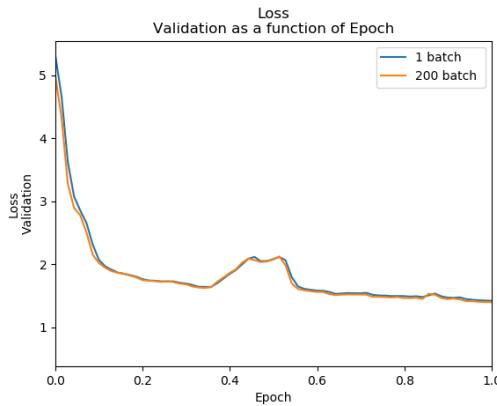
D.5.1 Problem

During training, a big performance spike appeared periodically. It is necessary to know why this spike appeared.

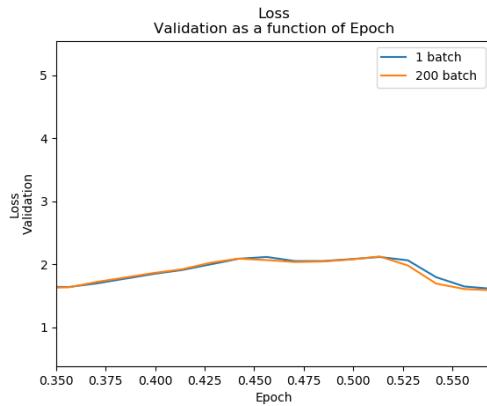
The different parts of the spike are :

- from line 24545 to line 31558 (35% to 45% of the corpus) : the increase of the BPC ;
- from line 31558 to line 36468 (45% to 52% of the corpus) : a stable part at a high value ;
- from line 36468 to line 38572 (52% to 55% of the corpus) : the decrease of the BPC.

Loss of 1 epoch (first epoch)



Zoom on spike



D.5.2 Analysis

By extracting the parts of the corpus corresponding to the parts of the spike, and scrolling through them, some recurrent elements appear : - lines beginning by kern, more specifically kern info and kern debug ; - lines containing a memory address, like 0 x91ffff , 0x0093 and 00000000fed18000, or an error code like 0x0100, - lines beginning by daemon, more specifically daemon err ;

The most interesting part of the spike is the increase of the BPC, were the performance deteriorate.

Given the repartition and percentages (see the next two sub-sections), the most likely causes for the spikes are :

- the memory address and hexadecimal codes ;
- the kernel messages (very repetitive, and containing memory address and hexadecimal codes).

Examples of Kernel messages Ces données ont été modifiées pour des raisons de confidentialité.

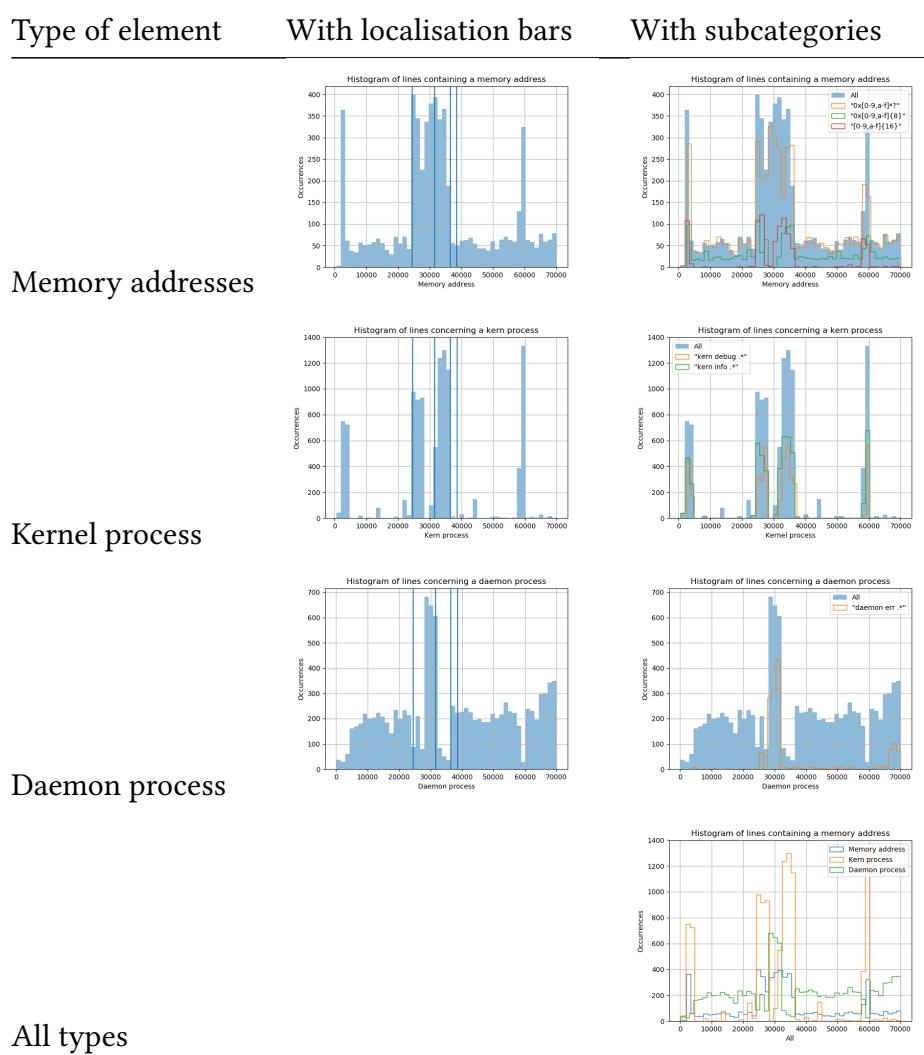
```

1 kern info kernel ABCD: FHJLN (abcd_id[0x00] fhjln_id[0x00] enabled)
2 kern info kernel ABCD: FHJLN (abcd_id[0x02] fhjln_id[0x02] enabled)
3 kern info kernel ABCD: FHJLN (abcd_id[0x04] fhjln_id[0x04] enabled)
4 kern info kernel ABCD: FHJLN (abcd_id[0x06] fhjln_id[0x06] enabled)
5 kern info kernel ABCD: FHJLN (abcd_id[0x08] fhjln_id[0x08] enabled)

```

Repartition of match in the corpus

The “localisation bars” (vertical blue lines) delimit to the different parts of the spike.



Percentages of match in each part of the corpus

Percentages of match are percentages on the total line number of the part of the corpus analysed.

```

1 --- spike up slope (lines 24545 to 31558, 35% to 45%) ---
2 Total lines: 7013
3 Matching "kern .*": 2888 (41%)
4 Matching "kern info .*": 1458 (20%)
5 Matching "kern debug .*": 1172 (16%)

```

```

6 Matching "daemon .*": 2048 (29%)
7 Matching "daemon err .*": 935 (13%)
8 Matching any memory pattern: 1728 (24%)
9 Matching "0x[0-9,a-f]{8}": : 196 (2%)
10 Matching "[0-9,a-f]{16}": : 250 (3%)
11 Matching "0x[0-9,a-f]*?": : 1478 (21%)
12
13 --- spike flat (lines 31558 to 36468, 45% to 52%) ---
14 Total lines: 4910
15 Matching "kern .*": 4218 (85%)
16 Matching "kern info .*": 2154 (43%)
17 Matching "kern debug .*": 1760 (35%)
18 Matching "daemon .*": 346 (7%)
19 Matching "daemon err .*": 174 (3%)
20 Matching any memory pattern: 1149 (23%)
21 Matching "0x[0-9,a-f]{8}": : 294 (5%)
22 Matching "[0-9,a-f]{16}": : 296 (6%)
23 Matching "0x[0-9,a-f]*?": : 853 (17%)
24
25 --- spike down slope (lines 36468 to 38572, 52% to 55%) ---
26 Total lines: 2104
27 Matching "kern .*": 27 (1%)
28 Matching "kern info .*": 15 (0%)
29 Matching "kern debug .*": 0 (0%)
30 Matching "daemon .*": 383 (18%)
31 Matching "daemon err .*": 15 (0%)
32 Matching any memory pattern: 90 (4%)
33 Matching "0x[0-9,a-f]{8}": : 34 (1%)
34 Matching "[0-9,a-f]{16}": : 5 (0%)
35 Matching "0x[0-9,a-f]*?": : 85 (4%)
36
37 --- whole spike (lines 24545 to 38572, 35% to 55%) ---
38 Total lines: 14027
39 Matching "kern .*": 7133 (50%)
40 Matching "kern info .*": 3627 (25%)
41 Matching "kern debug .*": 2932 (20%)
42 Matching "daemon .*": 2777 (19%)
43 Matching "daemon err .*": 1124 (8%)
44 Matching any memory pattern: 2967 (21%)
45 Matching "0x[0-9,a-f]{8}": : 524 (3%)
46 Matching "[0-9,a-f]{16}": : 551 (3%)
47 Matching "0x[0-9,a-f]*?": : 2416 (17%)
48
49 --- full corpus ---
50 Total lines: 70131
51 Matching "kern .*": 10972 (15%)
52 Matching "kern info .*": 5298 (7%)
53 Matching "kern debug .*": 4134 (5%)
54 Matching "daemon .*": 10828 (15%)
55 Matching "daemon err .*": 1504 (2%)
56 Matching any memory pattern: 5859 (8%)
57 Matching "0x[0-9,a-f]{8}": : 1586 (2%)
58 Matching "[0-9,a-f]{16}": : 893 (1%)
59 Matching "0x[0-9,a-f]*?": : 4987 (7%)
60
61 Matching "kern .*" outside of spike: 3839 (5%)

