STATAPP 49 : PETITS MODÈLES DE LANGUE

MATÉO AMAZO, ROMAIN DELHOMMAIS, VINCENT GIMENES, GUILLAUME LÉVY

Résumé. Cette note présente une étude sur l'évaluation des performances des modèles de langue dans des environnements de ressources limitées, en se concentrant sur l'utilisation du modèle GPT-2 sur des données d'entraînement provenant de pages Wikipédia en anglais.

1. Introduction

1.1. **Contexte du projet.**

Dans le domaine du traitement automatique du langage naturel (NLP), les modèles de langue jouent un rôle crucial dans la compréhension et la génération de texte. Cependant, ces modèles sont souvent associés à des exigences en ressources massives en termes de données d'entraînement et de puissance de calcul. Notre projet de recherche se concentre sur l'évaluation de la performance des modèles de langue dans des conditions de ressources limitées, en particulier avec des ensembles de données d'entraînement de taille réduite.

1.2. **Objectifs poursuivis.**

Cette note présente une vue d'ensemble de nos travaux, en mettant en lumière les défis et les découvertes clés. Nous abordons les implications de nos résultats pour des applications pratiques et explorons les pistes futures pour améliorer l'efficacité des modèles de langue dans des environnements contraints. Notre objectif est de rendre accessible à un large public les enseignements tirés de notre étude, ouvrant ainsi la voie à des avancées significatives dans le domaine du NLP avec des ressources limitées.

2. Compréhension des modèles de langue

Dans cette section, nous présentons de manière simplifiée les notions essentielles relatives aux modèles de langue, éléments centraux de notre étude sur l'évaluation des performances dans des environnements de ressources limitées.

2.1. **Les modèles de langue : fondamentaux.**

Les modèles de langue constituent une technologie clé du traitement automatique du langage naturel (NLP). Leur rôle est d'estimer la probabilité qu'une séquence de mots apparaisse dans un langage donné. Ces modèles sont largement utilisés dans diverses applications, de la traduction automatique à la génération de texte.

2.2. **Contraintes de ressources : un défi majeur.**

Dans notre étude, nous nous intéressons spécifiquement à l'évaluation des modèles de langue dans des environnements où les ressources sont limitées. Cette contrainte de ressources peut se manifester sous forme de volumes restreints de données d'entraînement ou de capacités de calcul limitées.
Comprendre ces concepts fondamentaux est crucial pour appréhender notre approche méthodologique et les résultats que nous présentons dans la suite de cette note.

3. Méthode expérimentale

Cette section décrit de manière concise notre approche méthodologique pour évaluer les performances de nos modèles de langue.

3.1. **Sélection des données d'entraînement.** Nous avons utilisé des sous-ensembles de pages Wikipédia en anglais pour constituer nos ensembles de données d'entraînement. Ces ensembles variaient en taille, représentant des volumes de données allant de 1 à 15 Go (en ordre de grandeur).

3.2. **Choix des modèles de langue.** Nous avons sélectionné un coeur de modèle unique, GPT-2 (Generative Pre-trained Transformer 2), en le dotant d'une gamme de configurations (de choix de paramètres) différentes. Nous avons dû concilier nos limitations en ressources de calcul avec l'inévitable précision requise pour obtenir des résultats intéressants.

3.3. **Méthodes d'évaluation.** Nous avons évalué les performances des modèles de langue à l'aide de métriques standard du NLP telles que la perplexité.
Cette approche méthodologique nous a permis de mener des expériences rigoureuses et comparables pour évaluer les performances de nos modèles de langue. Les résultats de nos expérimentations sont présentés dans la section suivante.

4. Résultats et conclusions

Nous ne présentons dans cette section que deux méthodes d'évaluation de nos modèles, par souci de concision et de clarté.

- D'une part, en lui soumettant des textes non utilisés pendant la phase d'entraînement, provenant de CNN et estimant sa capacité à prédire la fin d'un texte connaissant le début.
- D'autre part, en lui soumettant une liste prédéfinie de phrases ou débuts de phrases à compléter (des prompts), de longueur et complexité variables.

4.1. **Estimation quantitative.**
Pour la première évaluation, une métrique couramment utilisée pour évaluer la performance du modèle de langue sur un nouveau corpus est la perplexité, qui s'exprime sans unité.
La perplexité est, du point de vue du modèle, la probabilité de la séquence de texte naturel qui lui est présentée. Une perplexité basse signifie que le modèle considère que le vrai langage est vraisemblable. Pour préciser un peu„ il est considéré la probabilité de tout le texte, sous forme de probabilité jointe de tous les tokens.
Retenons enfin que la perplexité n'est qu'un indicateur parmi d'autres de la qualité du modèle, une façon nécessairement imparfaite de la mesurer.
La plupart des modèles que nous avons entraînés ont une perplexité comprise entre 400 et 500, indiquant une performance moyenne à acceptable, à l'exception du plus volumineux qui monte à 700, indiquant une performance dégradée. L'analyse qualitative ci-après confirme cette première impression.

4.2. **Estimation qualitative.**
En ce qui concerne la deuxième partie, par défaut le modèle entraîné se répète beaucoup et très rapidement. De nombreuses occurrences consécutives du mot *the* ou parfois de *of the United States* (pour le prompt *In the State of California*) ont été observées.
Un paramètre à spécifier lors de la génération du modèle permet de limiter la répétition dans la réponse. On peut ajuster la longueur de texte (en tokens) durant laquelle la répétition est interdite.

