Challenge de la transcription automatique de la parole: exemple du système *Deep Speech* de Baidu. (Laurent Besacier)

Introduction

Afin d'évaluer les ressources nécessaires pour construire un système de transcription automatique de la parole tel que ceux développés par les géants industriels du domaine, nous avons analysé en détail l'article qui décrit le système *Deep Speech 2*, récemment déployé par Baidu. Cet article (publié à la conférence ICML 2016) est en ligne sur: https://arxiv.org/abs/1512.02595. Le système permet, d'après les résultats expérimentaux obtenus sur des jeux de données standard, des performances de transcription proches des performances humaines ("*.../... in several cases, our system is competitive with the transcription of human workers when benchmarked on standard datasets.*"). En premier lieu, il est intéressant de noter que cet article de conférence rassemble **34 co-auteurs** (tous ne sont pas spécialistes de traitement de la parole et d'apprentissage profond car l'entraînement le déploiement opérationnel d'un tel système nécessite aussi des compétences en HPC, par exemple).

Taille des collections utilisées pour l'apprentissage

Deux systèmes de transcription automatique sont construits pour l'Anglais et le Mandarin à partir de 12000h et 9400h de parole transcrite, respectivement. Si on suppose une fréquence d'échantillonnage des signaux de 16khz, ceci représente des **données brutes de 2.5To** (hors "data augmentation") auquelles il faut ajouter des grandes quantités de textes (transcriptions des signaux + données pour l'apprentissage des modèles de langue) qui représentent quelques dizaines de Go supplémentaires.

Dimmension du réseau profond construit

L'architecture du système *Deep Speech 2* présente jusqu'à 12 couches (réseaux convolutionnels dans les couches "basses" et réseaux récurrents dans les couches "hautes"). Le **nombre total de paramètres du réseau est entre 18M** pour le modèle le plus petit (taux d'erreur de mots sur l'anglais 10.59%) **et 100M** pour le modèle le plus grand évalué (taux d'erreur de mots sur l'anglais 7.73%).

Besoins computationnels pour l'apprentissage

L'**apprentissage** du modèle nécessite des dizaines d'**exaFLOPs** (10^18) qui représenteraient **3 à 6 semaines pour s'executer sur un seul GPU**. Les auteurs décrivent un certains nombre d'optimisations qui permettent de revenir à un apprentissage qui nécessite environ 50 teraFLOP/s (10^12) lorsqu'il se déroule sur 16 GPU de type Titan X (3 à 5 jours d'entraînement). La figure ci-dessous indique le temps d'apprentissage pour 1 époque en fonction du nombre de processeurs GPUs disponibles. On voit qu'une époque d'apprentissage représente 2^18s (3 jours) sur un seul GPU.

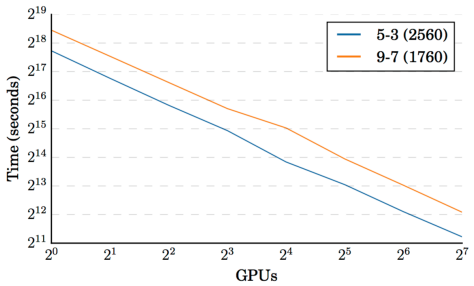

Figure 4: Scaling comparison of two networks—a 5 layer model with 3 recurrent layers containing 2560 hidden units in each layer and a 9 layer model with 7 recurrent layers containing 1760 hidden units in each layer. The times shown are to train 1 epoch. The 5 layer model trains faster because it uses larger matrices and is more computationally efficient.

(image issue de : https://arxiv.org/pdf/1512.02595.pdf)

Il est cependant important de noter que **les chiffres donnés ci-dessus concernent un seul apprentissage tandis que l'éxploration des diverses architectures et hyper-paramètres nécessite des allers-retours permanents entre apprentissage et évaluation**. Un modèle comme celui présenté par Baidu dans cet article a sans doute nécessité N (**N>500**) apprentissages sur les collections de données entière avant d'être réglé.

Besoins en mémoire

La mémoire ne sert pas seulement à charger les paramètres du réseau mais elle sert aussi, au cours de l'apprentissage, à garder les activations à travers chaque couche du réseau en vue de la retro-propagation. Par exemple, dans *Deep Speech 2*, **charger en mémoire un réseau de 70M de paramètres nécessite 280Mo**, mais **stocker en mémoire les activations pour un batch de 64 signaux (phrases de 7s en moyenne) nécessite 1.5Go** et des cas d'échec pourront survenir au cours de l'apprentissage si de longs signaux sont présents dans le batch.

Conclusion

L'article *Deep Speech 2* a été publié à ICML 2016. Après 18 mois, entraîner un système de transcription automatique tel que celui-ci **n'est pas possible pour la majorité des laboratoires académiques en France** (sans parler du déploiement pour traiter plusieurs dizaines de requêtes utilisateurs en parallèle).