```

D.5.3 Conclusion(s)

There are two possible conclusions :

- the kernel messages are the cause of the spike ;
- or the memory addresses and hexadecimal codes are the cause of the spike.

Kernel messages

If the kernel messages are the cause of the spike, the most likely explanation is that this part of the corpus represent a crash of the server or a major error. In that case, we must remove that part of the corpus from the training set, as it is not the “normal” evolution of the log.

Memory addresses and hexadecimal codes

If the memory addresses and hexadecimal codes are the cause of the spike, it should be because a succession of number is a very specific thing to learn. In that case, either we let the model learn the brute codes, or we replace every code by a “<hex>” character to ease the learning process. It is also possible to replace the different kind of code by a different character.

D.5.4 Improvements and next steps

To check whether the memory addresses and hexadecimal codes, or the kernel messages are the cause of the spike, trying to train the model while replacing every code by a “<hex>” character. If there is no improvement, then the codes are not the cause of the performance spike.

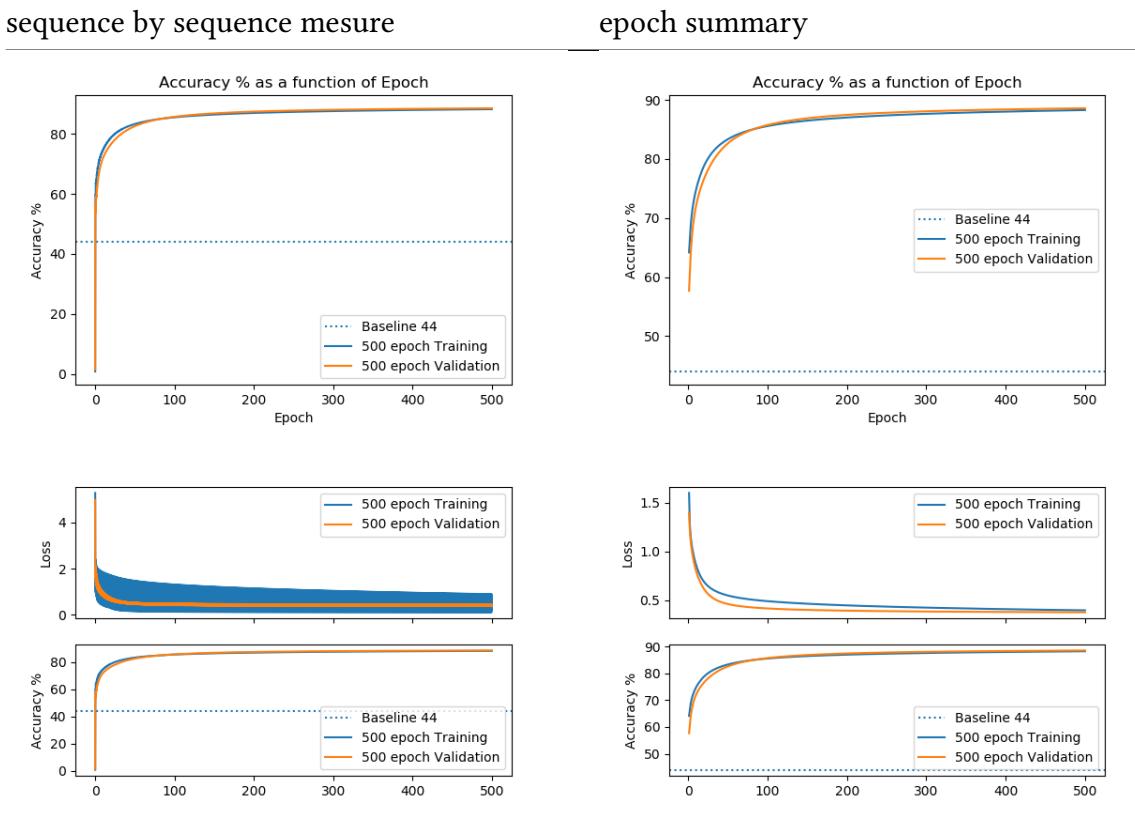
D.6 Rapport de la réunion avec les autres membres du projet

Team meeting report

2018/07/20 - SYNALP - Esteban MARQUER

D.6.1 Main points

- Performance spike issue : see report “2018_07_18-Performance_spike_analysis”
- solution chosen : replace hexadecimal codes with a special label per kind of code
- Long training : no over-fitting, 90% accuracy in 500 epochs



- Baseline accuracy : see first section of “General_information.md” for values.
 - Baseline accuracy is high with padding (about 50%), perhaps lines are too long (a lot of padding is needed)
 - Compared with performance, performance is still good
- GPU usage with completely loaded corpus : 30% to 60%
 - This is bad news, with such a model training should go at 99% all the time
 - It is necessary to locate the element slowing the process, try removing all unnecessary processes (logging, storage in memory, accuracy, ...)
- New corpus implementation (with buffer and iterators) : about 180s/epoch, 65s/epoch with old implementation (everything loaded in memory)
 - a good implementation of the data loading is critical

- perhaps pre-loading is too slow because of computations, a pre-processed version could help
- training the model multiple time on a loaded segment could bridge the gap between the two processes used
- using more processes could do the trick (one for loading only, one for processing, and one for training)
- using “binary”-sized batches (like 8, 64 or 1024) is said to achieve faster results, maybe a bit of speed can be gained there
- Development of a learning rate optimisation script : good, better if offline (good if both online and offline)
 - online stands for optimisation before, or/and during every training
 - offline stands for an analysis done a single time, aside from any training

D.6.2 Improvements and next steps

- Finish the development of learning rate optimisation.
- Try to make the corpus implementation clean and fast enough (with compared run times).
- Integrate the modifications of the corpus processing (memory address management)
- Use “binary”-sized batches ; 128 seems perfect, as it is between 50 and 200 (the bounds found when optimising batches).

D.7 Optimisation du taux d'apprentissage

Learning rate optimisation

2018/07/23 - SYNALP - Esteban MARQUER

D.7.1 General information

Learning rate is an hyperparameter in the training algorithm which changes the speed of the training and the performance of the model. Specifically, it is a coefficient of the gradient used to update the parameters of the model.

There are three main learning rates for a model :

- a learning rate that is too small (closer to 0) : the training is slow, and can get blocked in some local minima ;
- a learning rate that is too high (closer to +infinity, usually closer to 1) : the learning is faster and avoid local minima, but could diverge from the solution ;
- a balanced learning rate : what we want to find, the traing is fast yet does not diverge.

D.7.2 Optimisation process

Usually, learning rate optimisation is done with a logarithmic scale of the learning rate. The shape of the produced curves confirm the use of such a scale.

The optimisation process is driven by three parameters and a single metric.

The metric is the accuracy of the model on the validation corpus at the end of the training.

The parameters are : the two bounds of the learning rate, and the learning rate variation factor : the “learning rate multiplier”.

