

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

Encadrant universitaire : Jeanine Souquières

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 M. Samuel Cruz-Lara, Mme. ??? et Mme. ??? 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.

Avant-propos

La lecture du présent mémoire ne nécessite aucune connaissance préalable en Apprentissage Automatique ou en Apprentissage Profond.

Cependant, de nombreux termes techniques sont utilisés. La plupart est abrégée ou sous forme de sigle.

D'une part, ces abréviation sont habituellement en anglais dans la littérature du domaine. Aussi, pour maintenir la lisibilité pour les lecteurs initiés au domaine, les abréviations utilisées seront sous leur forme anglaise. Pour maintenir une cohérence dans les termes utilisés, les expression dont les versions anglaises et françaises Enfin, les termes techniques seront toujours introduits en français, accompagnées d'une explication, de la traduction anglaise et du sigle anglais. Par exemple : Exemple de Terme Technique (*Technical Expression Example* en anglais, TEE).

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

Table des matières

Remerciements	iii
Avant-propos	v
1 Introduction	3
1.1 Contexte & Enjeux	3
1.2 Objectifs	3
1.3 Plan	4
2 Cadre théorique	5
2.1 Introduction	5
2.2 Éléments théoriques généraux	5
2.3 Concepts et termes fondamentaux	6
2.4 État de l'art	6
3 Présentation du laboratoire, de l'équipe et des collaborateurs	7
3.1 Présentation du laboratoire et de l'équipe	7
3.2 Présentation du projet ITEA3-PAPUD, cas d'utilisation BULL	9
4 Projet	13
4.1 Architecture innovante de réseau de neurones pour l'élaboration de modèle du langage	13
4.2 Projet PAPUD	15
5 Réalisation	17
5.1 Projet GMSNN	17
5.2 Projet PAPUD	17
6 Conclusion	21
Annexes	23

1 Introduction

1.1 Contexte et enjeux du stage

D'une part, depuis quelques années, les "réseaux de neurones" ont connu une explosion de popularité. Ce qui se cache derrière cette hype est la combinaison de théories relativement anciennes et de avancées technologiques permettant la mise en œuvre desdites théories. Un des intérêts de ces méthodes d'Apprentissage Automatique (*Machine Learning* en anglais) est la capacité à apprendre à partir de données restreintes. Un des domaines exploitant les performances de ces outils est le TAL, au vu de ce et de ça. Une application au TAL de l'apprentissage différentiel est le LM qui nous intéressera particulièrement dans ce mémoire.

D'autre part, nous sommes à l'aire du "Big Data". Cela implique que les quantités de données exploitables produites de nos jours est très grande. Dans ce rapport, nous nous intéresseront à des volumes de données bien au-delà des volumes habituellement utilisés dans le domaine.

Ainsi, nous nous explorerons la question de l'application des méthodes du DL 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 NN 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 ou Réseau de Neurones (*Neural Network* en anglais) innovante, imaginée par notre maître de stage Mr Cerisara. Par la même occasion, ce stage est l'opportunité pour moi d'apprendre à manipuler les Réseaux de Neurones Artificiels.

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 (*Profiling and Analysis Platform Using Deep Learning*).

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

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.

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 de ce rapport.

Dans un premier temps, le domaine scientifique dans lequel s'est déroulé le stage sera expliqué. Dans un second temps, les concepts et termes fondamentaux utilisés dans ce mémoire seront définis. 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 Éléments théoriques généraux

Ce stage s'inscrit dans deux principaux domaines :

- l'apprentissage profond (*Deep Learning* en anglais);
- le Traitement Automatique des Langues (*Natural Language Processing* en anglais, NLP).

2.2.1 Apprentissage profond

L'apprentissage profond représente un ensemble de techniques de apprentissage automatique, basés sur ce que l'on appelle des Réseaux de Neurones.

L'apprentissage automatique (*Machine Learning* en anglais) est un ensemble de « méthodes [statistiques] permettant à une machine (au sens large) d'évoluer par un processus systématique, et ainsi de remplir des tâches difficiles ou problématiques par des moyens algorithmiques plus classiques ». D'après Wikipédia *Apprentissage automatique*. Wikipédia. Juin 2018. URL : https://fr.wikipedia.org/wiki/Apprentissage_automatique (visité le 10/08/2018).

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.

2.2.2 NLP

2.3 Concepts et termes fondamentaux

Réseau de Neurones Artificiels Un Réseau de Neurones Artificiels ou Réseau de Neurones (*Neural Network* en anglais) est

Apprentissage Profond (*Deep Learning* en anglais) 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, le modèle correspond au réseau de neurone.

2.3.1 LM

2.4 État de l'art

2.4.1 Projet GMSNN

2.4.2 Projet PAPUD

-> next parts

3 Présentation du laboratoire, de l'équipe et des collaborateurs

3.1 Présentation du laboratoire et de l'équipe

3.1.1 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.

3.1.2 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 :¹

- **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)**.

1. *Organisation.* LORIA. Août 2018. URL : <http://www.loria.fr/fr/presentation/organisation/> (visité le 03/08/2018).

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.

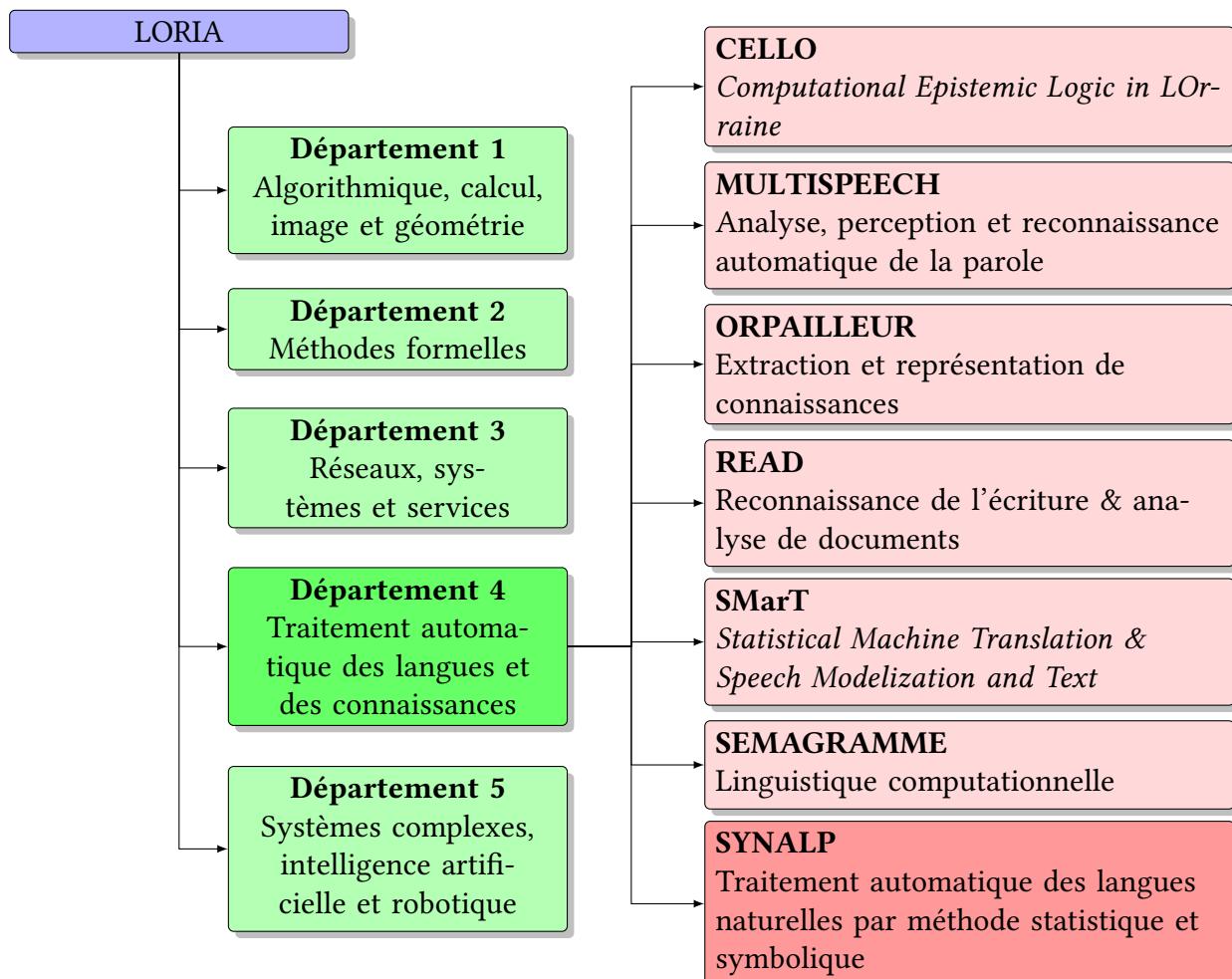


FIGURE 3.1 – Organigramme des départements du LORIA, et des équipes du département 4

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 laboratoire²

3.1.3 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 (NLP) et des connaissances.

2. Organigramme 2017 du Loria. LORIA. Août 2018. url : <http://www.loria.fr/wp-content/uploads/2016/05/organigramme-2017.pdf> (visité le 03/08/2018).

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 NLP, ainsi que sur les applications de ces approches.

Ainsi, les principaux sujets de recherche de l'équipe sont les Modèle de la Langue (*Language Model* en anglais, LM), les grammaires formelles, la sémantique computationnelle, le traitement de la parole, et les outils et ressources utilisés en NLP.

Ce stage s'inscrit en particulier dans la réalisation de LM, et l'élaboration d'outils et ressources utilisés en NLP. 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 laboratoire³. 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)⁴.

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

La seconde partie du stage s'intègre dans le projet ITEA3-PAPUD, en particulier dans le cas d'utilisation BULL. Nous verrons en détail les objectifs du projet dans la section 4.2 (page 15).

Le projet PAPUD (*Profiling and Analysis Platform Using Deep Learning*) est un projet de l'initiative ITEA (*Information Technology for European Advancement*) du réseau EUREKA.

3.2.1 EUREKA et ITEA3

« EUREKA est une initiative européenne, intergouvernementale, destinée à renforcer la compétitivité de l'industrie européenne. » D'après Wikipedia.⁵

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

3. Le Loria. LORIA. Juil. 2018. URL : <http://www.loria.fr> (visité le 31/07/2018).

4. About SYNALP. (en anglais). SYNALP. Juil. 2018. URL : <http://synalp.loria.fr/pages/about-synalp> (visité le 30/07/2018).

5. EUREKA. Wikipédia. Août 2018. URL : <https://fr.wikipedia.org/wiki/EUREKA> (visité le 01/08/2018).

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.

3.2.2 Projet PAPUD et cas d'utilisation BULL

C'est lors de la troisième vague d'appels à projets d'ITEA3 que le projet PAPUD a 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 3.1 – Informations générales sur le projet PAPUD, d'après le site de ITEA3⁶

3.2.3 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'ATOS⁷, 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 ;

7. *Produits*. Atos. Août 2018. URL : <https://atos.net/fr/produits> (visité le 01/08/2018).

- 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.⁸
Les

8. 16037 PAPUD. (en anglais). itea3.org. Juil. 2018. url : <https://itea3.org/project/papud.html> (visité le 31/07/2018).

4 Projet

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

4.1.1 Contexte

Nous avons vus dans le chapitre 2 les techniques générales du apprentissage profond

Nous nous plaçons ici dans le contexte de la réalisations de LM. Les modèles actuels, généralement basés sur les Réseau de Neurones Récursifs (*Recurrent Neural Network* en anglais, 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.

Cependant, peu de ces modèles profitent réellement de l'abondance des données disponibles. deux éléments responsables : lourd pré-traitement nécessaire, * plus de données ==> très très lourd peu ou pas d'amélioration de performance, pour plus de temps pour entraîner -> inutile

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). Par exemple, un RNN classique conserve en mémoire des informations datant d'au plus 20 entrées auparavant ; un Réseau récurrent à porte (*Gated Recurrent Unit* en anglais, GRU) peut se rappel d'informations vieilles de plus de 100 entrées ; et un Réseau récurrent à mémoire à court et long terme (*Long Short Term Memory* en anglais, LSTM) dépasse difficilement les 200 entrées.

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.

Solution

L'architecture proposée

Passer à l'échelle

L'élément

Adapter le modèle à un volume potentiellement infini de données

Objectif secondaire

Du char-lstm au char-gmsnn

Par la suite, nous désignerons ce projet par « projet Réseau de Neurones Récursifs Multi-Échelles Croissant (*Growing Multi-Scale Recurrent Neural Network* en anglais, GMSNN) ».

4.2 Projet PAPUD

Le projet ITEA3-PAPUD, cas d'utilisation BULL est comme présenté précédemment, un pr

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.

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

5 Réalisation

5.1 Projet GMSNN

5.1.1 Recherche documentaire

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

5.1.3 Implémentation du nouveau modèle basique

5.1.4 Optimisation et amélioration du nouveau modèle

5.1.5 Conclusions

5.2 Projet PAPUD

5.2.1 Recherche documentaire

5.2.2 Ré-implémentation simplifiée du modèle état de l'art

5.2.3 Implémentation du nouveau modèle basique

5.2.4 Optimisation et amélioration du nouveau modèle

5.2.5 Conclusions

Il est ais  d'ins rer du code dans un rapport. Il suffit de d finir le langage, la l gende   afficher et enfin un Label pour pouvoir y faire r f rence. Le r sultat est donn e dans le listing 5.1. Il est  g galement possible de changer les couleurs, pour cela il faut  diter le lstset dans la classe tnreport.cls.

```
1 void CEquation :: IniParser ()  
2 {  
3     if (!pP){ // if not already initialized ...  
4         pP = new mu::Parser;  
5     }
```

```

6 pP->DefineOprt( "%" , CEquation :: Mod, 6); // deprecated
7 pP->DefineFun( "mod" , &CEquation :: Mod, false );
8 pP->DefineOprt( "&" , AND, 1); //DEPRECATED
9 pP->DefineOprt( "and" , AND, 1);
10 pP->DefineOprt( " | " , OR, 1); //DEPRECATED
11 pP->DefineOprt( "or" , OR, 1);
12 pP->DefineOprt( "xor" , XOR, 1);
13 pP->DefineInfixOprt( "!" , NOT);
14 pP->DefineFun( " floor " , &CEquation :: Floor , false );
15 pP->DefineFun( " ceil " , &CEquation :: Ceil , false );
16 pP->DefineFun( " abs " , &CEquation :: Abs , false );
17 pP->DefineFun( " rand " , &CEquation :: Rand , false );
18 pP->DefineFun( " tex " , &CEquation :: Tex , false );
19
20 pP->DefineVar( "x" , &XVar );
21 pP->DefineVar( "y" , &YVar );
22 pP->DefineVar( "z" , &ZVar );
23 }
24 }
```

Fragmet de code 5.1 – Premier Exemple

Il est également possible d'afficher du code directement depuis un fichier source, le résultat de cette opération est visible dans le listing ??

