Semi-supervised Learning : projet de fin d'année

1 Contexte

Les années 2022 et 2023 sont marquées par l'apparition de datasets publics énormes. Le projet LAION a par exemple récemment publié un dataset contenant plus de 5 milliards d'images et de descriptions appairées. Dans ce contexte précis, il semblerait que les méthodes les plus efficaces semblent d'entraîner des modèles Transformers contenant des dizaines de milliards de paramètres et de parvenir à les entraîner de manière distribuée. L'idée de ce sujet est de réussir à exploiter ces modèles *Large Foundational* dans un contexte où la donnée labéllisée manque.

C'est précisément le cas du *few shot learning* dans lequel le nombre d'échantillons labelisés est petit, et où l'objectif reste quand même de minimiser le *generalization gap*. Dans ce sujet, ce petit échantillon labellisé va être également combiné avec un échantillon nettement plus large (x200) de données non-labellisées: on est dans le cadre du *semi-supervised learning*.

On vous propose de travailler sur le jeu de données STL10. Il s'inspire de l'ensemble de données CIFAR-10, mais avec quelques modifications. En particulier, chaque classe du dataset comporte moins d'exemples d'entraînement étiquetés que dans CIFAR-10, mais il contient un très grand ensemble d'exemples non étiquetés. Le défi principal consiste à utiliser les données non étiquetées (qui proviennent d'une distribution similaire mais différente des données étiquetées) pour construire une connaissance préalable utile. Les caractéristiques du dataset sont les suivantes:

- 10 classes: airplane, bird, car, cat, deer, dog, horse, monkey, ship, truck.
- des images en couleur avec une définition de 96x96 pixels, color.
- 500 images en training(10 pre-defined folds), 800 images en test par class (8000 au total).
- 100000 images non-labellisées.

2 Ce qu'il vous est demandé

Vous pouvez optionnellement travailler seul, en binôme ou en trinôme, à votre convenance. Vous devrez rendre un rapport technique sous la forme d'un notebook auto-contenu (soit exporté en pdf, soit un fichier .ipynb, soit sous la forme d'un partage colab). Votre rapport devra traiter les points suivants :

1. Entrainement **from scratch** d'un classifieur à convolutions entraîné sur le subset de STL avec labels. Ceci fournira une baseline pour la suite de vos expériences.
2. Proposition et étude d'une méthode permettant d'améliorer cette baseline, en s'inspirant (ou non) des pistes proposées en Section 3.

Votre rapport sera un document scientifique, c'est-à-dire:

- qu'il devra être écrit dans un bon français ou anglais,
- que vous devrez articuler une vraie démarche de réflexion, présenter vos choix et idées,
- qu'il ne doit pas être une simple succession de blocs de code mais inclure des commentaires sur les principales expériences, sur vos choix, etc,

- qu'il contiendra un abstract (qui doit résumer votre travail en quelques lignes), une introduction (pour présenter la problématique, et les différentes solutions déployées) et une conclusion (qui doit résumer votre travail et vos principaux résultats).

En outre, on vous demandera de limiter votre travail à une trentaine de pages (ou équivalent si vous ne rendez pas un pdf). Cette consigne est là pour vous éviter de produire des rapports trop longs, moins bien structurés, et pour vous amener plutôt à articuler une recherche plus réfléchie. Enfin, faites aussi attention à ne pas laisser de prints ou logs inutiles dans vos rapports (par exemple, montrez plutôt une courbe synthétisant l'évolution d'une training loss, plutôt que les prints de chaque époque).

3 Quelques pistes

Le choix des modèles et des métriques vous revient. Nous noterons avant tout la qualité et la rigueur du protocole expérimental proposé, plutôt que les performances brutes obtenues. Nous attendons donc une justification de la méthode choisie ainsi qu'une description détaillée. Nous vous proposons ces quelques idées:

1. Entraîner d'un réseau convolutionnel et/ou avec mécanisme d'attention (optionnel) sur le dataset labellisé.
2. Exploiter un réseau pré-entraîné et l'utiliser pour extraire des features. Par exemple: Masked auto-encoders, CLIP, OpenCLIP. Puis réaliser du *linear probing*.
3. *Self-supervised learning*: exploiter les données non-annotées (autoencoder par exemple).
4. *Self-training*: utiliser les prédictions d'un modèle pré-entraîné sur les données annotées pour labelliser les données non-annotées.

Pour l'évaluation: On attend une discussion sur le choix d'une ou plusieurs métriques bien choisie ainsi que des commentaires sur les résultats obtenus. Une discussion sur le lien entre le choix de l'erreur d'entraînement et la métrique d'évaluation est aussi attendue. L'utilisation de modèles de Computer Vision de types Vision Transformer semble possible sur ce dataset même en utilisant de simple Google Colabs, comme l'atteste notre benchmark en utilisant ce notebook: 80 inférences en moins de 0.2 secondes.

Pour le dataset: En ce qui concerne le dataset, on vous propose d'utiliser le dataset STL10. On peut le retrouver sur Pytorch et Tensorflow facilement, (STL10 with Tensorflow, STL10 with Pytorch). Une discussion sur l'utilisation des données (splitting, etc...) est attendue.

4 Axes d'évaluation

Clarté et organisation Comme indiqué ci-dessus, le notebook doit être organisé comme un rapport scientifique à part entière (abstract / conclusions / interprétation des résultats / recommandations)

Executabilité Le notebook doit être executable dans la mesure du possible. Dans certains cas (utilisation de données externe, modèle long à apprendre), il est légitime que ce ne soit pas le cas mais cela devra être indiqué clairement.

Démarche scientifique Nous laissons à votre discrétion le choix des métriques, du processus d'évaluation mais celles-ci doivent être motivées (aka train / test ou train / test / val ; auc vs accuracy ; etc...). Une prise de recul sur les données et sur les risques sera fortement appreciée.

Il est plus important d'exposer une bonne démarche que des bons modèles.

Modélisation Vous êtes libre du choix des modèles, du moment qu'ils restent pertinents et justifiés.

Résultats et métriques Il sera important que vous preniez du recul face à vos résultats, et que vous indiquiez clairement comment vous allez évaluer les performances vos modèles.

Il est tout à fait possible de proposer d'autres pistes dans ce projet. Nous valoriserons ces propositions sous la forme de points supplémentaires (mais il est conseillé de s'assurer que vous répondez d'abord bien aux cinq points ci-dessus qui suffiront à valider le cours).