The learning rate varies as follow in the psedo-python algorythm :

```
1 # the learning rate takes the highest value as start value, as it will
  # decrease over time
2 learning_rate = start_value
3
4 # until the learning rate as reach the stop value (the lowest of the two
  # bounds), we make it vary logarithmically
5 while learning_rate > stop_value:
6     performance = train_model(learning_rate)
7     save_model_performance(performance, learning_rate)
8
9     # the learning rate is updated
10    # example with a learning rate of 1 and a learning rate multiplier of 0.1:
11    # at first the learning rate is 1, then 0.1, then 0.01 ...
12    learning_rate = learning_rate * learning_rate_multiplier
13
14 # we compare the performance of the model with the different learning rates
15 compare_model_performance()
```

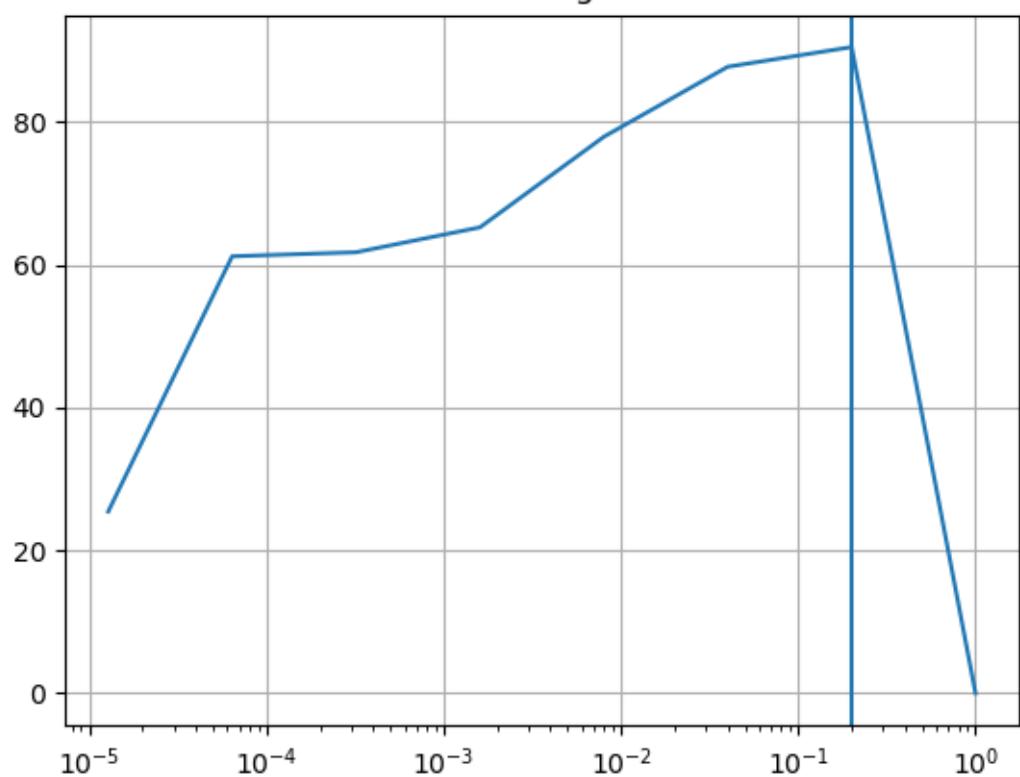
The closer to 0 the learning rate multiplier is, the faster the variation will be, and the closer to 1 it is, the slower the variation will be. To have a decent resolution in the curves, a multiplier close to 1 is crucial.

The training is done each time on a full epoch.

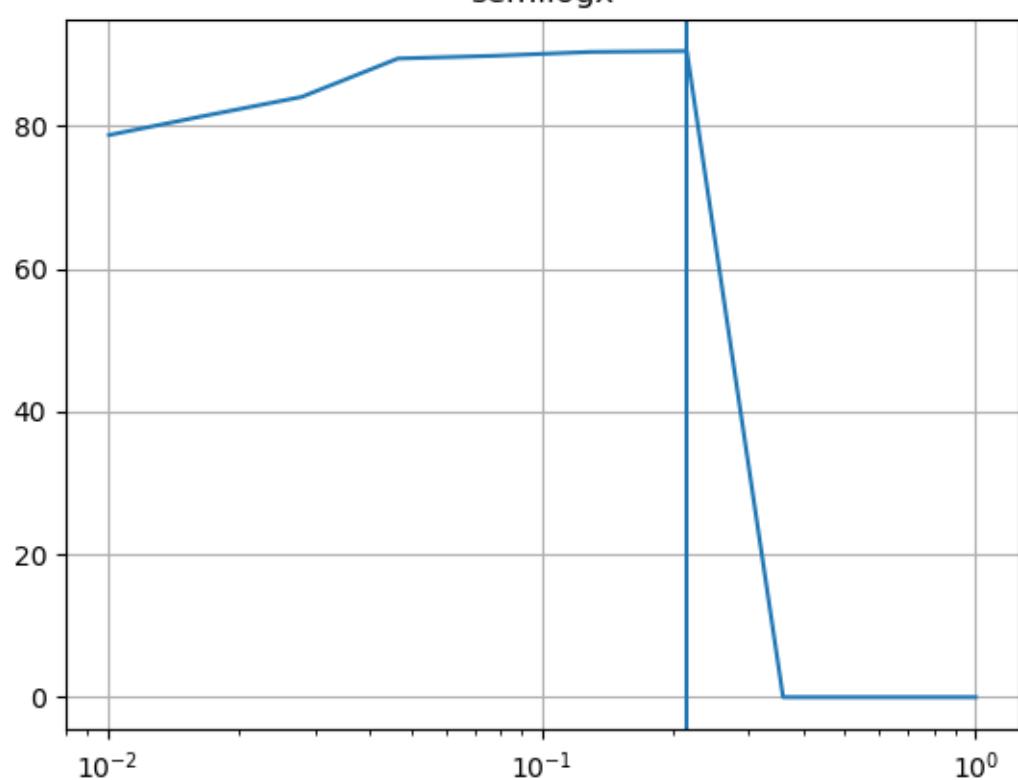
D.7.3 Results

The first plot is done with a learning rate between 1 and 10^{-5} , with a multiplier of 0.2. The second one is done with a learning rate between 1 and 10^{-2} , with a multiplier of 0.6 (the total of dots is 10).

semilogx



semilogx



There is a strange accuracy at the end of the first plot : the theoretically unreachable 0%. It is confirmed by the second plot, with multiple points having an accuracy of 0%. With a baseline accuracy of 44%, it is clear that the result diverge from what is expected. It is a case of divergence due to a learning rate that is too high.

The best learning rate found is 0.216 (0.6^3), giving the fastest learning (over 90% of accuracy in 1 epoch) without diverging.

D.7.4 Additional information

The current implementation consist of a script building a model, and finding the ideal learning rate for this model. It is an offline implementation of the model.

The way it is implemented is ideal for an online use too, as the only operation needed are the removal of the plotting part, and adding the reload of the model with an updated learning rate.

D.7.5 Improvements and next steps

Use the new learning rate in the training. As it is quite close to diverge, it would be advised to use a slightly lower learning rate like 0.2.

If inline optimisation is used, three possibilities seem viable : - choosing the learning rate every time we start a training, to adapt to the current hyperparameters ; - updating the learning rate every set number of epochs, to adapt the learning to the current state of the model; - doing each epoch with a set of learning rates, and every time choosing the best result; even if costly (every epoch is done multiple time), it should allow a really fast learning with lowered chances of divergence or over-fitting.

Personally, my preferred option is the second one.

D.8 Effets de l'optimisation du taux d'apprentissage

Learning rate optimisation effect on learning curve

2018/07/24 - SYNALP - Esteban MARQUER

D.8.1 Context

The previous results of learning rate optimisation lead to a potentially optimal learning rate of 0.2 (instead of 0.001).

A run with the new learning rate and every other thing identical was done to compare performance to previous 500 epoch run. That specific run had a learning rate of 0.001.

D.8.2 Results

The new learning rate has two effects : 1. convergence is achieved in less than 100 epoch, compared to previous learning achieving convergence in more than 400 epochs; 2. a small but constant gap between training performance and validation performance, but it is not over-fitting (if both training and validation are constant, we can not conclude that the model over-fits).