6 Conclusion

Annexes

Table des annexes

A Bibliographie / Webographie	27
B Listes des tables, des figures et des fragments de code	29
C Glossaire, acronymes et noms d'entités	31
D Rapports d'avancement du projet GMSNN	35
D.1 Informations sur les rapports contenus dans la présente annexe	35
D.2 Étude des problèmes de mémoire	36
D.3 Sauvegarde du modèle	41
D.4 Analysis and implementation of job save and restart	41
D.5 Solution tentative pour les fuites mémoires	45
D.6 Problèmes théoriques liés à l'entraînement batch par batch	47
D.7 Tentative de réduction du temps de calcul par utilisation de l'algorithme d'entraînement « <i>Truncated BPTT</i> »	50
D.8 Entraînement couche par couche	52
D.9 Premier test du modèle réimplémenté	54
D.10 Premier entraînement complet du modèle réimplémenté	55
D.11 Premier entraînement à 50 époques du modèle réimplémenté	57
D.12 Premier test du modèle multi-échelles	61
D.13 Comparaison des stratégies de fusion des résultats des différentes couches	65
D.14 Test des effets du changement de taille des paquets (<i>batch</i>)	70
D.15 Entraînement sur le corpus complet avec beaucoup de temps alloué	72
E Rapports d'avancement du projet PAPUD	79

E.1	Informations sur les documents contenus dans la présente annexe	79
E.2	Informations générales	80
E.3	Résultats de l'implémentation basique	82
E.4	Paquets (<i>batchs</i>) simultanés	84
E.5	Analyse du pic de performance	88
E.6	Rapport de la réunion avec les autres membres du projet	92
E.7	Optimisation du taux d'apprentissage	94
E.8	Effets de l'optimisation du taux d'apprentissage	98
E.9	Taille de <i>batch</i> binaire	100
E.10	Performances du lecteru de corpus multi-fichiers multi-processus	102
F	Copie de la convention de stage	107
G	Copie de l'avenant à la convention de stage	113

A Bibliographie / Webographie

- [1] *Cadre Théorique d'un Mémoire - Contenu et Exemple*. Scribbr. Fév. 2016. URL : <https://www.scribbr.fr/memoire/cadre-theorique-dun-memoire> (visité le 02/08/2018).
- [2] *Mémoire, rapport de stage : les règles de présentation*. l'Étudiant. Août 2018. URL : <https://www.letudiant.fr/jobsstages/nos-conseils/memoire-rapport-de-stage-les-regles-de-presentation.html> (visité le 03/08/2018).
- [3] *Apprentissage automatique*. Wikipédia. Juin 2018. URL : https://fr.wikipedia.org/wiki/Apprentissage_automatique (visité le 10/08/2018) (cf. p. 5, 31).
- [4] *Organisation*. LORIA. Août 2018. URL : <http://www.loria.fr/fr/presentation/organisation/> (visité le 03/08/2018) (cf. p. 7).
- [5] *Organigramme 2017 du Loria*. LORIA. Août 2018. URL : <http://www.loria.fr/wp-content/uploads/2016/05/organigramme-2017.pdf> (visité le 03/08/2018) (cf. p. 8).
- [6] *Le Loria*. LORIA. Juil. 2018. URL : <http://www.loria.fr> (visité le 31/07/2018) (cf. p. 9).
- [7] *About SYNALP*. (en anglais). SYNALP. Juil. 2018. URL : <http://synalp.loria.fr/pages/about-synalp> (visité le 30/07/2018) (cf. p. 9).
- [8] *EUREKA*. Wikipédia. Août 2018. URL : <https://fr.wikipedia.org/wiki/EUREKA> (visité le 01/08/2018) (cf. p. 9, 34).
- [9] *16037PAPUD*. (en anglais). itea3.org. Juil. 2018. URL : <https://itea3.org/project/papud.html> (visité le 31/07/2018) (cf. p. 11).
- [10] *Produits*. Atos. Août 2018. URL : <https://atos.net/fr/produits> (visité le 01/08/2018) (cf. p. 10).
- [11] *Language model*. en anglais. Wikipédia. Juil. 2018. URL : https://en.wikipedia.org/wiki/Language_model (visité le 10/08/2018) (cf. p. 31).
- [12] *Markdown*. Wikipédia. Juil. 2018. URL : <https://fr.wikipedia.org/wiki/Markdown> (visité le 10/08/2018) (cf. p. 32).
- [13] *Réseau de neurones récurrents*. Wikipédia. Août 2018. URL : https://fr.wikipedia.org/wiki/R%C3%A9seau_de_neurones_r%C3%A9currents (visité le 10/08/2018) (cf. p. 32).
- [14] Jason BROWNLEE. « A Gentle Introduction to Backpropagation Through Time ». en anglais. In : *Machine Learning Mastery* (juin 2017). URL : <https://machinelearningmastery.com/gentle-introduction-backpropagation-time> (visité le 06/07/2018) (cf. p. 50).

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

Liste des tableaux

3.1	Informations générales sur le projet PAPUD, d'après le site de ITEA3 ¹	10
-----	---	----

Liste des illustrations

3.1	Organigramme des départements du LORIA, et des équipes du département	4 . . . 8
-----	---	-----------

Liste des fragments de code

5.1	Premier Exemple 17
-----	-----------------	--------------

1. 16037 PAPUD. (en anglais). itea3.org. Juil. 2018. URL : <https://itea3.org/project/papud.html> (visité le 31/07/2018).

C Glossaire, acronymes et noms d'entités

Glossaire

apprentissage profond L'apprentissage profond (*Deep Learning* en anglais), est une méthode d'apprentissage automatique utilisant la technique des réseaux de neurones. *voir* apprentissage automatique & Réseau de Neurones Artificiels, 5, 6, 10, 13

apprentissage automatique L'apprentissage automatique (*Machine Learning* en anglais) est un ensemble de « méthodes [statistiques] permettant à une machine (au sens large) d'évoluer par un processus systématique, et ainsi de remplir des tâches difficiles ou problématiques par des moyens algorithmiques plus classiques ». D'après Wikipédia *Apprentissage automatique*. *voir* PAPUD & ITEA3, 3, 5

Char-GMSNN-LM Un Modèle du Langage au niveau du Caractère basé sur un Réseau de Neurones Multi-Échelles Croissant. C'est un modèle du langage basé sur un réseau de neurones artificiels. Ce réseau de neurones est un GMSNN utilisé comme Char-LM. *voir* Char-LM & GMSNN

Char-LM Un Modèle de la Langue au niveau du Caractère (*Character-level Language Model* en anglais) est un Modèle de la Langue qui prédit non pas le prochain mot à partir des mots précédents, mais le prochain caractère à partir des caractères précédents *voir* LM

cloud 10

data-centers 10

état de l'art 17

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 see 35, 79

GMSNN (le nombre de couches augmente selon le nombre d'entrées) *voir* MSNN

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

grammaires formelles 9

GRU *voir* RNN, 13

LM 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. Autrement dit, c'est une représentation servant à prédire le prochain mot à partir des mots précédents. D'après Wikipédia *Language model*. en anglais. Wikipédia. Juil.

2018. URL : https://en.wikipedia.org/wiki/Language_model (visité le 10/08/2018). 9

LSTM voir RNN, 13

Markdown « Markdown est un langage de balisage léger créé en 2004 par John Gruber avec Aaron Swartz. Son 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. »¹ 35

matrice 31

modèle Un modèle en apprentissage automatique est la représentation du monde construite lors de l'apprentissage afin de répondre au problème à résoudre. voir apprentissage automatique, 6

MSNN voir RNN

NLP Le Traitement Automatique des Langues (*Natural Language Processing* en anglais, NLP) est une discipline qui s'intéresse au traitement des informations langagières par des moyens formels ou informatiques 5

Réseau de Neurones Un Réseau de Neurones Artificiels ou Réseau de Neurones (*Neural Network* en anglais) est voir apprentissage automatique, 3, 4, 5, 13

RNN Un Réseau de Neurones Artificiels Récursifs, plus simplement Réseau de Neurones Récursifs (*Recurrent Neural Network* 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. *Réseau de neurones récurrents*. Wikipédia. Août 2018. URL : https://fr.wikipedia.org/wiki/R%C3%A9seau_de_neurones_r%C3%A9currents (visité le 10/08/2018) voir Réseau de Neurones Artificiels, 13

sémantique computationnelle 9

traitement de la parole 9

1. *Markdown*. Wikipédia. Juil. 2018. URL : <https://fr.wikipedia.org/wiki/Markdown> (visité le 10/08/2018).

Acronymes

Char-GMSNN-LM *Glossaire* : Char-GMSNN-LM

Char-LM *Glossaire* : Char-LM

GMSNN *Glossaire* : GMSNN

GPU *Glossaire* : GPU

GRU *Glossaire* : GRU, 13

LM *Glossaire* : LM, 9, 13

LSTM *Glossaire* : LSTM, 13

MSNN *Glossaire* : MSNN

NLP *Glossaire* : NLP, 5, 8, 9

RNN *Glossaire* : RNN, 13

Entités et sigles

AREQ Assemblée des Responsables des Équipes 7

ATOS 10

BULL 9, 10, 11, 15

CNRS Centre National de la Recherche Scientifique 7, 8

EUREKA « EUREKA est une initiative européenne, intergouvernementale, destinée à renforcer la compétitivité de l'industrie européenne. » D'après Wikipedia.² 9, 34

Gitlab Gitlab 35

INRIA Institut National de Recherche en Informatique et en Automatique 7

ITEA3 troisième instance d'ITEA (*Information Technology for European Advancement*), une initiative de recherche, développement et innovation du réseau EUREKA. voir EUREKA, 9, 10, 15, 29

LORIA Laboratoire Lorrain d'Informatique et ses Applications 7, 8, 9, 29

PAPUD *Profiling and Analysis Platform Using Deep Learning* 3, 4, 9, 10, 11, 15, 29

projet GMSNN projet basé sur une proposition innovante d'architecture de réseau de neurones, faite par M. Christophe Cerisara voir PAPUD & ITEA3, 14, 35

projet PAPUD projet ITEA3-PAPUD, cas d'utilisation BULL voir PAPUD & ITEA3, 1, 15, 79

SYNALP SYNALP (*SYmbolic and statistical NATural Language Processing*) est une équipe de recherche du département 4 du LORIA voir LORIA, 4, 7, 8, 9

Université de Lorraine Université de Lorraine 7, 8

2. EUREKA.

D Rapports d'avancement du projet GMSNN

D.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.

D.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.

D.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.

D.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.

D.2 Étude des problèmes de mémoire

Analysis of memory usage

Analysis report

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

D.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.

D.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

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

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

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

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

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

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

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   (
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)
```

D.3 Sauvegarde du modèle

D.4 Analysis and implementation of job save and restart

Implementation report

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

D.4.1 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.

D.4.2 End solution : saving using pytorch utilities

The saving solution currently implemented is the `torch.save(trainer, file)` and `trainer = torch.load(file)` and the alternative 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.

D.4.3 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)
```

D.4.4 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

D.4.5 Partial save

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

D.4.6 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.