Une fois cela fait, les réponses sont syntaxiquement bien plus satisfaisantes, bien que leur sémantique puisse encore laisser à désirer. Nous donnons ci-dessous un exemple typique. Le prompt que nous avons donné est `"In a galaxy far far away..."`, la suite est la réponse du modèle.

```
  "In a galaxy far, far away...": "In a galaxy far, far away... there are no visible
differences in any of its visible stars, but no more than the luminosity of a galaxy.
\"\n\n\"The luminosity of galaxies is a natural phenomenon\" (a) or \"massive\" in which
galaxies are absorbed by a single atom; the \"massive black\" nuclei and stars are absorbed
into stars by the two. These are only scattered by a small fraction of the stars, and the
galaxies are absorbed by the Milky Way galaxy.\n\nThe universe is not generally",
```

Analysons la réponse du modèle à notre prompt.

1. **Cohérence thématique.** Le modèle commence par une référence à l'expression emblématique "In a galaxy far, far away…" associée à l'univers de Star Wars, ce qui évoque immédiatement un contexte spatial et fantastique. Cependant, la réponse prend rapidement une tournure plus scientifique en abordant le concept de luminosité des galaxies et la structure de l'univers, ce qui crée une légère dissonance thématique.
2. **Complexité du langage.** Le texte généré utilise un langage relativement complexe et technique, avec des termes comme "luminosity", "galaxies", "massive black", et "Milky Way galaxy". Cela suggère que le modèle a une certaine compréhension des concepts astronomiques, bien qu'il puisse parfois les utiliser de manière inattendue ou incorrecte.
3. **Manque de cohérence.** Malgré la présence de termes scientifiques, le texte manque de cohérence et de logique dans sa structure et son développement. Les phrases semblent être assemblées de manière disjointe, avec des transitions abruptes entre les idées. Par exemple, la transition entre la luminosité des galaxies et la "massive black" est abrupte et peu naturelle.
4. **Conclusions ambiguës.** La dernière phrase "The universe is not generally" est une conclusion ambiguë qui n'est pas la suite logique des idées précédentes. Cela met en doute la capacité du modèle à fournir une conclusion cohérente et satisfaisante en réponse au prompt donné.

La fin abrupte de la réponse est un artefact de notre méthode de génération : nous donnons au modèle une longueur fixée à l'avance pour sa réponse, engendrant ce type de coupure.

5. Limitations et perspectives

Bien que notre étude ait apporté des éclaircissements sur l'efficacité des modèles de langue dans des environnements de ressources limitées, il est important de reconnaître plusieurs limitations inhérentes à notre approche et à nos résultats.

1. **Biais des données d'entraînement.** Notre modèle de langue a été entraîné exclusivement sur des pages Wikipédia en anglais, ce qui pourrait introduire des biais dans ses performances et sa capacité de généralisation. Les données de Wikipédia peuvent ne pas représenter pleinement la diversité des styles de langage et des domaines de connaissances, limitant ainsi la portée de nos conclusions.
2. **Généralisation limitée.** En raison de la nature spécifique de nos données d'entraînement, notre modèle pourrait avoir des difficultés à généraliser au-delà du domaine couvert par les

articles de Wikipédia. Cela pourrait limiter son utilité dans des applications réelles où une compréhension diversifiée et contextuelle du langage est requise.

3. **Taille limitée du modèle.** Nous avons utilisé un modèle de langue de taille modérée (GPT-2) en raison des contraintes de ressources. Cependant, cette taille limitée pourrait avoir un impact sur les performances du modèle, en limitant sa capacité à capturer des nuances subtiles du langage et à produire des réponses plus précises et pertinentes.

6. Conclusion

En résumé, bien que notre étude ait contribué à la compréhension des performances des modèles de langue dans des environnements de ressources limitées, ces limitations soulignent la nécessité de la prudence dans l'interprétation des résultats et de la poursuite de la recherche pour améliorer la robustesse et la généralisation de ces modèles.

Bibliographie

[1] D. Friedman, A. B. Dieng, The Vendi Score: A Diversity Evaluation Metric for Machine Learning, Transactions on Machine Learning Research (n.d.). https://par.nsf.gov/biblio/10427561

[2] G. Xu, J. Li, G. Gao, H. Lu, J. Yang, D. Yue, Lightweight Real-Time Semantic Segmentation Network With Efficient Transformer and CNN, IEEE Transactions on Intelligent Transportation Systems 24 (2023) 15897–15906. https://doi.org/10.1109/TITS.2023.3248089

[3] S. Biderman, et al., Pythia: a suite for analyzing large language models across training and scaling, in: Proceedings of the 40th International Conference on Machine Learning, JMLR.org, 2023

[4] A. Vaswani, et al., Attention is All you Need, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017: p. . https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[5] G. Francopoulo, Pruning Texts with NLP and Expanding Queries with an Ontology: TagSearch, in: 2003: pp. 319–321. https://doi.org/10.1007/978-3-540-30222-3_30

[6] R. Tang, Y. Lu, J. Lin, Natural Language Generation for Effective Knowledge Distillation, in: 2019: pp. 202–208. https://doi.org/10.18653/v1/D19-6122

[7] A. P. Pasarkar, A. B. Dieng, Cousins Of The Vendi Score: A Family Of Similarity-Based Diversity Metrics For Science And Machine Learning, (2023)

[8] HuggingFace, Training a causal language model from scratch, (n.d.). https://huggingface.co/learn/nlp-course/en/chapter7/6 (accessed November 22, 2024)

[9] R. Mehrotra, Topic Modelling using LDA and LSA in Sklearn, (2022). https://www.kaggle.com/code/rajmehra03/topic-modelling-using-lda-and-lsa-in-sklearn (accessed February 15, 2024)

[10] Weights and biases, (n.d.). https://docs.wandb.ai/quickstart (accessed March 15, 2024)

[11] OpenAI, GPT-2 partial release statement, (n.d.). https://openai.com/index/better-language-models/ (accessed May 12, 2024)

Département de statistiques, Ensae, Palaiseau