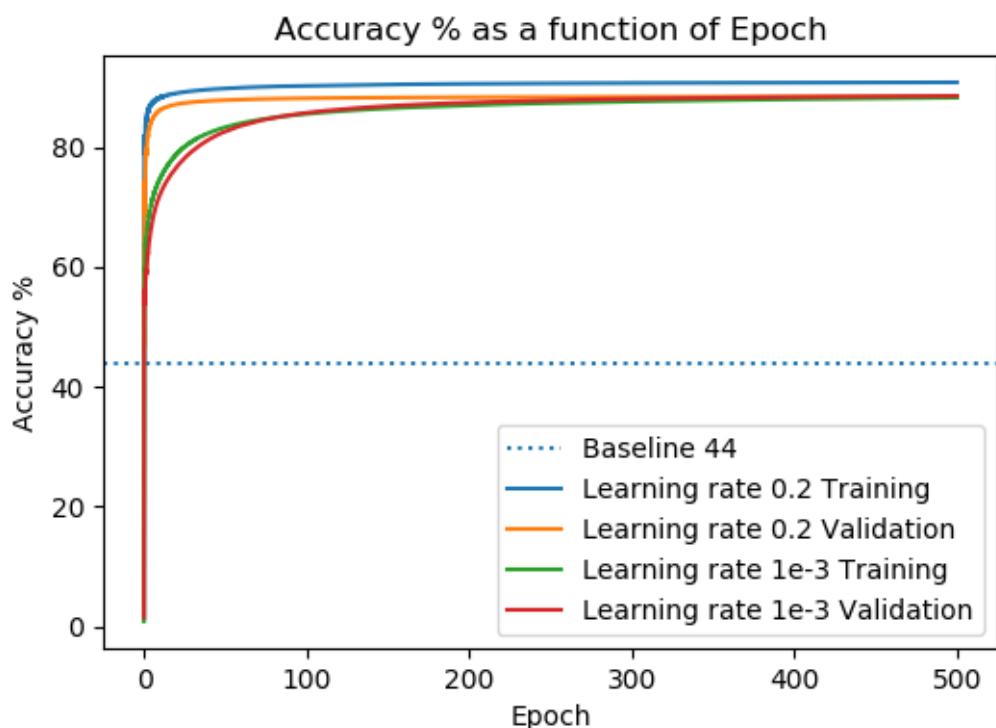


FIGURE D.4 – accuracy

D.8.3 Conclusion

The new learning rate is better than the previous one, it will be kept.

D.8.4 Improvements and next steps

[OPTIONAL] Use inline optimisation to update the learning rate every set number of epochs (once convergence is reached).

D.9 Taille de *batch* binaire

Binary batch size effect on run speed

2018/07/26 - SYNALP - Esteban MARQUER

D.9.1 Context

It has been said that batches using binary sizes (64, 256, 1024, ...) perform faster than non-binary sizes.

D.9.2 Paradigm

To verify this phenomenon and perhaps improve the training speed, a comparative experiment has been done with a batch size of 128 and a batch size of 200.

D.9.3 Results

There is no notable effect, except the effect predicted by the batch size comparison done previously, and stating that batches with a size of 200 are faster than with a size of 128.

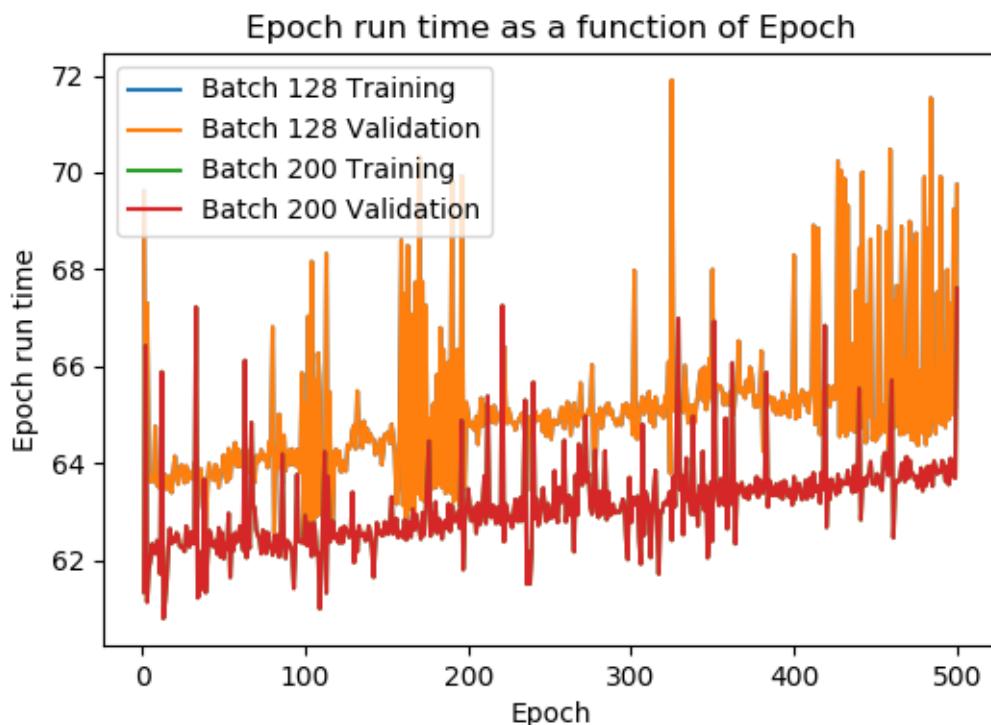


FIGURE D.5 – time per epoch

D.9.4 Conclusion

The most straightforward conclusion is that either the effect of binary sizes is negligible if not non-existent, or there is a hidden parameter changing the true size of the data (for example a set of bytes used to store metadata together with the regular data).

As the previous batch size (200) performs better, we will keep it.

D.9.5 Next steps and improvements

Investigating a potentially non-existent effect (at least with the current model) does not seem efficient considering the potential gain in computation time (compared to improving the naive model).

No further work will be done on that now (at least for now).

D.10 Performances du lecteur de corpus multi-fichiers multi-processus

Learning rate optimisation effect on

2018/07/26 - SYNALP - Esteban MARQUER

D.10.1 Context

Given the large amount of data to process, and the way it is structured in many small archives, a way to load and pre-process them efficiently had to be implemented.

D.10.2 Multi-file-multi-process corpus concept

Multi-process system

To avoid the need to completely pre-processing the data and storing it, and to shift the computational weight of the loading process, the different tasks are splitted between multiple processes.

At first, the idea was to use three processes, with one for the loading, one for the pre-processing, and the last one for the transformation into tensor. All those process fetch data from a multiprocess-safe queue and store the result in an output queue, used by the next process as an input.

A representation of the initial concept :

```
1 HDD --> Loader --> [ raw data queue ]
2      --> Processing --> [ processed data queue ]
3      --> Tensorising --> [ tensors queue ]
4      --> model
```

Given the performance of that system, the slowest part was the processing. Moreover, the processing could be splitted in multiple modules : removing the end-of-line characters, replacing patterns by tags and splitting into characters, transforming the data via a dictionary, and padding/cropping the sequence.

A representation of the modular processing concept :

```
1 HDD --> Loader --> [ raw Q. ]
2      --> Process 1 --> [ partially processed Q. 1 ]
3      --> Process 2 --> [ partially processed Q. 2 ]
4      ...
5      --> Process n --> [ processed Q. n ]
6      --> Tensorising --> [ tensors queue ]
7      --> model
```

This architecture allows to change the order of the pre-processing modules, and to add or remove some of them.

By splitting those different tasks on multiple processes, an efficient processing is achieved. Moreover, it is way easier to add pre-processing steps.

Multi-file system

By considering a set of files as a single sequence of line, and by loading only the one file containing the current data, combined by efficient line-by-line sequential reading, a light and fast loading is achieved.

D.10.3 Tests

Paradigm

While setting up the unitary tests for the corpus implementations, three main situations have produced useful insight on the performances of the new implementations.

Everything was run on my laptop, with other processes running that may have hindered the performance (for example, a heavy IDE).

Process-specific tests were run after every element was proven to do their expected job. Those tests added the processes and the data queues and information on queue filling and process state.