D.4.7 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.msnn_backup = self.model.msnn.backup()
8
9     def __exit__(self, type, value, traceback):
10        # Restoring training linked data
11        self.model.msnn.restore(self.msnn_backup)
12
13
14 with BackupContextManager(model):
15     """Do computations"""
```

D.5 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

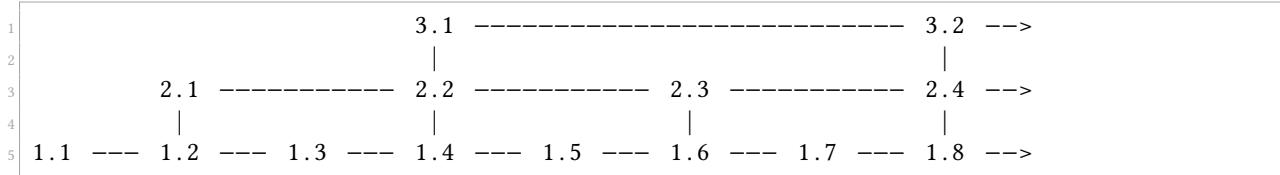
D.5.1 Abstract

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

D.5.2 Problem

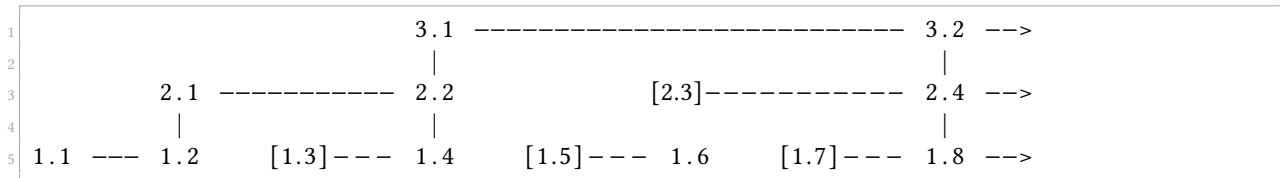
It seems simply detaching a hidden state is not enough :

With a graph :

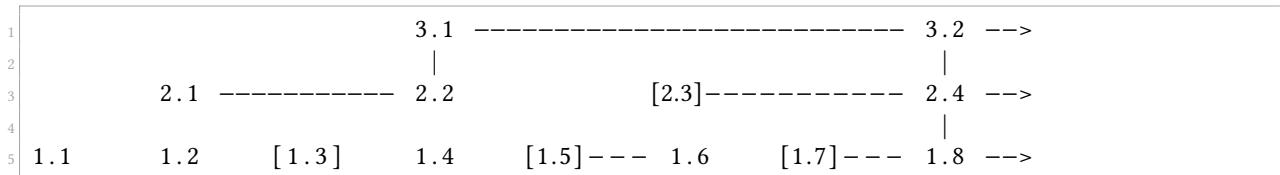


where all nodes are hidden states, and each layer is a line, with a transmission rate of 1 transmission every 2 hidden state.

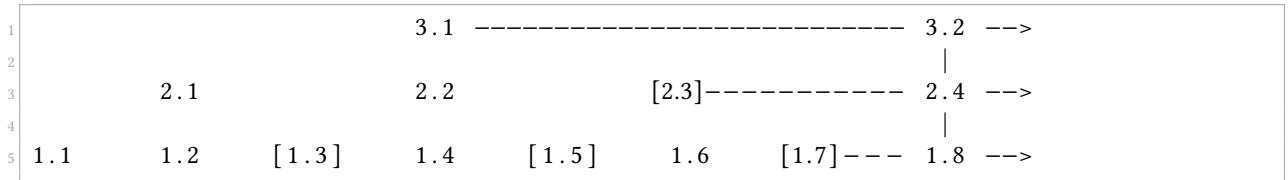
When detaching every 2 hidden, with [i,j] a detached node, the graph becomes :



But it should be :



Or with a more aggressive strategy :



D.5.3 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

D.6 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

D.6.1 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.

D.6.2 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

D.6.3 Analysis of the different strategies

Spreading the corpus along the batch dimension

The worst possible strategies seem to be D.6.2spreading 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 D.6.2, 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.

D.6.4 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.

D.7 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

D.7.1 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).

D.7.2 Notations

The following notations are from an introduction article on BPTT :¹

- **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.

1. Jason BROWNLEE. « A Gentle Introduction to Backpropagation Through Time ». en anglais. In : *Machine Learning Mastery* (juin 2017). URL : <https://machinelearningmastery.com/gentle-introduction-backpropagation-time> (visité le 06/07/2018).

D.7.3 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

D.7.4 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()
```

D.8 Entrainement couche par couche

Layer by layer training

Analysis report

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

D.8.1 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.

D.8.2 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).

D.9 Premier test du modèle réimplémenté

Test run of `detrnn.py`

Test report

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

D.9.1 Abstract

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

D.9.2 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

D.9.3 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](#)

D.9.4 Potential ameliorations & next steps

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

D.10 Premier entraînement complet du modèle réimplémenté

Test run of `detrnn.py`

Test report

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

D.10.1 Abstract

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

D.10.2 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

D.10.3 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

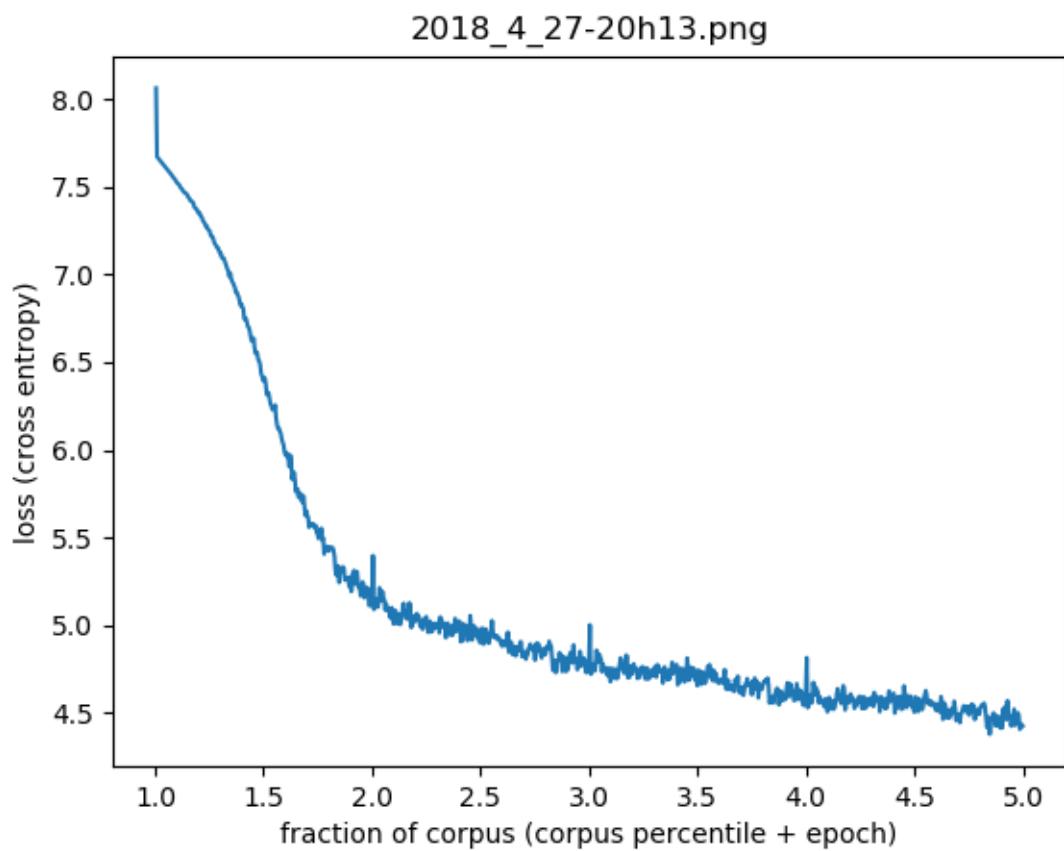


FIGURE D.1 – plot

Details

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

D.10.4 Potential ameliorations & next steps

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

D.11 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

D.11.1 Abstract

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

D.11.2 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

D.11.3 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).

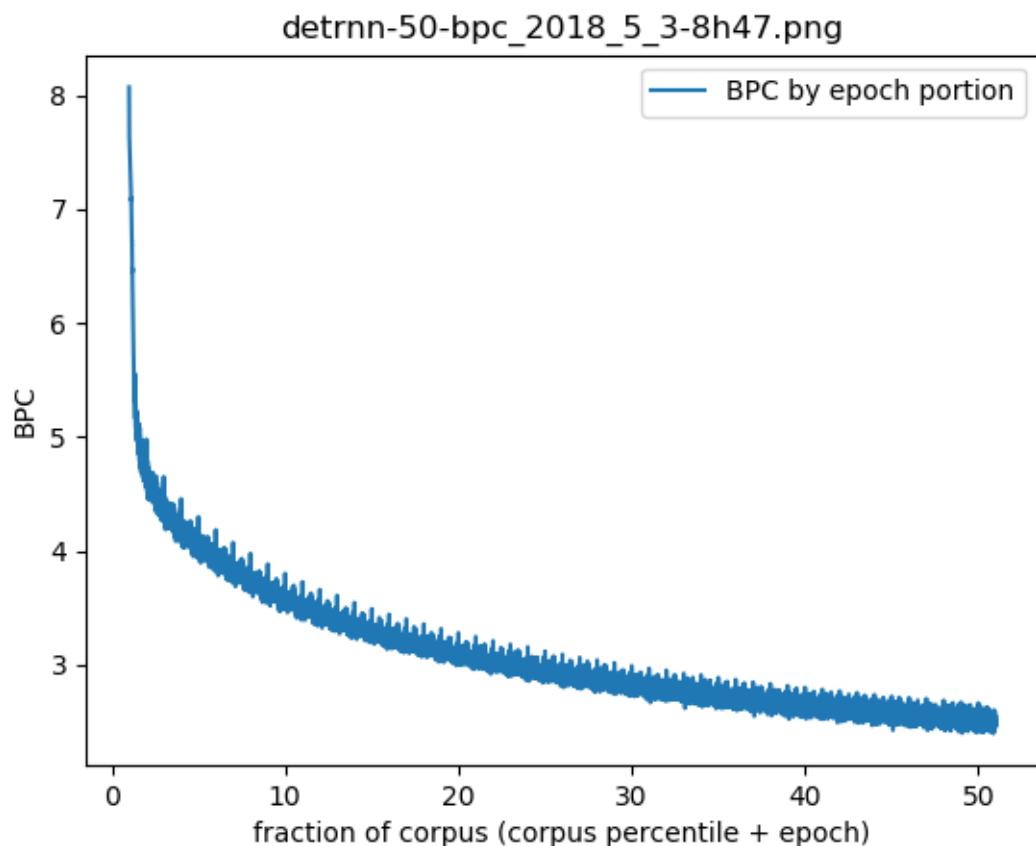


FIGURE D.2 – BPC

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

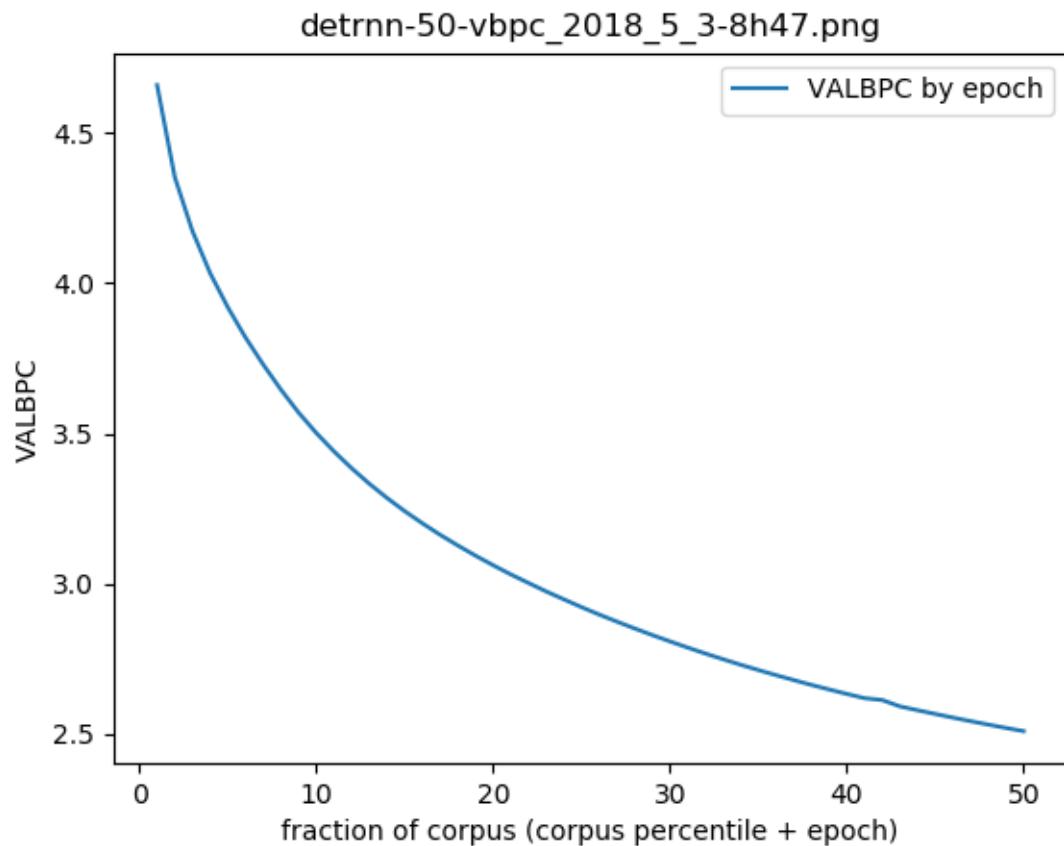


FIGURE D.3 – ValBPC

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

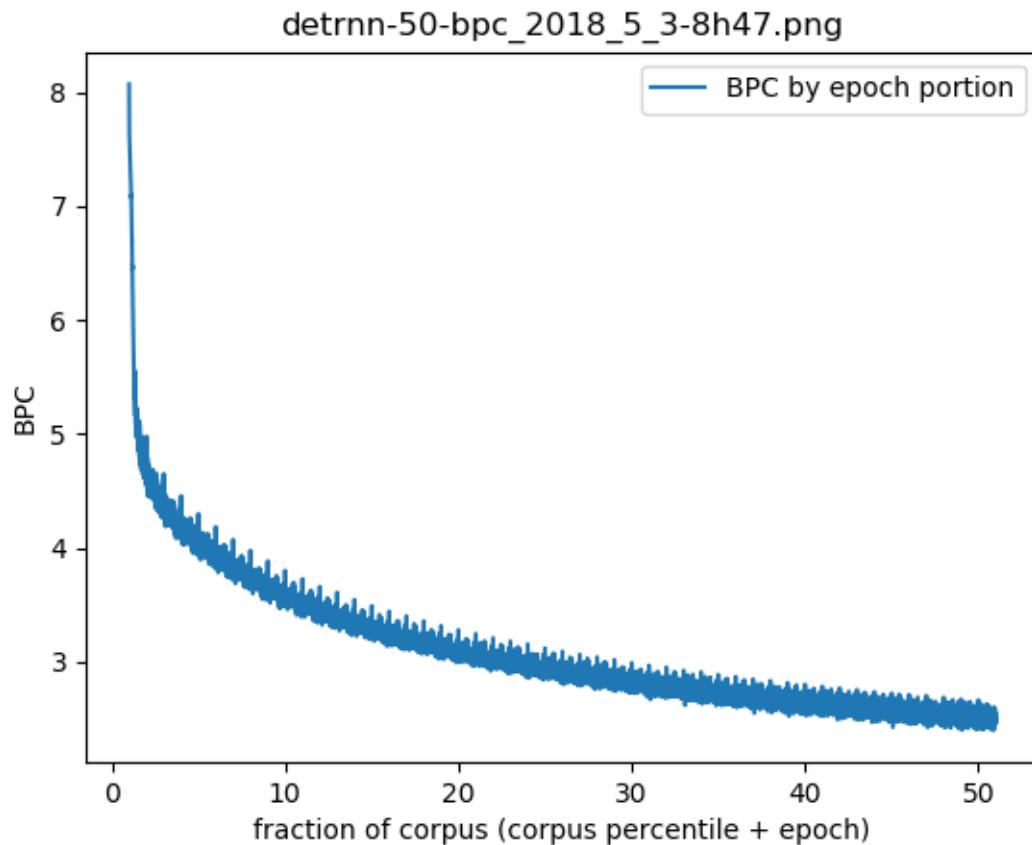


FIGURE D.4 – Loss

Logs

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

D.11.4 Potential ameliorations & next steps

Next step is to test the growing model.

