IFT6285 – Devoir 4

Correction de mots

Maxime MONRAT
Juan Felipe DURAN

Université de Montréal
Automne 2021

1. Le modèle a été entraîné sur un PC équipé d'un processeur 8 cœurs à 3.89 GHz. On remarque que le temps de traitement augmente de manière assez linéaire en fonction du nombre de tranches considérées pour l'entrainement, chaque tranche supplémentaire prenant environ 25min de plus à être entraînée.

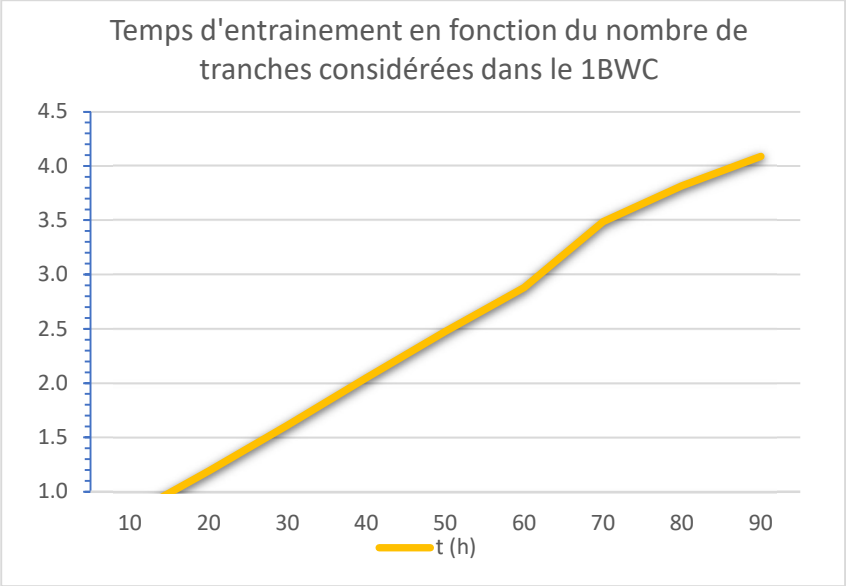
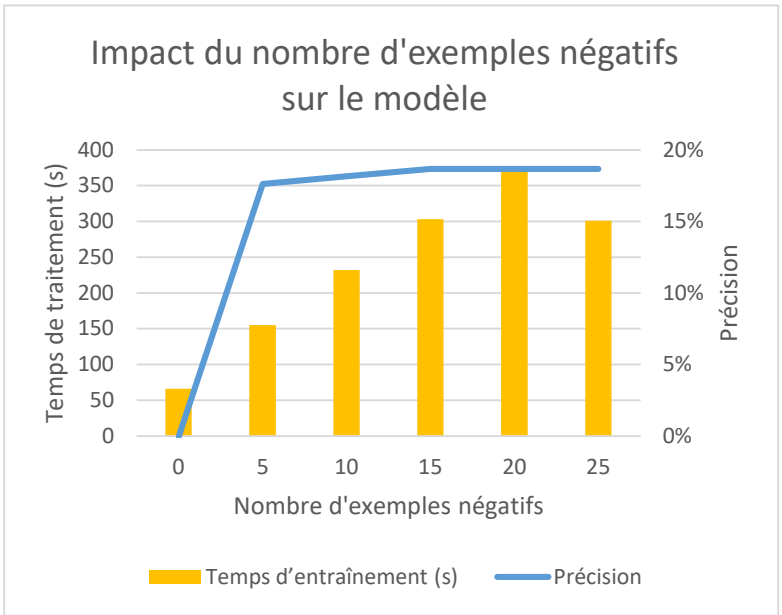
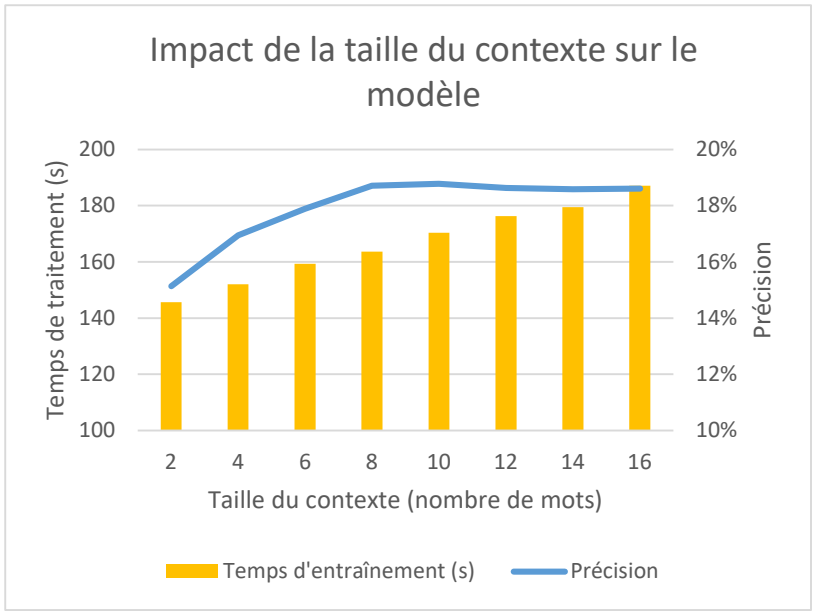
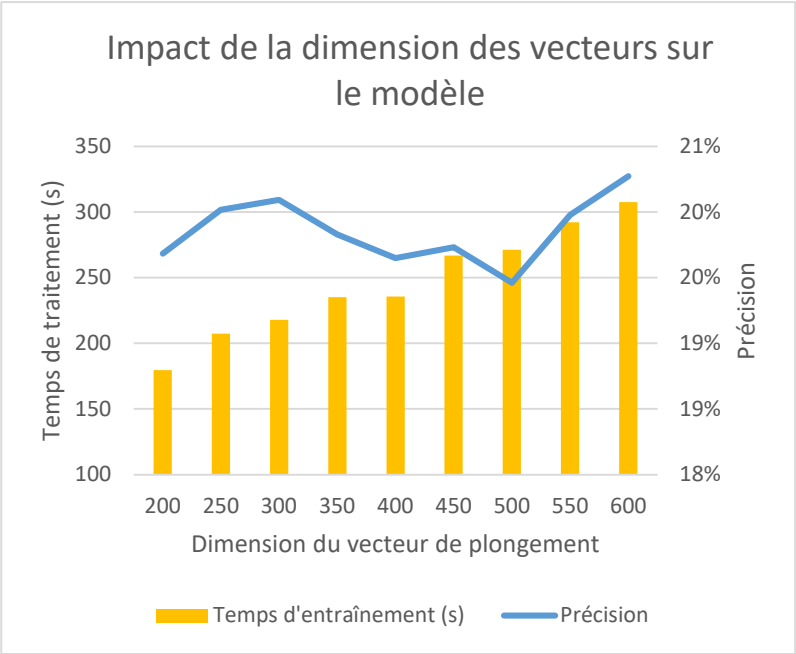


Figure 1: Courbe présentant le temps d'entraînement (h) en fonction du nombre de tranches du corpus considérées

2. Nous avons analysé l'impact des différents méta-paramètres sur un modèle entraîné sur les 5 premières tranches du corpus (exemples négatifs, taille du contexte et dimension du vecteur). Afin d'évaluer la précision de notre modèle hors-contexte, nous avons effectué un test sur les analogies en utilisant la fonction `evaluate_word_analogies` de Gensim, en utilisant le test set de Google questions-words.txt.







On remarque que de manière générale utiliser des plus grandes valeurs pour ces trois méta-paramètres fait augmenter le temps de traitement, mais permet également des gains importants en précision. Évidemment, ces données sont à moduler avec le fait qu'en situation réelle, la performance du modèle doit être évaluée sur un set de test en lien avec les objectifs d'utilisation.

Au vu de nos données, notre modèle optimal utilise **20 exemples négatifs**, un **contexte de 10 mots** et plonge le vocabulaire dans un espace vectoriel de **dimension 300**.