Results

Multiple test were done with very light, light treatment of the data, and real-situation training (the training is a way to process the data).

Printing only the status of the process at each batch Total time per epoch 12 s

```
1 { 'example': '28032/67923',
2   'batch': '218/527',
3   'iterator status':
4     'Process MultiFile status: process: alive; output queue: 1023/1024
5     Process EndLine status: process: alive; output queue: 1023/1024
6     Process Regex status: process: alive; output queue: 1024/1024
7     Process Dictionary status: process: alive; output queue: 2/1024
8     Process CropPad status: process: alive; output queue: 10/1024
9     Process Batch status: process: alive; output queue: 0/32'}
```

Printing the status of the process at each batch and printing the data Total time per epoch 47 s

```
1 { 'example': '30208/67923',
2   'batch': '235/527',
3   'iterator status':
4     'Process MultiFile status: process: alive; output queue: 1024/1024
5       Process EndLine status: process: alive; output queue: 1024/1024
6       Process Regex status: process: alive; output queue: 1023/1024
7       Process Dictionary status: process: alive; output queue: 916/1024
8       Process CropPad status: process: alive; output queue: 1024/1024
9       Process Batch status: process: alive; output queue: 32/32'}
```

Printing the status of the process at each batch and training the model Total time per epoch about 300 s, the old implementation needed about 200 s with the corpus full loaded in GPU RAM. CPU usage (usage mainly by the program sub-processes) : from 60% to 100%

```
1 { 'example': '26800/67923',
2   'batch': '134/338',
3   'iterator status':
4     'Process MultiFile status: process: alive; output queue: 1024/1024
5       Process EndLine status: process: alive; output queue: 1022/1024
6       Process Regex status: process: alive; output queue: 1017/1024
7       Process Dictionary status: process: alive; output queue: 905/1024
8       Process CropPad status: process: alive; output queue: 975/1024
9       Process Batch status: process: alive; output queue: 64/64'}
```

D.10.4 Conclusions

The first test shows the without processing the data, we can find where the data is “blocked”, and so find the slowest process in the bunch. Here, the slowest process is the transformation into ids.

As it is a simple dicitonary reading, which is already a very fast process in Python, if it is the slowest process of the chain, we can conclude that the basic performance of this implementation is very good.

When we do even a tny bit of processing (like printing the data), the data is blocked at the end of the chain, meaning the printing process is slower than the loading/pre-processing/tensorising process.

In a true training situation, we can confirm that processing the data is slower than pre-processing it. The only process slowing the whole training is the transfer to GPU RAM, which is not in a separate process yet (due to specificities of pytorch tensor mangement).

The implementation produce equivalent results with multipple files.

Globally, this new implementation should be enough to load and preprocess the data for the model, with both small and large datasets.

D.10.5 Next steps and improvements

- It should be possible to delegate the transfer to GPU RAM to a specific process (given the explanations on pytorch manual). This should reduce the gap between the speed achieved with pre-loaded data andthe speed of the new corpus system.
- The direct next step is to test the new corpus implementation on the computer cluster.
- Then, changing the current small corpus by a larger multi-file corpus will be possible.

E Copie de la convention de stage



CONVENTION DE STAGE

ENTRE

L'établissement d'enseignement supérieur :

Université de Lorraine, établissement public à caractère scientifique, culturel et professionnel, sis 34 Cours Léopold – CS 25233 – 54 052 NANCY Cedex, siren n° 130 015 506 00012, représenté par son Président, Monsieur Pierre Mutzenhardt,

Représenté par : (nom du (de la) signataire de la convention) : Antoine TABBONE

Qualité du représentant : Directeur de l'UFR Mathématiques et Informatique

Composante /UFR/ : UFR Mathématiques et Informatique

Adresse : 56 bis, boulevard de Scarpone, 54000 Nancy

Tel : 03 72 74 16 18

L'organisme d'accueil :

Nom : LORIA UMR 7503

Adresse : Campus Scientifique BP 239, 54506 Vandoeuvre-les-Nancy

Tél : 03 83 59 20 00

Fax : _____

Mél : dirrection@loria.fr

Représenté par : (nom du signataire de la convention) : Jean-Yves MARION

Qualité du représentant : Directeur du LORIA

Nom du service dans lequel le stage sera effectué : Équipe SYNALP

Lieu du stage : (si différent de l'adresse de l'entreprise) _____

Et l'étudiant stagiaire :

Nom : MARQUER Prénom : Esteban

Sexe : F M né(e) le : 08 / 06 / 1997

Adresse : 6, Rue Cyfflé, 54000, Nancy

Tél : 06 78 09 35 84

Mél : marquer.esteban7@etu.univ-lorraine.fr

Intitulé complet de la formation ou du cursus suivi dans l'établissement d'enseignement supérieur et son volume horaire :

Licence L3 MIASHS: MIAGE / Sciences cognitives Volume horaire : 600 h de présence / an

SUJET DE STAGE : Interpréter les couches cachées des réseaux récurrents en TAL

DATES DE STAGE : Du 23 / 04 / 2018 Au 27 / 07 / 2018

DUREE TOTALE DU STAGE * : 3 Heures ou Semaines ou Mois (rayer la mention inutile) soit en JOURS : 66

*7 heures = 1 jour et 22 jours = 1 mois

Encadrement du stagiaire assuré par :

L'établissement d'enseignement supérieur en la personne de :

Nom : THOMANN

Prénom : Laurent

Fonction : Professeur. Responsable des stages

Tél : 03 72 74 16 24

Mél : Laurent.thomann@univ-lorraine.fr

Caisse primaire d'assurances maladie à contacter en cas d'accident (lieu de domicile de l'étudiant sauf exception) :

¹ Article L612-9 du code de l'éducation : La durée du ou des stages effectués par un même stagiaire dans une même entreprise ne peut excéder six mois par année d'enseignement.

Article 1 : Objet de la convention

La présente convention règle les rapports de l'organisme d'accueil (entreprise, organisme public, association...) avec l'établissement d'enseignement supérieur et le stagiaire.

Article 2 : Objectif du stage

Le stage correspond à une période temporaire de mise en situation en milieu professionnel au cours de laquelle l'étudiant(e) acquiert des compétences professionnelles et met en œuvre les acquis de sa formation en vue de l'obtention d'un diplôme ou d'une certification et de favoriser son insertion professionnelle. Le/la stagiaire se voit confier une ou des missions conformes au projet pédagogique défini par son établissement d'enseignement et approuvées par l'organisme d'accueil.

Le programme est établi par l'établissement d'enseignement et l'organisme d'accueil en fonction du programme général de la formation dispensée.

Activités confiées : état de l'art sur les réseaux récurrents (RNN)
et leurs méthodes d'analyse ; récupération de programmes RNN
en pytorch ; analyse et interprétation des informations apprises
dans les couches cachées

Compétences à acquérir ou à développer :

maîtrise du deep learning ; programmation python

Article 3 : Modalité du stage

La durée hebdomadaire maximale de présence du (de la) stagiaire dans l'entreprise sera de 35 heures.

Le stage est à :
 Temps complet Temps partiel

Si temps partiel, préciser la quotité : _____

Si le (la) stagiaire doit être présent(e) dans l'organisme d'accueil la nuit, le dimanche ou un jour férié, l'organisme doit indiquer ci-après les cas particuliers : _____

Article 4 : Statut du stagiaire – Accueil et encadrement

L'étudiant(e), pendant la durée de son stage dans l'organisme d'accueil, conserve son statut antérieur; il (elle) est suivi(e) régulièrement par l'enseignant référent désigné dans la présente convention ainsi que par l'établissement. L'organisme d'accueil nomme un tuteur organisme d'accueil chargé d'assurer le suivi et d'optimiser les conditions de réalisation du stage. L'étudiant(e) pourra revenir à l'établissement pendant la durée du stage, pour y suivre certains cours demandés explicitement par le programme, participer à des réunions, les dates étant portées à la connaissance de l'organisme d'accueil par l'établissement et être autorisé, le cas échéant, à se déplacer.

Toute difficulté survenue dans la réalisation et le déroulement du stage ou, qu'elle soit constatée par le/la stagiaire ou par le tuteur de stage, doit être portée à la connaissance de l'enseignant référent et de l'établissement d'enseignement afin d'être résolue au plus vite.

