Université de Lorraine - IDMC
Masters SC & TAL – M1
2020-2021

Fiche de projet tutoré / Project form

Model ensembling for NLP

Encadrement / Supervisors
1. SYNALP team, LORIA laboratory
1. main supervisor: Christophe Cerisara cerisara@loria.fr

Description / Description
1. global project: Machine learning models for NLP are typically trained on large text corpora with the help of a GPU cluster. However, this dominant paradigm nowadays faces more and more pressing challenges: e.g., all data has to be accessible from the central GPU cluster, which is not always possible with private or industrial secret data; furthermore, the ecological cost of these GPU farms, as well as the economical cost to maintain and renew them is prohibitive for many actors. A potential and partial solution to these issues consist in training multiple local small models and merge them. Model ensembling is an old set of methods that is again gaining traction to perform such a merging of models. The aim of this project is to answer the following research questions:
- what are the best model ensembling methods when considering modern deep learning models and typical NLP benchmarks ?
- how much does the performance of the resulting model degrades with ensembling, as compared to standard training ?

2. biblio. UE 705 (semestre 7)

The first part of the project consists in studying the literature about model ensembling and also about the current NLP benchmarks and deep learning models that are successful in these benchmarks. For the first part, you will typically look at averaging and stacking methods, among others. Note that because we assume that each model does not have access to the full dataset, some ensembling methods are not applicable. For the second part, you will typically investigate modern NLP benchmarks such as SentEval and/or SuperGLUE and/or XNLI and/or MLQA, and study whether and how they can be adapted to the proposed decentralized approach.

3. réalisation. UE 805 (semestre 8)

For the implementation, you will reuse the existing code of both the benchmarks and models; this code is in python and often in pytorch. You will need to master python, but it is not a requirement to understand the details of pytorch, as you will mostly reuse existing code. You will have though to manipulate the inputs and outputs of the baseline models, retrain multiple times these models and their combinations, compare and analyze the results with regard to the state-of-the-art.

Informations diverses : matériel nécessaire, contexte de réalisation /
Various information: material, context of realization

The student will need a personal laptop with python and pytorch installed. If and when the laptop will not be powerful enough to run multiple experiments, then I'll give the students access to other machines.

Livrables et échéancier / Deliverable and schedule

Milestones:
M1- Biblio report on model ensembling
M1- Biblio report on modern NLP benchmarks and their "best" models
M2- Download, test and try to reproduce the published results for 1 or 2 such open source benchmarks
M3- Split the data into N parts, retrain locally, try and merge models with 1 or 2 model ensembling methods, compare
M4- Final report and code

Bibliographie /References (max. 4-5)
[*il ne s'agit pas de la bibliographie complète qui sera fournie aux étudiants au début du projet mais d'une bibliographie indicative pour aider à cerner le sujet*]

Benchmarks:
- https://arxiv.org/pdf/1911.02116.pdf
- https://arxiv.org/abs/1803.05449
- https://super.gluebenchmark.com

Model ensembling:
- https://machinelearningmastery.com/stacking-ensemble-for-deep-learning-neural-networks/
- https://arxiv.org/abs/1902.11175