D.12 Premier test du modèle multi-échelles

Test run of `detrnn.py`

Test report

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

D.12.1 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.

D.12.2 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 :

$$1 \quad (10 \text{ min} / 0.2 \text{ epoch}) * 4 \text{ epoch} = 200 \text{ min} = 3 \text{ h } 20 \text{ min}$$

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
```

D.12.3 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.

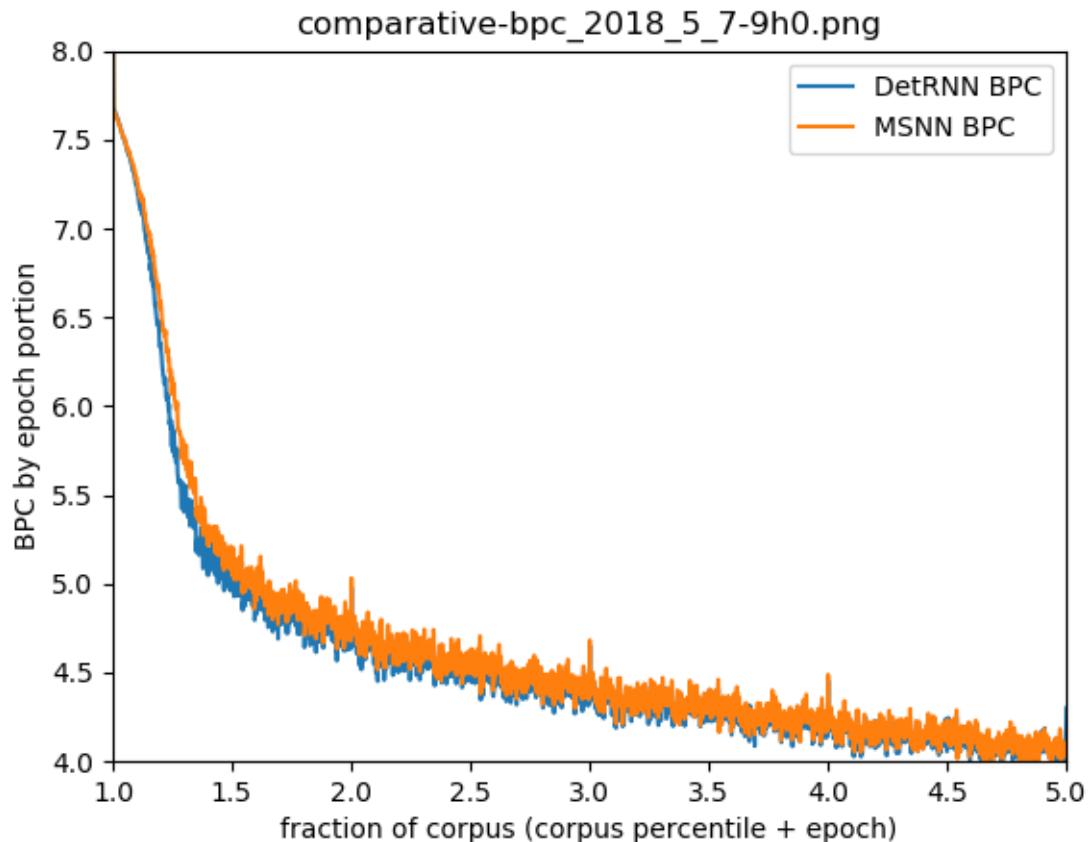


FIGURE D.5 – Comparative BPC

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).

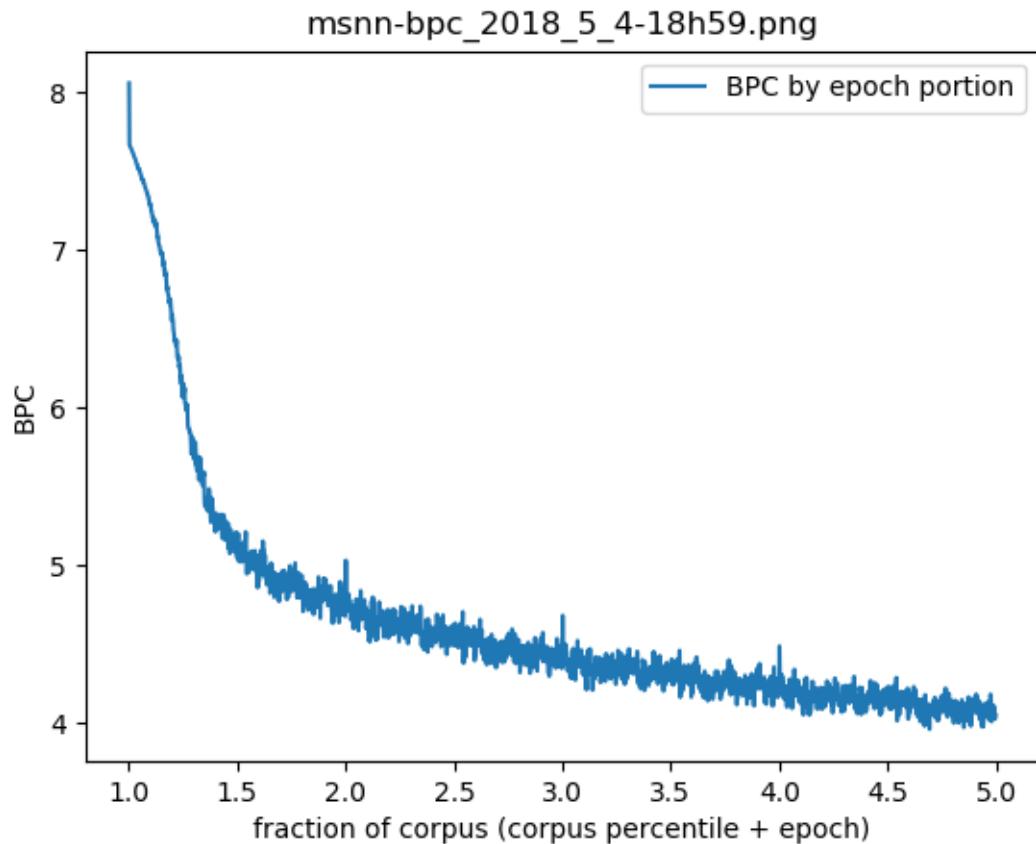


FIGURE D.6 – BPC

Logs

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

D.12.4 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.

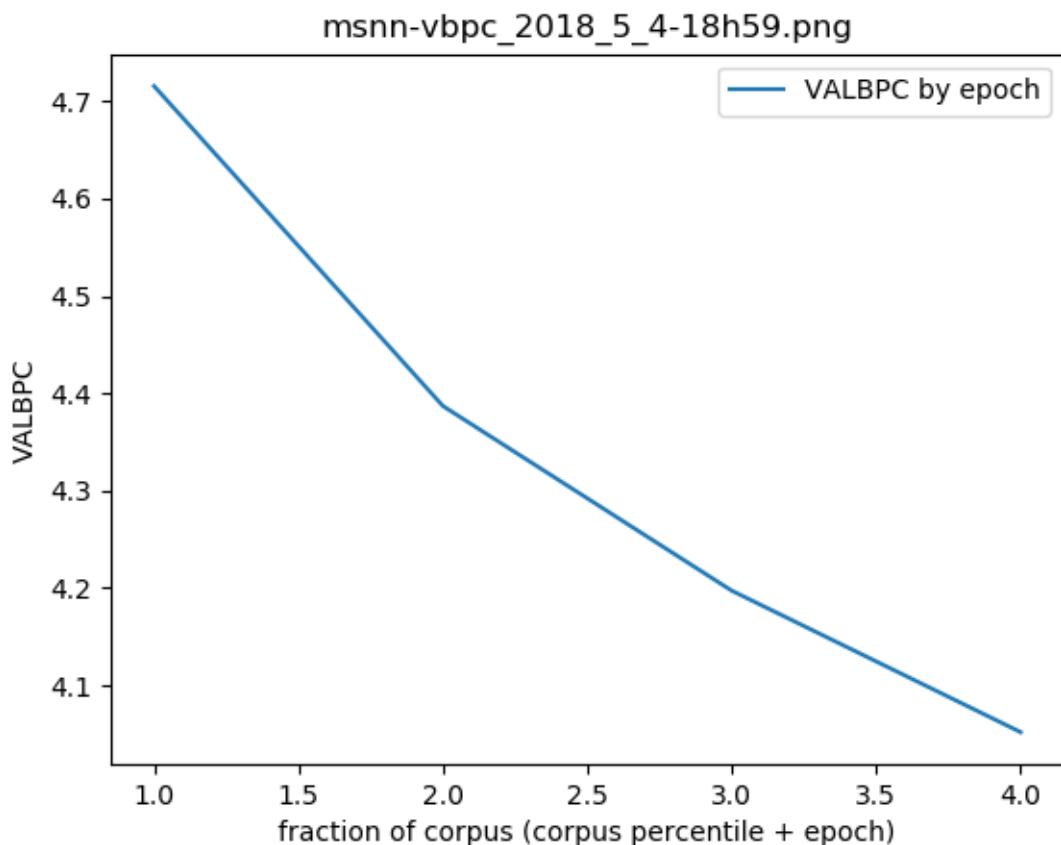


FIGURE D.7 – ValBPC

D.13 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

D.13.1 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.

D.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.

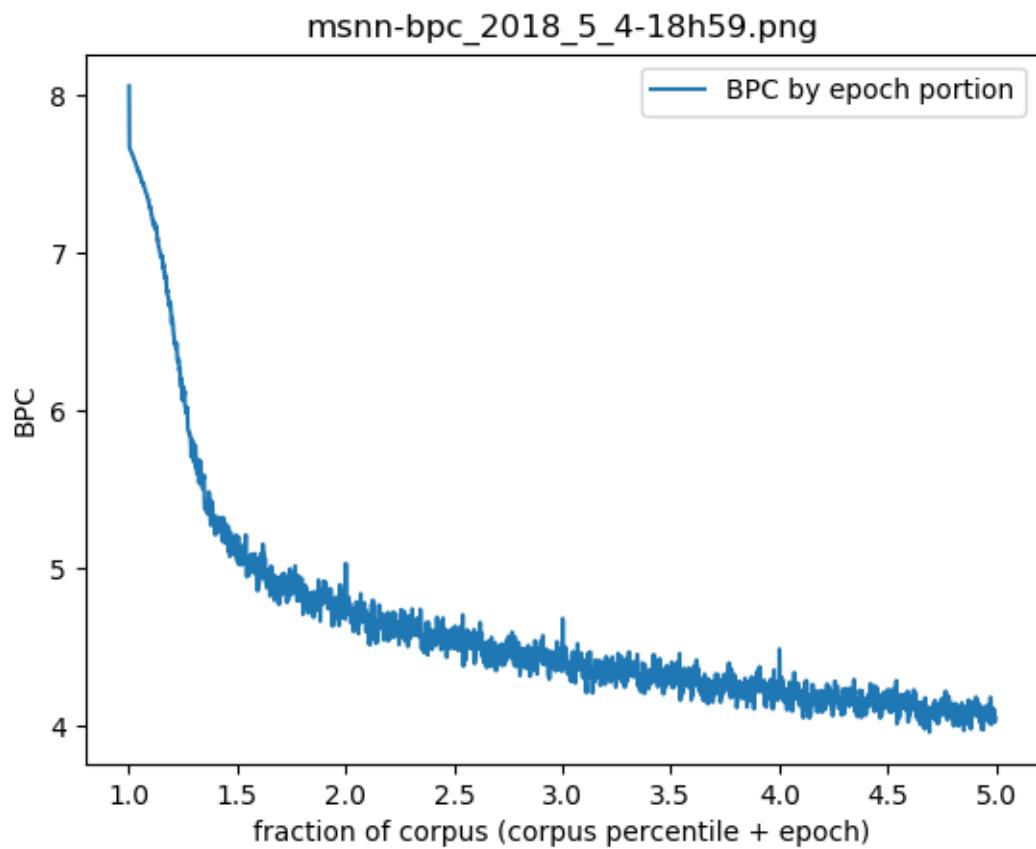


FIGURE D.8 – Loss

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
```

D.13.3 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.