Modalités d'encadrement : _____

Article 5 : Gratification – Avantages en nature Remboursement de frais

Lorsque la durée du stage est supérieure à deux mois consécutifs ou non, celui-ci fait obligatoirement l'objet d'une gratification sauf en cas de règles particulières applicables dans certaines collectivités d'outre-mer françaises et pour les stages relevant de l'article L4381-1 du code de la santé publique.

Le montant horaire de la gratification est fixé à 15 % du plafond horaire de la sécurité sociale défini en application de l'article L.241-3 du code de la sécurité sociale. Une convention de branche ou un accord professionnel peut définir un montant supérieur à ce taux.

La gratification est due à compter du premier jour du premier mois de la période de stage.

La gratification ne peut être cumulée avec une rémunération versée par l'administration ou l'établissement public d'accueil au cours de la période concernée.

La gratification est due au stagiaire sans préjudice du remboursement des frais engagés par le/la stagiaire pour effectuer son stage et des avantages offerts, le cas échéant, pour la restauration, l'hébergement et le transport.

L'organisme peut décider de verser une gratification pour les stages dont la durée est inférieure ou égale à deux mois.

En cas de suspension ou de résiliation de la présente convention, le montant de la gratification due au/à la stagiaire est proratisé en fonction de la durée du stage effectué.

La durée donnant droit à gratification s'apprécie compte tenu de la présente convention et de ses avenants éventuels, ainsi que du nombre de jours de présence effective du/de la stagiaire dans l'organisme.

Montant de la gratification (si différent du montant légal)

Modalités de versement de la gratification : _____

Le/la stagiaire bénéficie des protections et droits mentionnés aux articles L.1121-1, L.1152-1 et L.1153-1 du code du travail, dans les mêmes conditions que les salariés.

Le/la stagiaire a accès au restaurant d'entreprise ou aux titres-restaurants prévus à l'article L.3262-1 du code du travail, dans les mêmes conditions que les salariés de l'organisme d'accueil. Il/elle bénéficie également de la prise en charge des frais de transport prévue à l'article L.3261-2 du même code.

Les stagiaires accèdent aux activités sociales et culturelles mentionnées à l'article L.2323-83 du code du travail dans les mêmes conditions que les salariés.

Liste des avantages offerts : _____

Les trajets effectués par les stagiaires d'un organisme de droit public entre leur domicile et leur lieu de stage peuvent être pris en charge dans les conditions fixées par le décret n°2010-676 du 21 juin 2010 instituant une prise en charge partielle du prix des titres d'abonnement correspondant aux déplacements effectués par les agents publics entre leur résidence habituelle et leur lieu de travail.

la stagiaire accueilli(e) dans un organisme de droit public et qui effectue une mission dans ce cadre bénéficie des dispositions du décret n°2006-781 du 3 juillet 2006 fixant les conditions et les modalités de règlement des frais occasionnés par déplacements temporaires des personnels civils de l'Etat.

Est considéré comme sa résidence administrative le lieu du stage indiqué dans la présente convention.

Autres avantages :

Autres avantages : _____

Article 6 : Protection sociale

Pendant la durée du stage, l'étudiant(e) reste affilié(e) à son système de sécurité sociale antérieur : il(elle) conserve son statut étudiant. Les stages effectués à l'étranger doivent avoir été signalés préalablement au départ de l'étudiant(e) et avoir reçu l'agrément de la Sécurité Sociale. Les dispositions suivantes sont applicables sous réserve de conformité avec la législation du pays d'accueil et de celle régissant le type d'organisme d'accueil :

6.1 Gratification inférieure ou égale au produit de 15 % du plafond horaire de la sécurité sociale par le nombre d'heures de stage effectuées au cours du mois considéré :

Dans ce cas, conformément à la législation en vigueur, la gratification de stage n'est pas soumise à cotisation sociale. L'étudiant(e) continue à bénéficier de la législation sur les accidents de travail au titre de l'article L 412-8-2 du code de la Sécurité Sociale, régime étudiant. En cas d'accident survenant à l'étudiant(e), soit au cours des travaux dans l'organisme, soit au cours du trajet, soit sur les lieux rendus utiles pour les besoins de son stage et pour les étudiant(e)s en médecine, en chirurgie dentaire ou en pharmacie qui n'ont pas un statut hospitalier, du stage hospitalier effectué dans les conditions prévues au b du 2o de l'article L. 412-8, l'organisme d'accueil envoie la déclaration à la Caisse Primaire d'Assurance Maladie (voir adresse en première page) en mentionnant l'établissement comme employeur, avec copie à l'établissement.

6.2 Gratification supérieure au produit de 15 % du plafond horaire de la sécurité sociale par le nombre d'heures de stage effectuées au cours du mois considéré :

Les cotisations sociales sont calculées sur le différentiel entre le montant de la gratification et 15 % du plafond horaire de la Sécurité Sociale pour une durée légale de travail hebdomadaire de 35 heures. L'étudiant(e) bénéficie de la couverture légale en application des dispositions des articles L 411-1 et suivants du code de la Sécurité Sociale. En cas d'accident survenant à l'étudiant(e), soit au cours des travaux dans l'organisme, soit au cours du trajet, soit sur des lieux rendus utiles pour les besoins de son stage, l'organisme d'accueil effectue toutes les démarches nécessaires auprès de la Caisse Primaire d'Assurance Maladie et informe l'établissement dans les meilleurs délais.

6.3 Protection Maladie du stagiaire à l'étranger :

1) Protection issue du régime étudiant(e) français :

- Pour les stages au sein de l'Espace Economique Européen (EEE) effectués par les étudiant(e)s de nationalité d'un pays membre de l'Union Européenne, l'étudiant doit demander la Carte Européenne d'Assurance Maladie (CEAM).

- Pour les stages effectués au Québec par les étudiant(e)s de nationalité française, l'étudiant doit demander le formulaire SE401Q (104 pour les stages en entreprise, 106 pour les stages en université).

- Dans tous les autres cas de figure :

Les étudiant(e)s qui engagent des frais de santé à l'étranger peuvent être remboursé(e)s auprès de la mutuelle qui leur tient lieu de Caisse de Sécurité Sociale étudiante, au retour, et sur présentation des justificatifs : le remboursement s'effectue alors sur la base des tarifs de soins français, des écarts importants peuvent exister. Il est donc fortement recommandé à l'étudiant(e) de souscrire une assurance Maladie complémentaire spécifique, valable pour le pays et la durée du stage, auprès de l'organisme d'accueil de son choix (mutuelle étudiante, mutuelle des parents, compagnie privée ad hoc...).

Exception : si l'organisme d'accueil fournit à l'étudiant(e) une couverture Maladie en vertu des dispositions du droit local (voir 2 ci-dessous), alors l'étudiant(e) peut choisir de bénéficier de cette protection Maladie locale. Avant d'effectuer un tel choix, il vérifiera l'étendue des garanties proposées.

2) Protection issue de l'organisme d'accueil :

En cochant la case appropriée, l'organisme d'accueil indique ci-après s'il fournit une protection Maladie au stagiaire, en vertu du droit local :

OUI (celle-ci s'ajoute au maintien, à l'étranger, des droits issus du régime français étudiant)

NON (la protection découle alors exclusivement du maintien, à l'étranger, des droits issus du régime français étudiant)

Si aucune case n'est cochée, le 6.3 s'applique.

6.4 Protection Accident du Travail du stagiaire à l'étranger :

1) Pour pouvoir bénéficier de la législation française sur la couverture accident de travail, le présent stage doit :

- Etre d'une durée au plus égale à 6 mois, prolongations incluses.
- Ne donner lieu à aucune rémunération susceptible d'ouvrir des droits à une protection accident de travail dans le pays étranger (une indemnité ou gratification est admise à hauteur de 15 % du plafond horaire de la sécurité sociale pour une durée légale hebdomadaire de 35 heures sous réserve de l'accord de la Caisse Primaire d'Assurance Maladie).
- Se dérouler exclusivement dans l'entreprise partie à la présente convention.
- Se dérouler exclusivement dans le pays étranger cité.

Lorsque les conditions ne sont pas remplies, l'organisme d'accueil s'engage à cotiser pour la protection du stagiaire et à faire les déclarations nécessaires en cas d'accident de travail.

