Synthèse d'article « Convolutional Sequence to Sequence learning »

Alexis Durieux
INSA Rouen Normandie

Ce document est un résumé de l'article « Convolutional Sequence to Sequence learning (Gehring, Auli, Grangier, Yarats, & Dauphin, 2017). L'objectif est ici de définir et rassembler les innovations, détails d'implémentation et résultats de ces recherches dans le domaine du *sequence to sequence learning*

Introduction

Depuis l'émergence du *deep learning* et des *architectures profondes*, les réseaux récurrents de type **LSTM** et **GRU** sont particulièrement utilisées pour les données séquentielles car elles permettent d'utiliser l'information présente dans la temporalité des données grâce aux états cachés des itérations précédentes. Néanmoins, ces réseaux récurrents sont très profonds et lourds et leur apprentissage est également difficilement parallélisable. L'architecture proposée dans (Gehring et al., 2017) propose de résoudre les problèmes de *sequence to sequence* learning par l'intermédiaire d'un modèle doté de décodeurs et encodeurs de séquence entièrement convolutionnels ainsi que d'un système à attention. L'avantage annoncé de ces modèles est qu'ils permettent d'encoder une information hiérarchique sur les séquences en entrée et en sortie grâce au modèle à attention. De plus, l'utilisation de couches convolutionnelles permet de concevoir des implémentations d'apprentissage grandement parallélisables.

Architecture proposée

L'architecture proposée est la suivante. Dans un premier temps, les auteurs utilisent un encodeur entièrement convolutionnel puis un décodeur entièrement convolutionnel couplé d'un modèle à attention. Les couches convolutionnelles sont dotées d'une non-linéarité en sortie de convolution. Dans ce modèle, les non-linéarités utilisées sont des **GLU**(Dauphin, Fan, Auli, & Grangier, 2016) Grâce à l'empilement des couches, les convolutions en une dimension successives permettent de donner du contexte à la séquence courante. Entre les couches de convolution pour le décodeur un *padding* est utilisé afin de revenir à la taille de la séquence d'entrée. Le système à attention utilisé permet à partir de l'état courant du décodeur et une représentation vectorielle d'encodeurs précédents de déterminer quelles entrées précédentes ont été prises en compte et d'adapter la propagation courante. Une illustration des modèles à attention dans le contexte du *sequence to sequence* learning est disponible à l'adresse en bas de page [1]

Apprentissage

Lors de l'apprentissage, plusieurs techniques sont utilisées afin de contrôler au mieux ce dernier. Une attention particulière est notamment accordée à l'initialisation des poids du réseau de neurones. En effet, toutes les couches sont initialisées grâce à des lois normales de moyenne 0. Cependant pour les couches connectées ou non derrière à une non-linéarité *GLU*, les variances utilisées sont variables. Elles dépendent du nombre de connexions d'entrée au neurone ainsi que de la présence ou non d'une non-linéarité en sortie de neurone. Cette méthode permet de contenir la variance des entrées. L'apprentissage est également stabilisé par la relative constance de la variance dans le réseau de neurone. Ceci est achevé par normalisation de certaines partie du réseau lors de l'apprentissage, notamment les entrées conditionnelles générées par le modèle à attention.

Résultats

Sur les tâches de *WMT 14* [2]. Anglais-Allemand et Anglais-Français, cette architecture nommée **ConvS2S** domine l'était de l'art actuel basé sur des **LSTM**. Les scores des résultats sont exprimés selon la métrique **BLEU** [3]. On remarque donc que les résultats issus de l'architecture entièrement convolutionelle sont une amélioration par rapport au précédent état de l'art. De plus, grâce à la légèreté de ses couches convolutionelles, les prédictions sont beaucoup plus rapides que lors de la propagation d'un réseau récurrent.

TABLE 1
Résultats sur des tâches WMT

Model	WMT'14 English-German	WMT'14 English-French
ConvS2S	26.43	41.16
Wu et al. (2016 - GNMT + RL)	26.30	41.44

1. https ://github.com/facebookresearch/fairseq/blob/master/fairseq.gif
2. WMT = Word Machine translation
3. https ://en.wikipedia.org/wiki/BLEU

Conclusion

Cette nouvelle architecture pour le traitement de données temporelle est intéressante de par le fait qu'elle propose une décomposition hierarchique du signal plus que temporelle, notamment pour les phrases. De plus, les convolutions permettent de paralléliser grandement les calculs ce qui permet d'accélérer les apprentissages. Néanmoins, les tests effectués présentés dans le papier sont peu nombreux et bien qu'ils démontrent une amélioration quant à l'état de l'art, il reste à prouver que ces architectures sont généralisables au pan tout entier du *sequence to sequence learning*.

Références

Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D. (2016). Language modeling with gated convolutional networks. *CoRR*, *abs/1612.08083*. Consulté sur `http://arxiv.org/abs/1612.08083`

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017, mai). Convolutional Sequence to Sequence Learning. *ArXiv e-prints*.