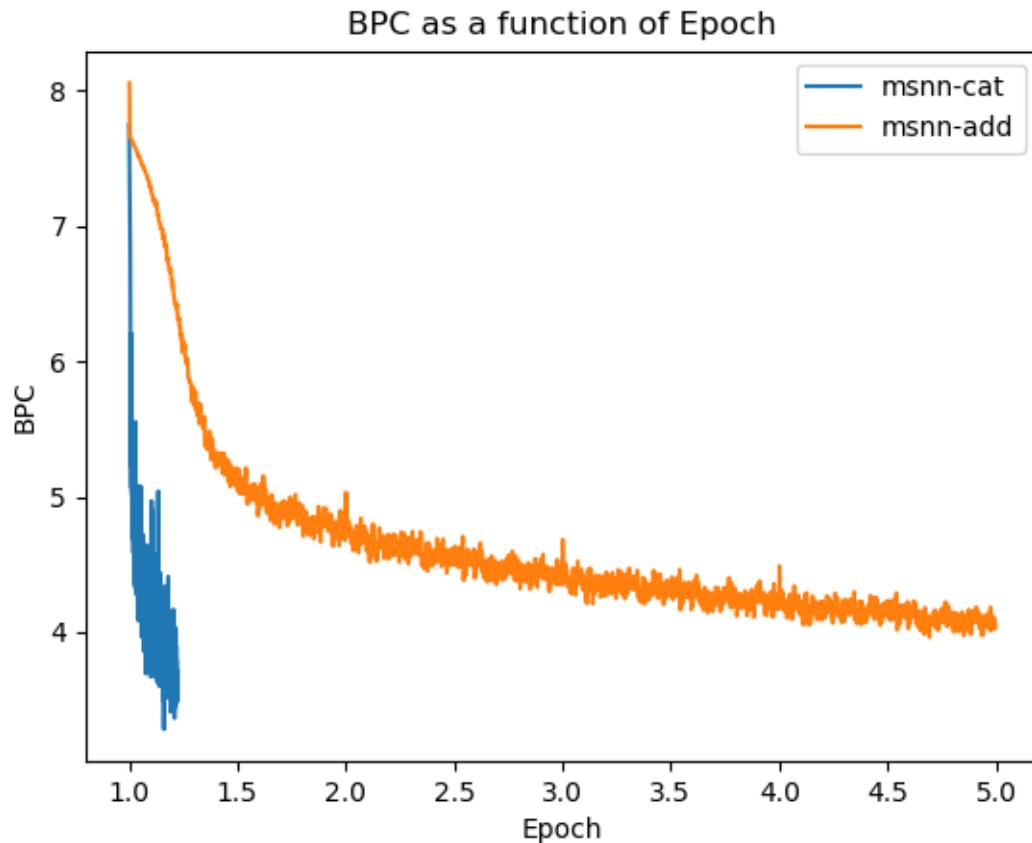


FIGURE D.9 – Comparative BPC

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.

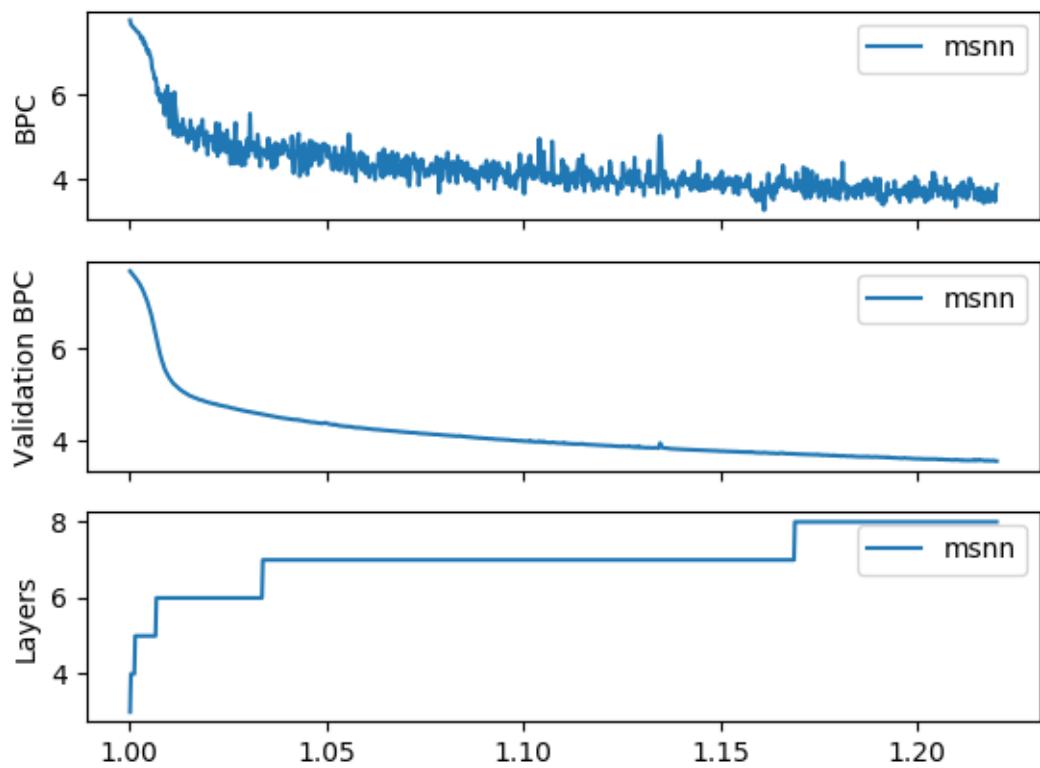


FIGURE D.10 – Comparative BPC

D.13.4 Potential ameliorations & next steps

Next step is to try to reduce run time.

D.14 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

D.14.1 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.

D.14.2 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
```

D.14.3 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

D.14.4 Potential ameliorations & next steps

Next step is to continue run time reduction.

D.15 Entrainement 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

D.15.1 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.

D.15.2 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.

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.

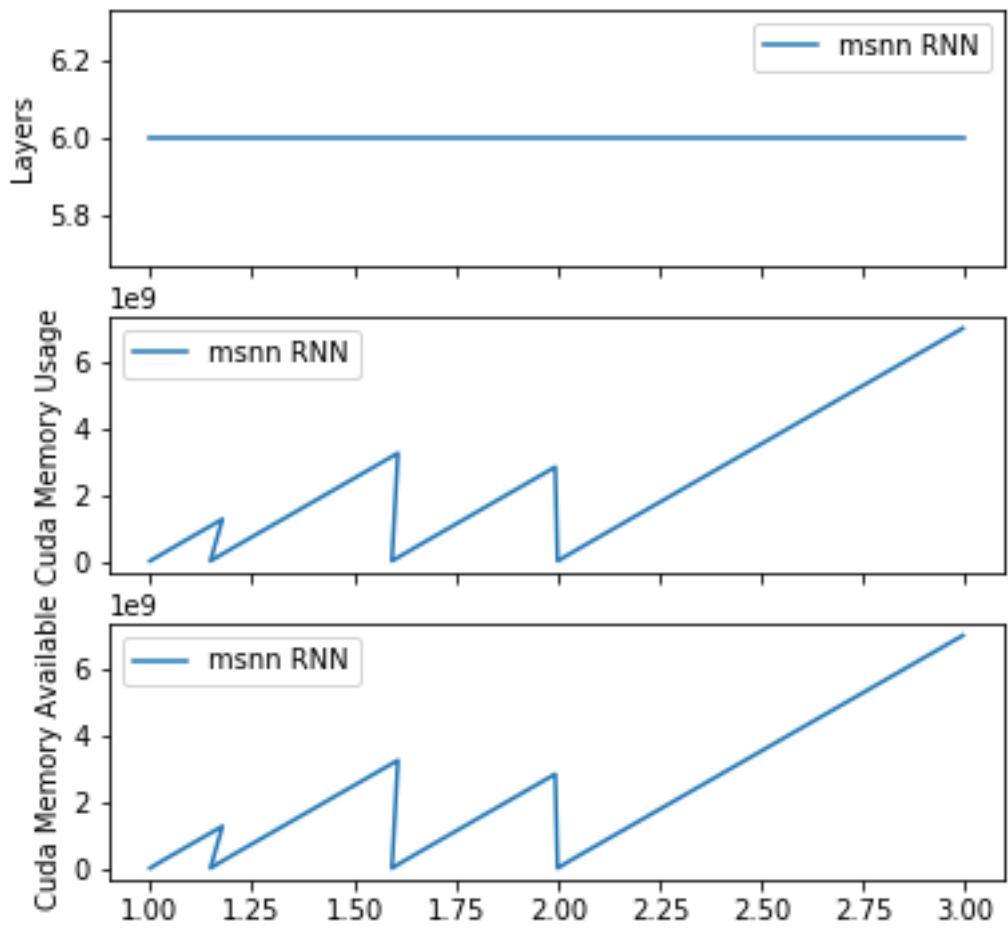


FIGURE D.11 – Memory usage

BPC / Validation BPC

BPC and Validation BPC

```
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
```

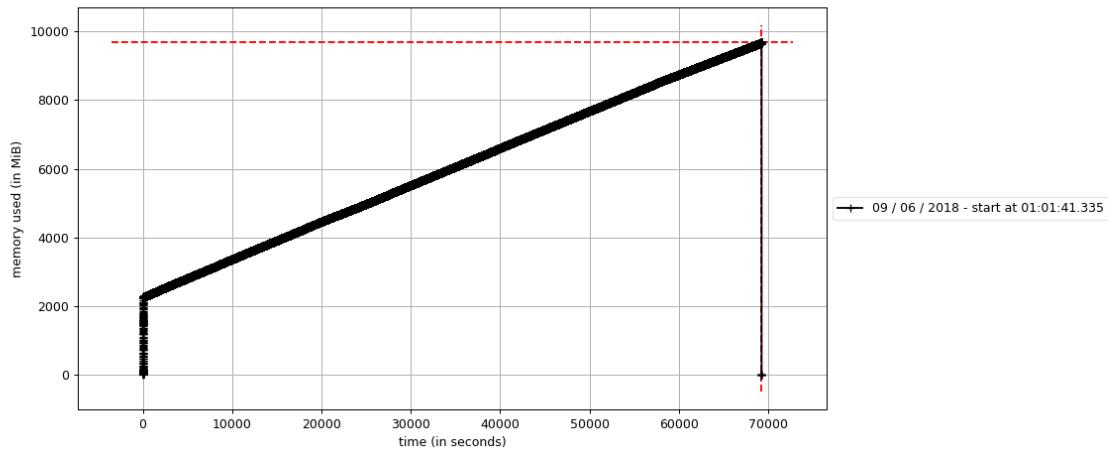


FIGURE D.12 – RAM third run

Job restart

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

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, ...).

D.15.3 Next steps

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

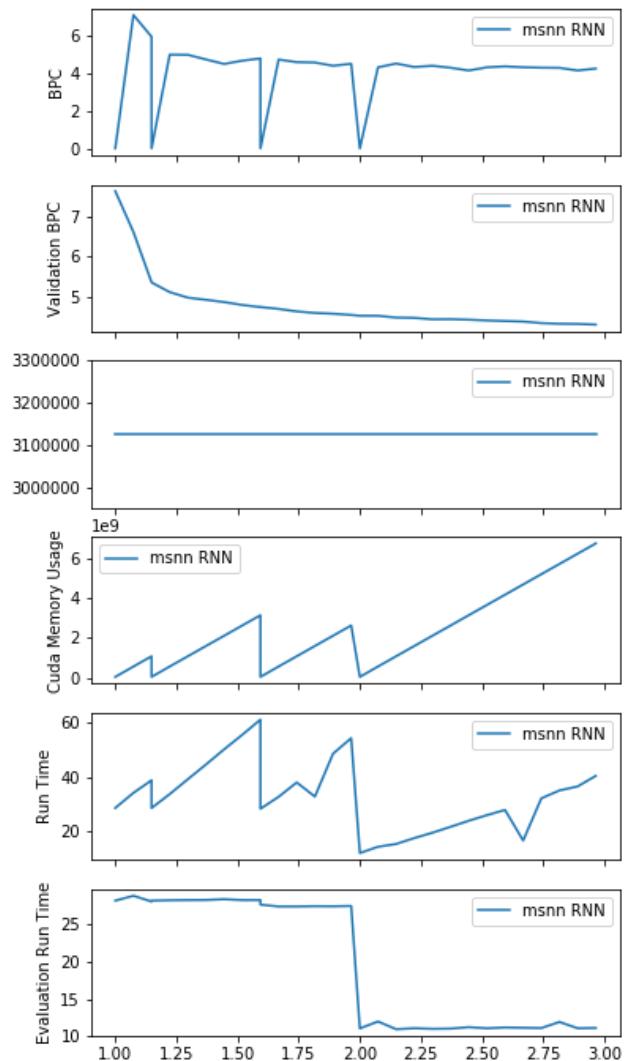


FIGURE D.13 – Memory and computation time

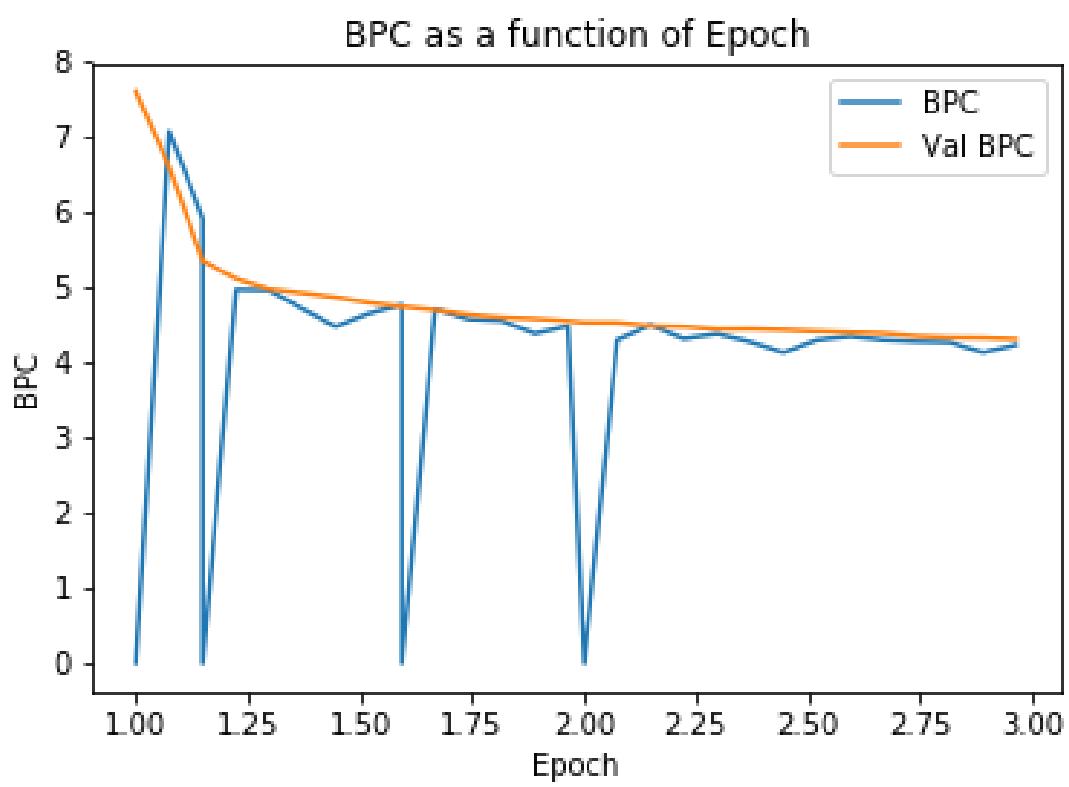


FIGURE D.14 – BPC

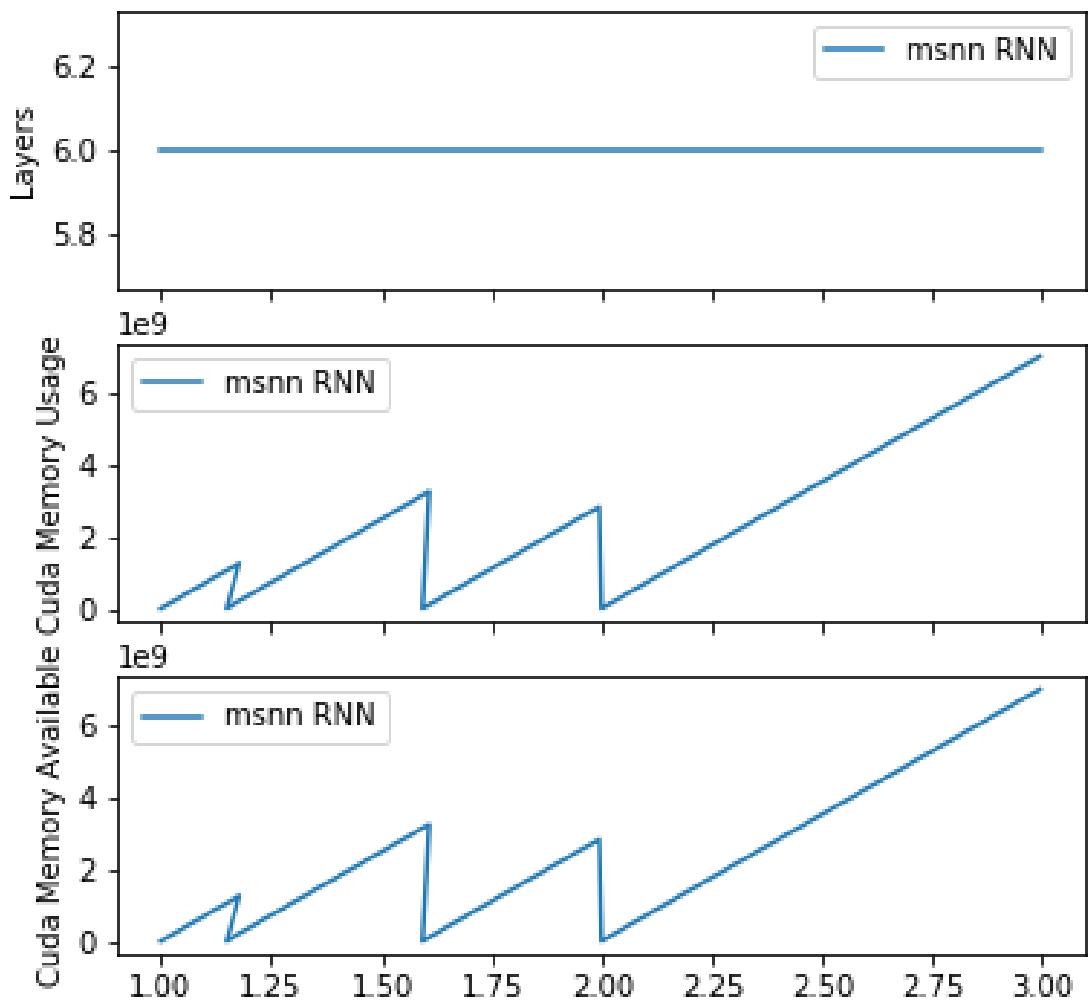


FIGURE D.15 – Memory usage

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

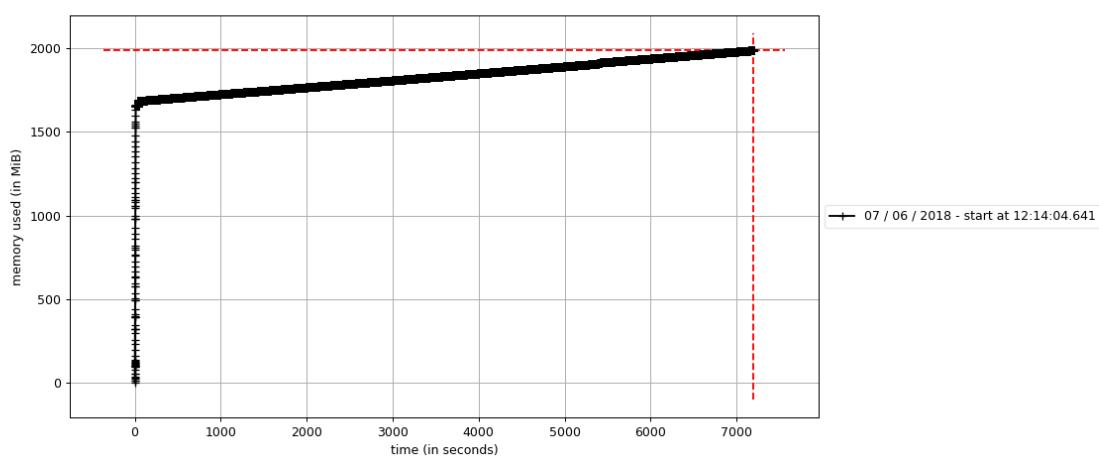


FIGURE D.16 – RAM first run