2) La déclaration des accidents de travail incombe à l'établissement qui doit être informé par l'organisme d'accueil par écrit dans un délai de 48 heures.

3) La couverture concerne les accidents survenus :

- Dans l'enceinte du lieu du stage et aux heures de stage.
- Sur le trajet aller-retour habituel entre la résidence du stagiaire sur le territoire étranger et le lieu du stage.
- Sur le trajet aller-retour (début et fin de stage) du domicile du stagiaire situé sur le territoire français et le lieu de résidence à l'étranger.
- Dans le cadre d'une mission confiée par l'organisme d'accueil et obligatoirement par ordre de mission.

4) Pour le cas où l'une seule des conditions prévues au point 6.4 1/ n'est pas remplie, l'organisme d'accueil s'engage par la présente convention à couvrir le stagiaire contre le risque d'accident de travail, de trajet et les maladies professionnelles et à en assurer toutes les déclarations nécessaires.

5) dans tous les cas,

- Si l'étudiant(e) est victime d'un accident du travail durant le stage, l'organisme d'accueil doit impérativement signaler immédiatement cet accident à l'établissement.
- Si l'étudiant(e) remplit des missions limitées en-dehors de l'organisme d'accueil ou en en-dehors du pays du stage, l'organisme d'accueil doit prendre toutes les dispositions nécessaires pour lui fournir les assurances appropriées.

Article 7 : Responsabilité civile et assurances

L'organisme d'accueil et l'étudiant(e) déclarent être garantis au titre de la responsabilité civile. Quelle que soit la nature du stage et le pays de destination, le(la) stagiaire s'engage à se couvrir par un contrat d'assistance (rapatriement sanitaire, assistance juridique etc.) et par un contrat d'assurance individuel accident. Lorsque l'organisme d'accueil met un véhicule à la disposition du(de la) stagiaire, il lui incombe de vérifier préalablement que la police d'assurance du véhicule couvre son utilisation par un étudiant. Lorsque dans le cadre de son stage, l'étudiant(e) utilise son propre véhicule ou un véhicule, prêté par un tiers, il(elle) déclare expressément à l'assureur dudit véhicule cette utilisation qu'il(elle) est amené à faire et le cas échéant s'acquitte de la prime y afférente.

Article 8 : Discipline

Durant son stage, l'étudiant(e) est soumis à la discipline et au règlement intérieur (qui doit être porté à la connaissance de l'étudiant(e)) de l'organisme, notamment en ce qui concerne les horaires, et les règles d'hygiène et de sécurité en vigueur dans l'organisme d'accueil. Toute sanction disciplinaire ne peut être décidée que par l'établissement. Dans ce cas, l'organisme d'accueil informe l'établissement des manquements et lui fournit éventuellement les éléments constitutifs. En cas de manquement particulièrement grave à la discipline, l'organisme d'accueil se réserve le droit de mettre fin au stage de l'étudiant(e) tout en respectant les dispositions fixées à l'article 9 de la présente convention.

Article 9 : Absence et Interruption du stage

Toute difficulté survenue dans le déroulement du stage devra être portée à la connaissance de tous les intéressés afin d'être résolue au plus vite.

En France (sauf en cas de règles particulières applicables dans certaines collectivités d'outre-mer françaises), en organisme de droit privé, en cas de grossesse, de paternité ou d'adoption, le/la stagiaire bénéficie de congés et d'autorisations d'absence d'une durée équivalente à celle prévues pour les salariés dans les organismes de droit privé aux articles L.1225-16 à L.1225-28, L.1225-35, L.1225-46 du code du travail.

Pour les stages dont la durée est supérieure à deux mois et dans la limite de la durée maximale de 6 mois, des congés ou autorisations d'absence sont possibles.

NOMBRE DE JOURS DE CONGES AUTORISES / ou modalités des congés et autorisations d'absence durant le stage :

Pour toute autre interruption temporaire du stage (maladie, absence injustifiée...) l'organisme d'accueil avertit l'établissement d'enseignement par courrier.

Toute interruption du stage, est signalée aux autres parties à la convention et à l'enseignant référent. Une modalité de validation est mise en place le cas échéant par l'établissement d'enseignement supérieur. En cas d'accord des parties à la convention, un report de la fin du stage est possible afin de

permettre la réalisation de la durée totale du stage prévue initialement. Ce report fera l'objet d'un avenant à la convention de stage.

Un avenant à la convention pourra éventuellement être établi en cas de prolongation du stage sur demande conjointe de l'organisme d'accueil et du(de la) stagiaire, dans le respect de la durée maximale du stage fixée par la loi (6 mois).

Interruption définitive :

En cas de volonté d'une des trois parties (organisme d'accueil, établissement, étudiant(e)) d'interrompre définitivement le stage, celle-ci devra immédiatement en informer les deux autres parties par écrit. Les raisons invoquées seront examinées en étroite concertation. La décision définitive d'interruption du stage ne sera prise qu'à l'issue de cette phase de concertation.

Article 10 : Devoir de réserve et confidentialité

Le devoir de réserve est de rigueur absolue. Les étudiant(e)s stagiaires prennent donc l'engagement de n'utiliser en aucun cas les informations recueillies ou obtenues par eux pour en faire l'objet de publication, communication à des tiers sans accord préalable de l'organisme d'accueil, y compris le rapport de stage. Cet engagement vaudra non seulement pour la durée du stage mais également après son expiration. L'étudiant(e) s'engage à ne conserver, emporter, ou prendre copie d'aucun document ou logiciel, de quelque nature que ce soit, appartenant à l'organisme d'accueil, sauf accord de ce dernier.

Nota : Dans le cadre de la confidentialité des informations contenues dans le rapport, l'organisme d'accueil peut demander une restriction de la diffusion du rapport, voire le retrait de certains éléments très confidentiels.

Les personnes amenées à en connaître sont contraintes par le secret professionnel à n'utiliser ni ne divulguer les informations du rapport.

Article 11 : Propriété intellectuelle

Conformément au code de la propriété intellectuelle, si le travail du stagiaire donne lieu à la création d'une œuvre protégée par le droit d'auteur ou la propriété industrielle (y compris un logiciel), si l'organisme d'accueil souhaite l'utiliser et que le stagiaire est d'accord, un contrat devra être signé entre le stagiaire (auteur) et l'organisme d'accueil. Devront notamment être précisés l'étendue des droits cédés, l'éventuelle exclusivité, la destination, les supports utilisés et la durée de la cession, ainsi que, le cas échéant, le montant de la rémunération due à l'étudiant au titre de la cession. Cette clause s'applique également dans le cas des stages dans les Organismes publics.

Article 12 : Recrutement

S'il advenait qu'un contrat de travail prenant effet avant la date de fin du stage soit signé avec l'organisme d'accueil, la présente convention deviendrait caduque ; l'« étudiant(e) » ne relèverait plus de la responsabilité de l'établissement d'enseignement. Ce dernier devrait impérativement en être averti avant la signature du contrat.

Article 13 : Fin de stage – Rapport – Evaluation

A l'issue du stage, l'organisme d'accueil délivre au stagiaire une attestation de stage et remplit une fiche d'évaluation de l'activité du stagiaire mentionnant au minimum la durée effective du stage et, le cas échéant le montant de la gratification perçue qu'il retourne à l'établissement d'enseignement supérieur.

Le(la) stagiaire devra produire cette attestation à l'appui de sa demande éventuelle d'ouverture de droits au régime général d'assurance vieillesse prévue à l'art. L.351-17 du code de la sécurité sociale ;

A l'issue du stage, les parties à la présente convention sont invitées à formuler une appréciation sur la qualité du stage.

Le(la) stagiaire transmet au service compétent de l'établissement d'enseignement un document dans lequel il(elle) évalue la qualité de l'accueil dont il(elle) a bénéficié au sein de l'organisme d'accueil. Ce document n'est pas pris en compte dans son évaluation ou dans l'obtention du diplôme ou de la certification

A l'issue de son stage l'étudiant devra : (préciser la nature du travail à fournir éventuellement en joignant une annexe)

Préciser le cas échéant les modalités de validation du stage :

Nombre de crédits ECTS : 6

Le tuteur organisme d'accueil ou tout autre membre de l'organisme d'accueil appelé à se rendre à l'établissement dans le cadre de la préparation, du déroulement et de la validation du stage ne peut prétendre à une quelconque prise en charge ou indemnisation de la part de l'établissement.

