Problème de traduction : **Language Model** vs **Sequence to Sequence**

17 avril 2024

1 Objectif

On cherche à comparer la performance, à nombre de paramètres égaux, de deux modèles traditionnels du NLP dans le problème de la traduction automatique. La première méthode, classique, consiste à utiliser un modèle dit **sequence to sequence** et à l'entraîner sur des couples de phrases source/cible. La seconde méthode, plus récente, consiste à entraîner un **modèle de langue** avec un decoder-only Transformer sur des exemples où l'on concatène la phrase source et la partie de la phrase cible en mode teacher-forcing.

2 Déroulement prévisionnel

Étape 0 (facultative) - 1 mois : Formation NLP accélérée (video David Louapre comme introduction, articles Pirmin à récupérer, LSTM, mécanisme d'attention, The Illustrated transformer, séminaire Pirmin une brève histoire du NLP).

Objectif : améliorer sa culture générale du NLP

Étape 1 - 2 mois : Construction d'un seq2seq LSTM avec et sans attention. On peut s'inspirer du code ici.

Répéter la tâche de traduction avec un decoder-only LSTM avec et sans attention (séparateur entre les deux langues choisies).

Objectif : se faire les mains sur la partie code + développer des intuitions sur les tâches (similarité/différences, quels sont les points de tension genre le vocabulaire du LM/seq2seq, etc...)

Etape 2 - 3 mois : Construction d'un transformer seq2seq et d'un transformer decoder-only sur les mêmes données qu'à l'étape 1.

Objectif : comparer les performances des deux architectures, déterminer les différences de traduction d'un aspect linguistiques.

3 Aspect Technique

3.1 Evaluation

On évalue les modèles avec le score BLEU, on essaie d'obtenir les mêmes résultats que l'état de l'art en fonction du/des jeu(x) de données choisi(s).

3.2 Jeu de données

Probablement utile de choisir plusieurs jeux de données de langue différentes (pour espérer faire une étude linguistique des phrases générées). Beaucoup de ressources au Workshop on statistical Machine Translation (WMT).

3.3 Puissance de calcul

Utilisation des machines de onepoint OU utilisation de Jean Zay.

4 Travail similaire

L'article Language Models are Good Translators pose la même question que nous : "can we really accomplish the machine translation task with a single language model ?". Les chercheurs expliquent que le LM et le seq2seq obtiennent les mêmes résultats à nombre de paramètres égaux.

Critique de l'article : déjà ils ont peu de choses à raconter à part la comparaison de BLEU score qui sont globalement équivalent.

Ils tentent de rajouter dans la loss un terme qui force le modèle à reconstruire la source mais ils ont bien du mal à le justifier : "may reduce the representation gap between source and target sentences".

Ils justifient l'utilisation de plusieurs jeux de données pour faire une comparaison de scaling en fonction de leurs tailles, ce qui paraît absurde. C'est intéressant de voir si certaines propriétés linguistiques différentes sont capturés et c'est intéressant de déterminer comment les modèles scalent plutôt.

Ils justifient l'utilité du même espace pour les deux langues en utilisant le problème de pivot (on veut traduire A $\rightarrow$ C mais on a pas de données, on traduit A $\rightarrow$ B puis B $\rightarrow$ C) : "Intuitively, pivot-based translation tasks can benefit from LM4MT model that learns a shared representation across different languages".

Aucune étude linguistique, aucun exemple des sorties générés dans l'article, c'est vraiment très bizarre.

Qu'est ce qu'on apporterait en plus lors du stage ? ...
Objectif minimal du stage ? Où il peut se démarquer