```
\da3/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
```

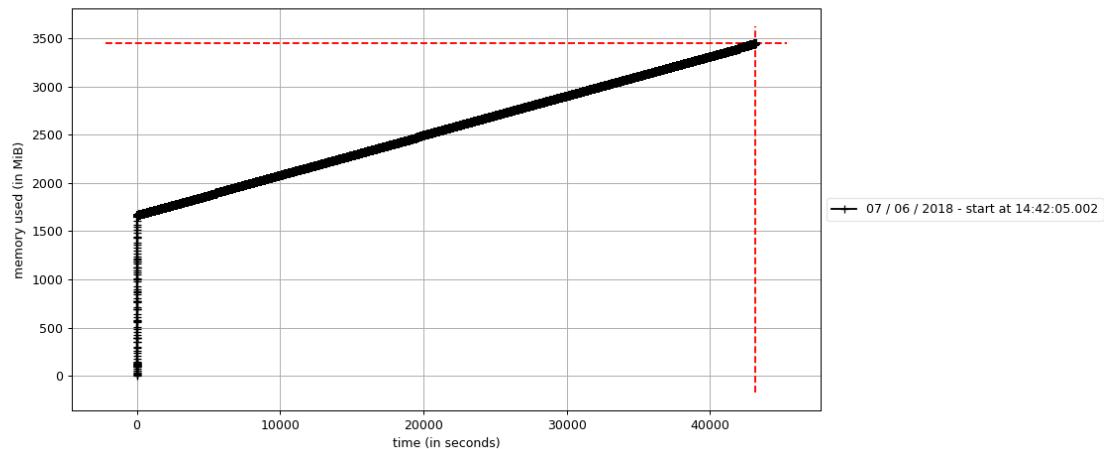


FIGURE D.17 – RAM second run