Un avenant à la convention pourra éventuellement être établi en cas de prolongation de stage faite à la demande de l'organisme et de l'étudiant(e). En aucun cas la date de fin de stage ne pourra être postérieure au 30/09 de l'année en cours.

L'accueil successif de stagiaires, au titre de conventions de stage différentes, pour effectuer des stages dans un même poste n'est possible qu'à l'expiration d'un délai de carence égal au tiers de la durée du stage précédent. Cette disposition n'est pas applicable lorsque ce stage précédent a été interrompu avant son terme à l'initiative du stagiaire.

Article 14 : Droit applicable – Tribunaux compétents

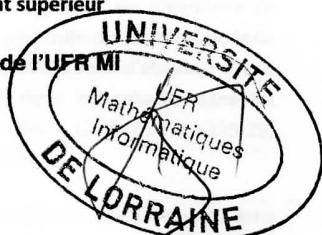
La présente convention est régie exclusivement par le droit français. Tout litige non résolu par voie amiable sera soumis à la compétence de la juridiction française compétente.

A Nancy le 12/04/18

Pour l'établissement d'enseignement supérieur

(nom et signature du représentant)

Antoine TABBONE, directeur de l'UFR MI



Pour l'organisme d'accueil

(nom et signature du représentant)

Le directeur du LORIA

Jean-Yves MARION



Pour l'étudiant

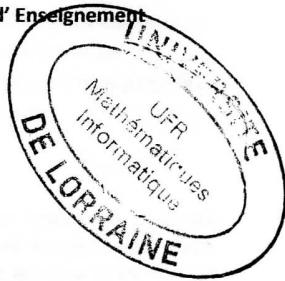
(nom et signature)

MARQUE Escobar

VISAS :

L'enseignant référent de l'Etablissement d'Enseignement Supérieur
(nom et signature)

Laurent THOMANN



Tuteur Organisme d'accueil

(nom et signature)

CERISARA Christophe Cenier

Annexe 1 : Charte des stages / Annexe 2 : Fiches d'évaluation / Annexe 3 à fournir par l'étudiant(e) : Attestation de responsabilité civile



F Copie de l'avenant à la convention de stage

AVENANT A LA
CONVENTION DE STAGE

Signée le 12.10.18

ENTRE

L'établissement d'enseignement supérieur :

Université de Lorraine, établissement public à caractère scientifique, culturel et professionnel, sis 34 Cours Léopold – CS 25233 –

54 052 NANCY Cedex, siret n° 130 015 506 00012, représenté par son Président, Monsieur Pierre Mutzenhardt,

Représenté par : (nom du (de la) signataire de la convention) : Antoine TABBONE

Qualité du représentant : Directeur de l'UFR Mathématiques et Informatique

Composante /UFR/ : UFR Mathématiques et Informatique

Adresse d'envoi de la convention à compléter obligatoirement par la composante :

56 bis, boulevard de Scarpone, 54 000 Nancy

L'organisme d'accueil :

Nom :

LORIA UMR 7503

Adresse : Campus Scientifique BP239, 54 506 Vandoeuvre-les-Nancy

Tél : 03 83 59 20 00 Fax : Mél : direction@loria.fr

Représenté par : (nom du signataire de la convention) : Jean-Yves MARION

Qualité du représentant : Directeur du LORIA

Nom du service dans lequel le stage sera effectué : Equipe SYNALP

Lieu du stage : (si différent de l'adresse de l'entreprise)

Et l'étudiant stagiaire :

Nom : MARQUER Prénom : Esteban

Sexe : F M né(e) le : 08/06/1997

Adresse : 6, rue CYFFÉ, 54 000 Nancy

Tél : Mél :

Intitulé complet de la formation ou du cursus suivi dans l'établissement d'enseignement supérieur et son volume horaire :

Licence L3 MIASHS : MIAGE / Sciences Cognitives Volume horaire : 600h de présence/an

SUJET DE STAGE :

Interpréter les couches cachées des réseaux récurrents en TAL

DATES DE STAGE : Du 27/04/2018 Au 03/08/2018

DUREE DU STAGE * : 1 Heures ou Semaines ou Mois (rayer la mention inutile) soit en JOURS : 5

*7 heures = 1 jour et 22 jours = 1 mois

Encadrement du stagiaire assuré par :

L'établissement d'enseignement supérieur en la personne de :

Nom : THOMANN

Prénom : Laurent

Fonction : Professeur, Responsable des Stages

Tél : 03 72 74 16 24

Mél : laurent.thomann@univ-lorraine.fr

L'organisme d'accueil en la personne de :

Nom : CERISARA

Prénom : Christophe

Fonction : Responsable équipe SYNALP

Tél : 03 54 95 86 25

Mél : cerisara@loria.fr

PREAMBULE

Vu la loi n°2014-788 du 10 juillet 2014 tendant au développement, à l'encadrement des stages et à l'amélioration du statut des stagiaires,

Vu le décret n°2014-1420 du 27 novembre 2014 relatif à l'encadrement des périodes de formation en milieu professionnel et des stages,

Les parties ont signé une convention de stage en date du 12/04/2015. Elles souhaitent aujourd'hui préciser les modifications suivantes.

Article 1 – Modification de l'article 5 :

L'alinéa 2 de article est modifié comme suit :

Article 5 – Gratification - Avantages

(...)

Le montant horaire de la gratification est fixé à 15 % du plafond horaire de la sécurité sociale défini en application de l'article L.241-3 du code de la sécurité sociale. Une convention de branche ou un accord professionnel peut définir un montant supérieur à ce taux.

(...)

Article 2 – Modification de l'article 6 :

La présente convention règle les rapports de l'organisme d'accueil avec l'établissement d'enseignement et le/la stagiaire.

L'article est modifié comme suit :

Article 6 – Régime de protection sociale

(...)

6.1 Gratification inférieure ou égale à 15 % du plafond horaire de la sécurité sociale :

(...)

6.2 – Gratification supérieure à 15 % du plafond horaire de la sécurité sociale :

Les cotisations sociales sont calculées sur le différentiel entre le montant de la gratification et 15 % du plafond horaire de la Sécurité Sociale.

(...)

6.4 Protection Accident du Travail du stagiaire à l'étranger

I) Pour pouvoir bénéficier de la législation française sur la couverture accident de travail, le présent stage doit :

(...)

- ne donner lieu à aucune rémunération susceptible d'ouvrir des droits à une protection accident de travail dans le pays d'accueil ; une indemnité ou gratification est admise dans la limite de 15 % du plafond horaire de la sécurité sociale (cf point 5), et sous réserve de l'accord de la Caisse Primaire d'Assurance Maladie ;

Article 3 – Modification de l'article 9 :

La présente convention règle les rapports de l'organisme d'accueil avec l'établissement d'enseignement et le/la stagiaire.

L'article est modifié comme suit :

Article 9 – Congés – Interruption du stage

(...)

Pour toute autre interruption temporaire du stage (maladie, absence injustifiée...) l'organisme d'accueil avertit l'établissement d'enseignement par courrier.

Toute interruption du stage, est signalée aux autres parties à la convention et à l'enseignant référent. Une modalité de validation est mise en place le cas échéant par l'établissement d'enseignement supérieur. En cas d'accord des parties à la convention, un report de la fin du stage est possible afin de permettre la réalisation de la durée totale du stage prévue initialement. Ce report fera l'objet d'un avenant à la convention de stage.

(...)

Article 4 –Dispositions finales :

Les autres dispositions de la convention de stage initiales restent inchangées.

Fait à Nancy Le _____

POUR L'ÉTABLISSEMENT D'ENSEIGNEMENT

Nom et signature du représentant de l'établissement

STAGIAIRE (OU SON REPRESENTANT LEGAL LE CAS ÉCHÉANT)

Nom et signature

MARION YVES STEPHAN

POUR L'ORGANISME D'ACCUEIL

Nom et signature du représentant de l'organisme d'accueil

Le directeur du LORIA

L'enseignant référent du stagiaire

Nom et signature

Jean-Yves MARION

Le tuteur de stage de l'organisme d'accueil

Nom et signature

CERISARA Christophe
Perison