```
\da3/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
```

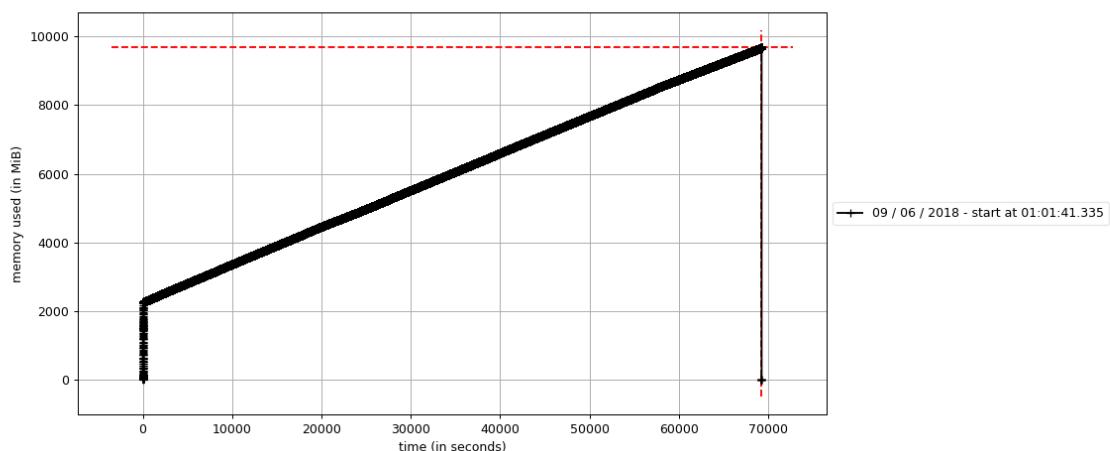


FIGURE D.18 – RAM third run

E Rapports d'avancement du projet PAPUD

E.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.

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

Pour les mêmes raisons que pour l'annexe précédente (décrise dans la section D.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.

E.2 Informations générales

General information

E.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

E.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>

E.3 Résultats de l'implémentation basique

Results of the basic implementation of the model

2018/07/09 - SYNALP - Esteban MARQUER

E.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.

E.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.

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

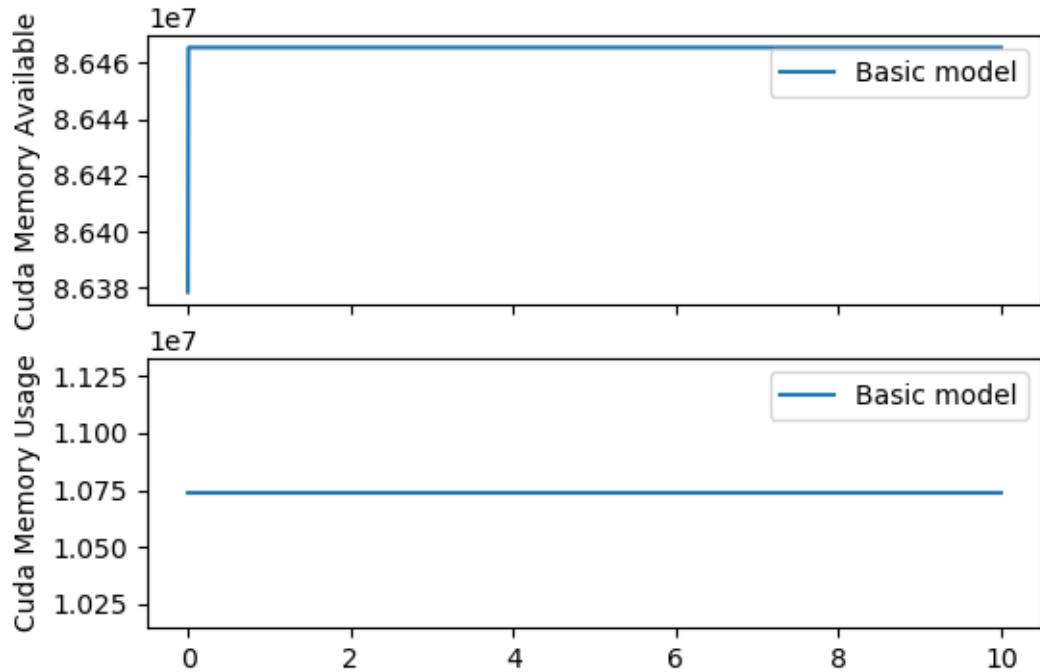


FIGURE E.1 – memory usage

specific examples.

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

E.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.

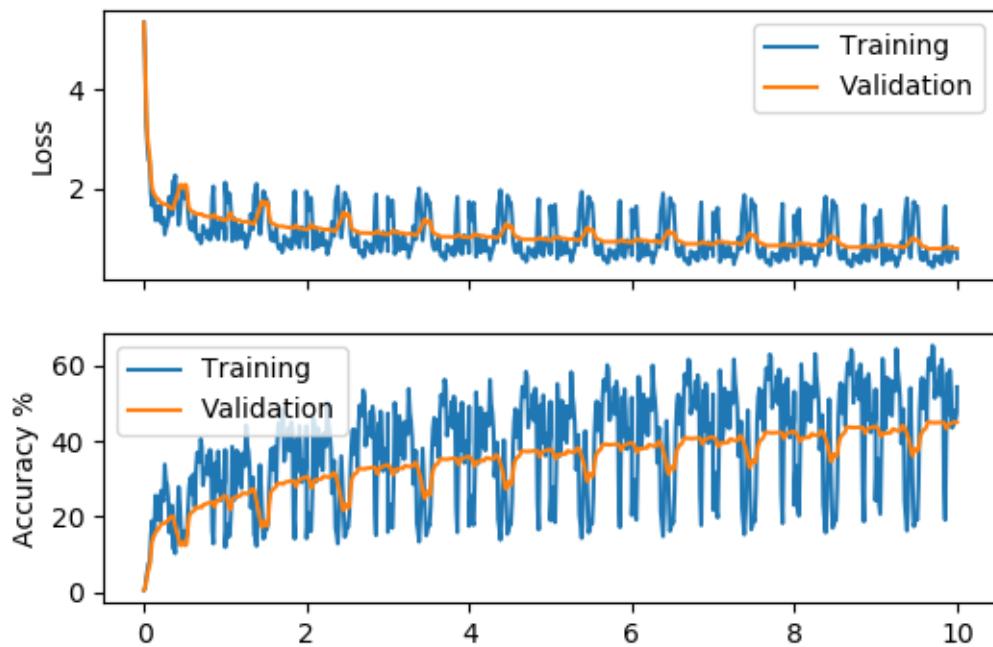


FIGURE E.2 – loss

E.4 Paquets (*batchs*) simultanés

Increasing the number of simultaneous examples

2018/07/10 - SYNALP - Esteban MARQUER

E.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.

E.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.

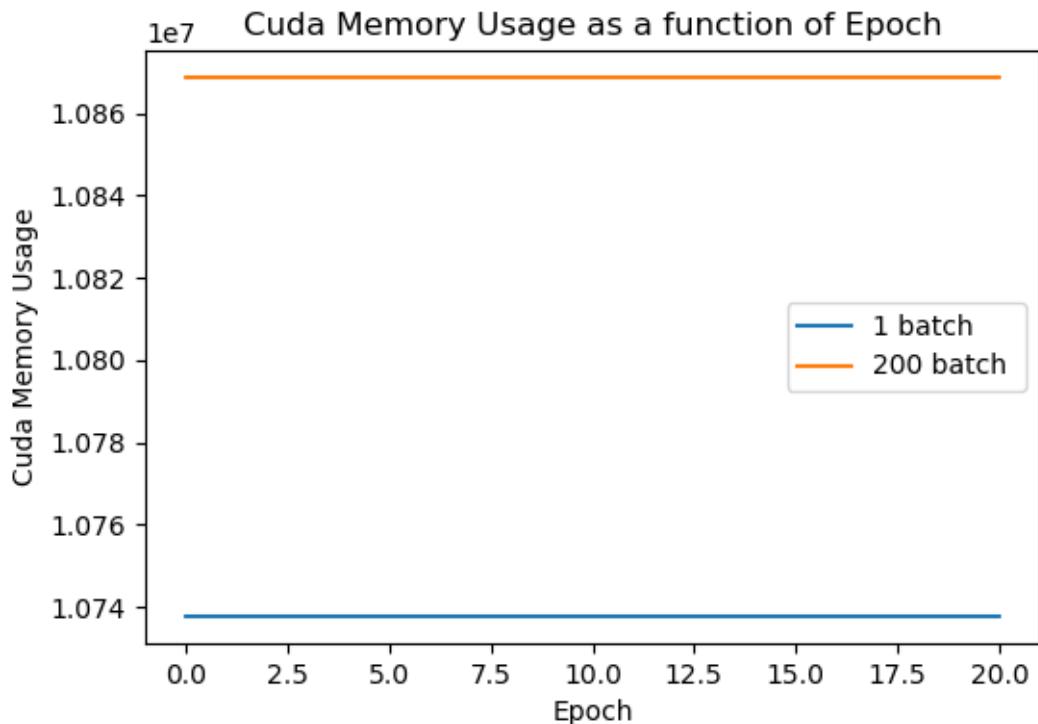
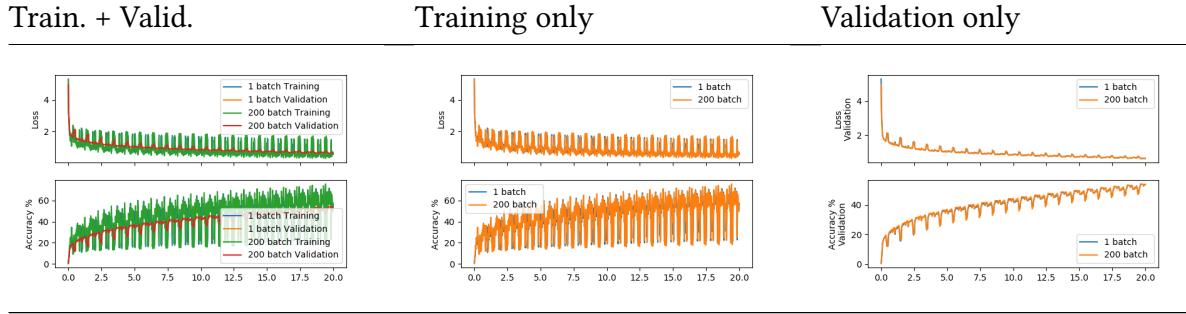


FIGURE E.3 – memory usage

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.



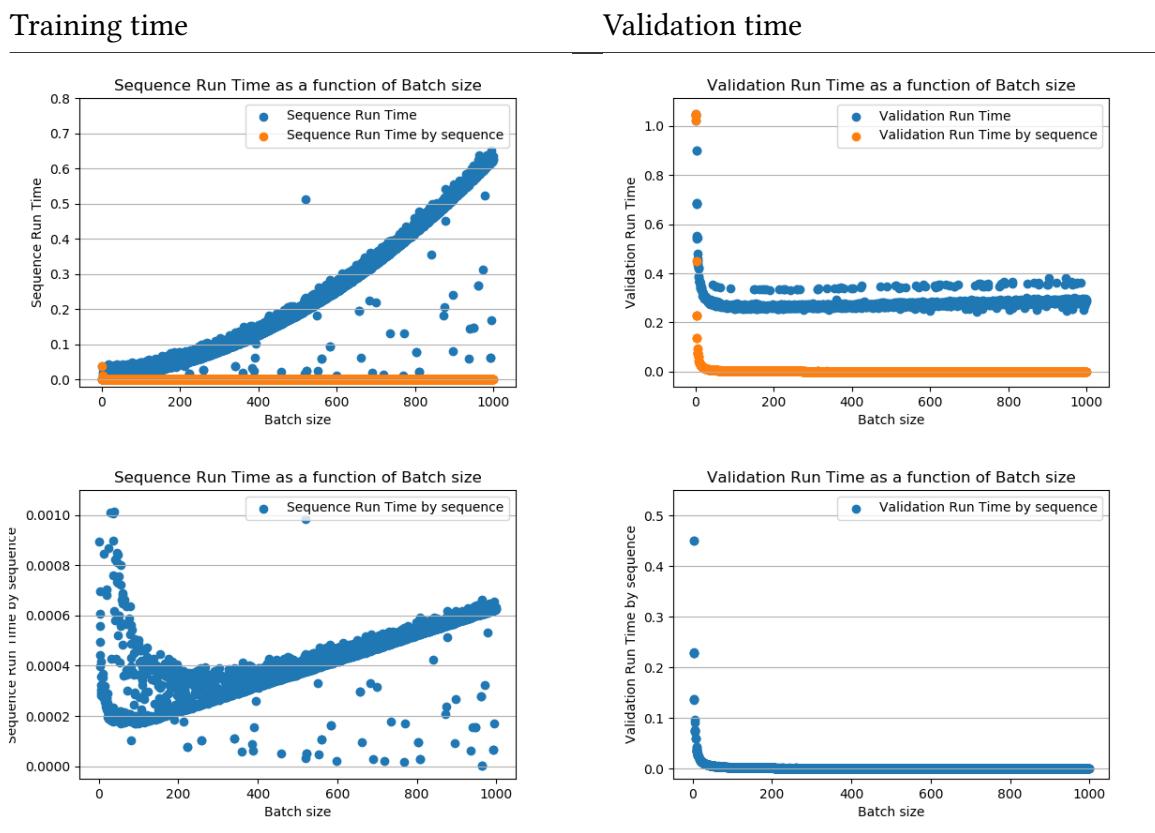
Computation time

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.



E.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.

E.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.

E.5 Analyse du pic de performance

Performance spike analysis

2018/07/18 - SYNALP - Esteban MARQUER

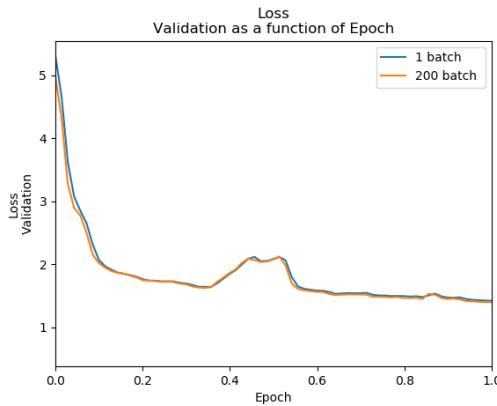
E.5.1 Problem

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

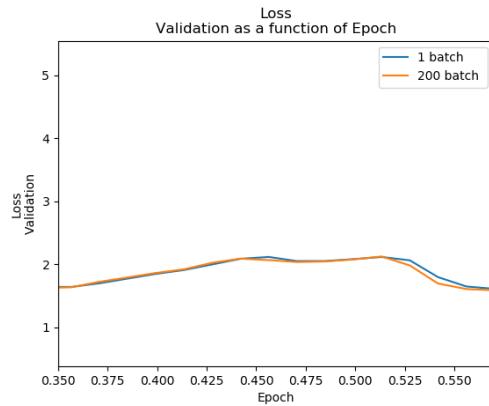
The different parts of the spike are :

- 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



E.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 x91fffff , 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

```
1 kern info kernel ACPI: LAPIC (acpi_id[0x00] lapic_id[0x00] enabled)
2 kern info kernel ACPI: LAPIC (acpi_id[0x02] lapic_id[0x02] enabled)
```

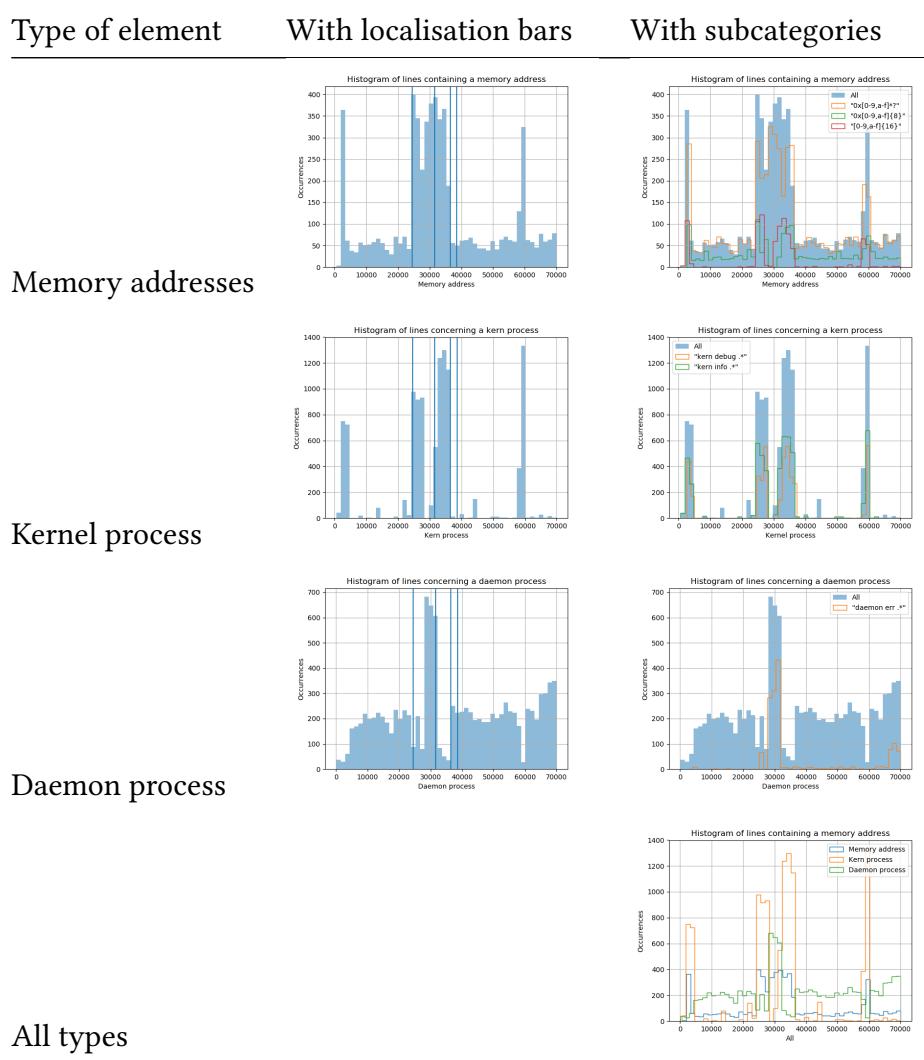
```

3 kern info kernel ACPI: LAPIC (acpi_id[0x04] lpic_id[0x04] enabled)
4 kern info kernel ACPI: LAPIC (acpi_id[0x06] lpic_id[0x06] enabled)
5 kern info kernel ACPI: LAPIC (acpi_id[0x08] lpic_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%)

```

E.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.

E.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.

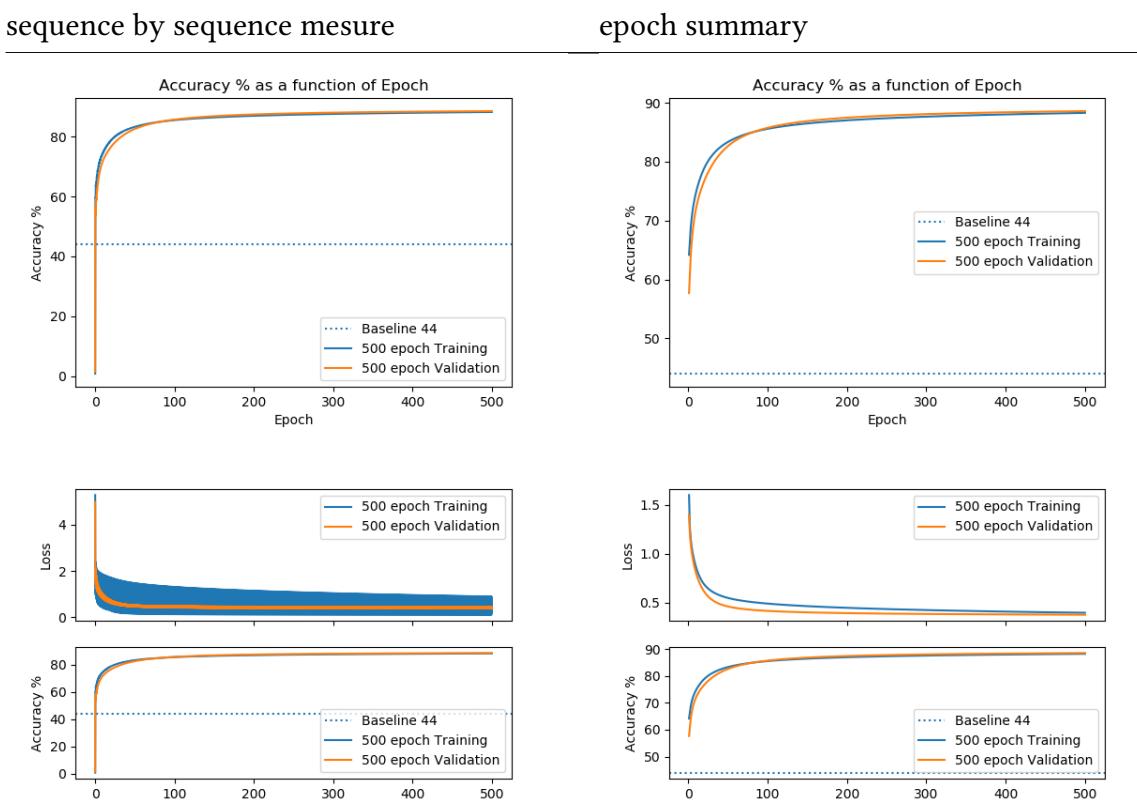
E.6 Rapport de la réunion avec les autres membres du projet

Team meeting report

2018/07/20 - SYNALP - Esteban MARQUER

E.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 new, with such a model training should go at 99% al 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

E.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).

E.7 Optimisation du taux d'apprentissage

Learning rate optimisation

2018/07/23 - SYNALP - Esteban MARQUER

E.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.

E.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 :

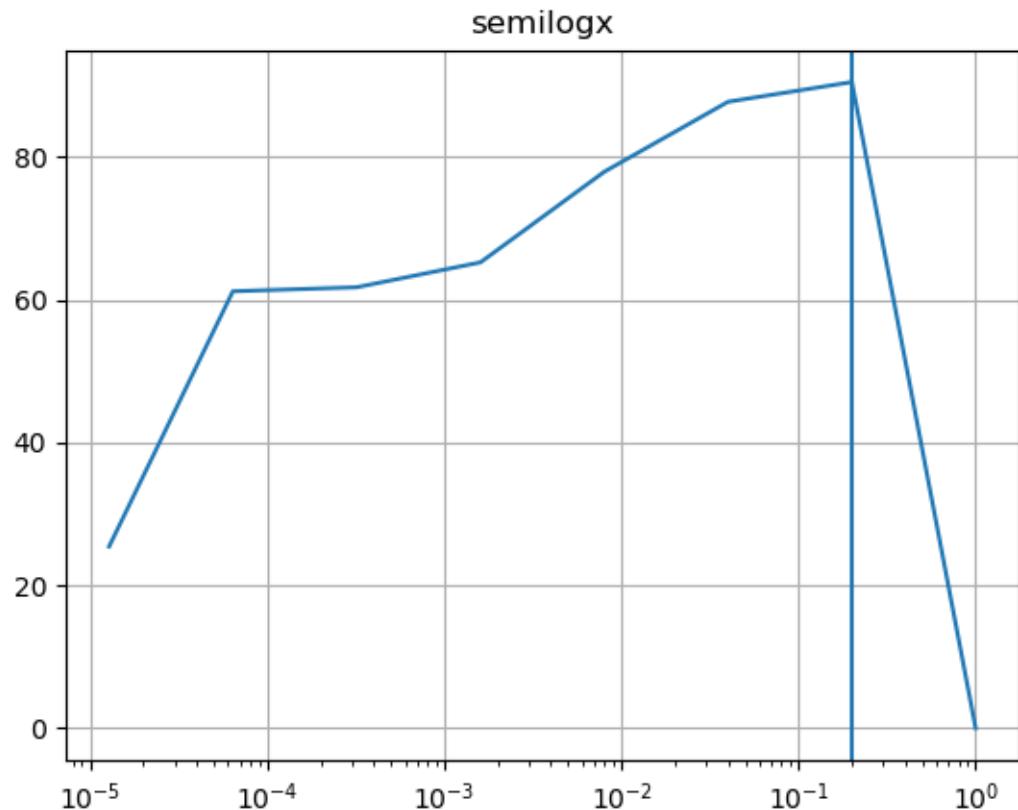
```
1 # the learning rate takes the highest value as start value, as it will
2 # decrease over time
3 learning_rate = start_value
4
5 # until the learning rate as reach the stop value (the lowest of the two
# bounds), we make it vary logarithmically
6 while learning_rate > stop_value:
7     performance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

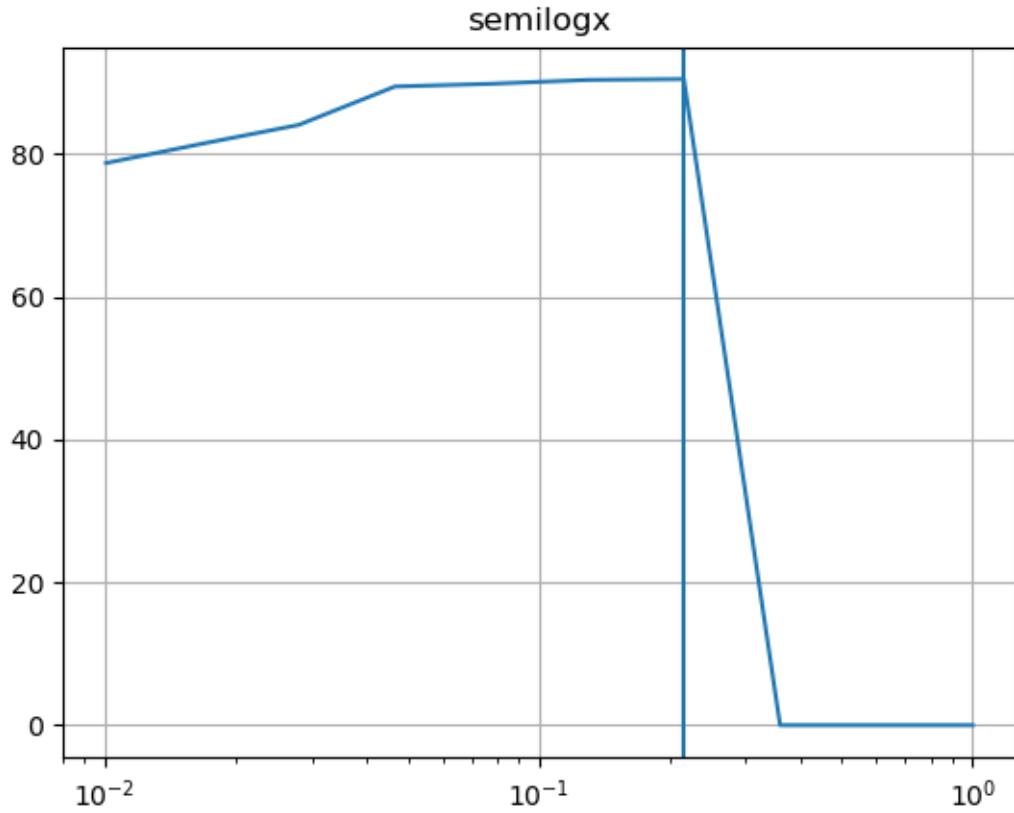
The closer to 0 the learning rate multiplier is, the faster the variation will be, and the closer to 1 it is, the slower the variation will be. To have a decent resolution in the curves, a multiplier close to 1 is crucial.

The training is done each time on a full epoch.

E.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).





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.

E.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.

E.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.

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

Learning rate optimisation effect on learning curve

2018/07/24 - SYNALP - Esteban MARQUER

E.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.

E.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).

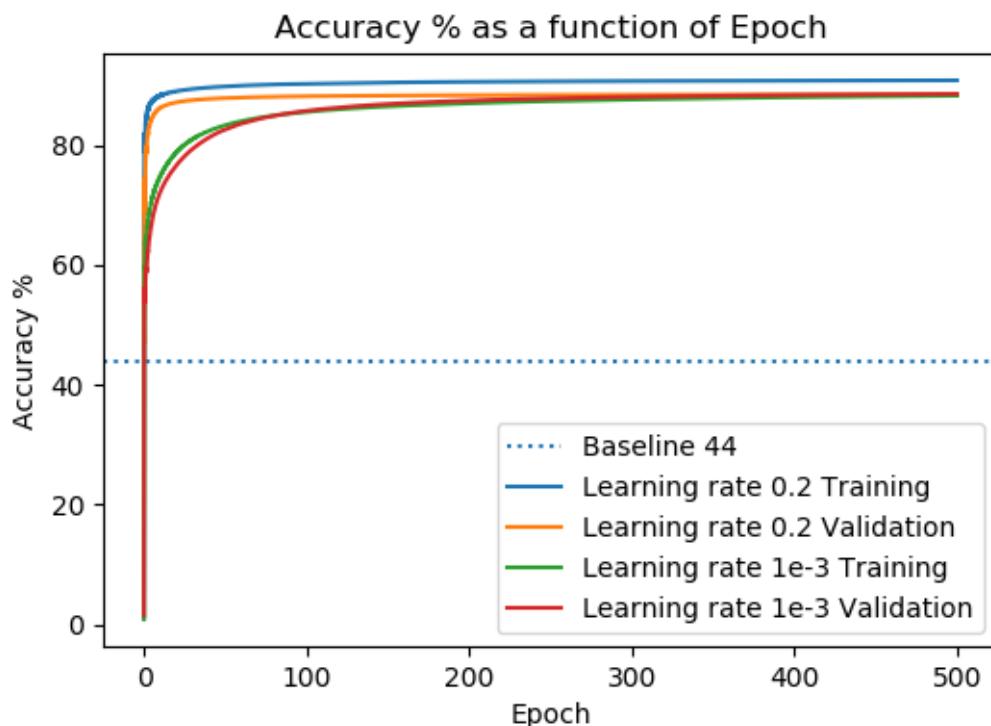


FIGURE E.4 – accuracy

E.8.3 Conclusion

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

E.8.4 Improvements and next steps

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

E.9 Taille de *batch* binaire

Binary batch size effect on run speed

2018/07/26 - SYNALP - Esteban MARQUER

E.9.1 Context

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

E.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.

E.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.

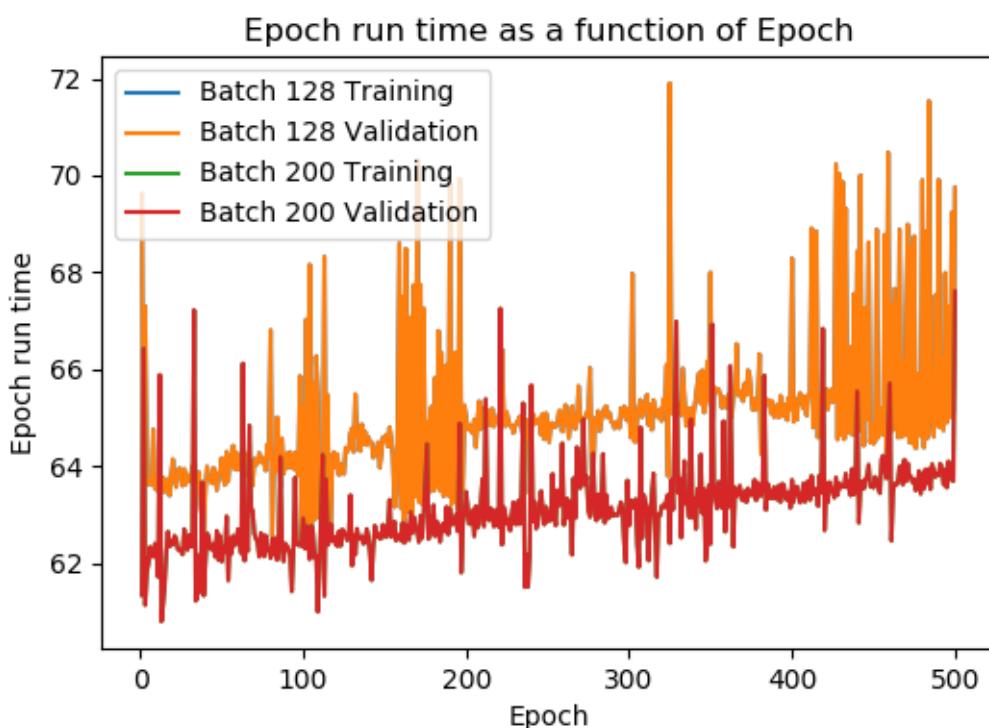


FIGURE E.5 – time per epoch

E.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.

E.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).

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

Learning rate optimisation effect on

2018/07/26 - SYNALP - Esteban MARQUER

E.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.

E.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.

E.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'}
```

E.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.

E.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.

F 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 :

Autre avantage : _____

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

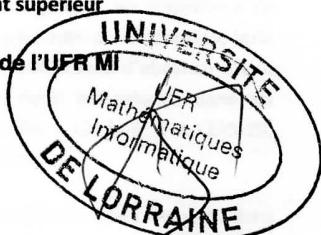
La présente convention est régie exclusivement par le droit français. Tout litige non résolu par voie amiable sera soumis à la compétence de la juridiction française compétente.

A Nancy le 12/04/18

Pour l'établissement d'enseignement supérieur

(nom et signature du représentant)

Antoine TABBONE, directeur de l'UFR MI

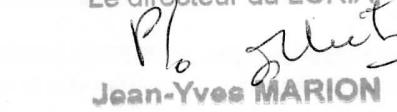


Pour l'organisme d'accueil

(nom et signature du représentant)

Le directeur du LORIA

Jean-Yves MARION



Pour l'étudiant

(nom et signature)

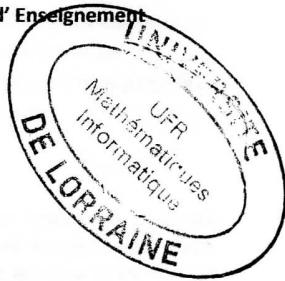
MARQUE Escobar



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

G Copie de l'avenant à la convention de stage



UNIVERSITÉ
DE LORRAINE

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/ : VER 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 : Estrhan

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
Jean-Yves MARION

L'enseignant référent du stagiaire

Nom et signature

CERISARA Christophe
Perison

Le tuteur de stage de l'organisme d'accueil

Nom et signature