Deep Learning for Images: projet de fin d'année

1 Contexte

La reconnaissance faciale est un thème qui revient sur le devant de la scène en France (cf ce récent article dans Les Echos qui traite du débat de l'utilisation de la reconnaissance faciale en France, notamment dans le cadre des Jeux Olympiques de 2024). En outre, bien que la reconnaissance faciale soit un domaine relativement ancien, le domaine des attaques adverses sur ces systèmes attire toujours de nombreuses études.

Le but de votre projet sera de réaliser et d'évaluer un système d'attaque adverse dans le domaine de l'identification faciale, dans un contexte closed-set (*i.e.* on n'évaluera pas les modèles sur un jeu de données 'test'). Pour ce sujet, une attaque adverse consistera en la génération d'images visuellement très similaires aux images de la gallerie de training, mais parvenant à tromper le modèle et à réduire sa performance. Pour cela, vous devrez présenter et utiliser (au moins) deux modèles distincts :

1. un classifieur **C** (qui sera par la suite le système à feindre/attaquer).
2. un modèle adverse **A** qui viendra attaquer le classifieur **C**.

2 Ce qu'il vous est demandé

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 répondre (ou discuter au moins) des questions suivantes :

1. Entrainement d'un classifieur *ad-hoc* spécifiquement entraîné sur un dataset de visages (par exemple LFW), et la comparaison avec un classifieur *pre-trained* (par exemple VGG in keras). On attend également une discussion concernant l'évaluation de ces modèles.
2. En ce qui concerne les attaques adverses, nous attendons à ce que les attaques soient visuellement non-discernables à l'oeil nu. Ceci implique une discussion autour des limites autorisées en termes de distances sur les images (norme infinie, L2, ou autre).
3. Enfin, nous attendons également une discussion concernant l'évaluation de la qualité de ce modèle adverse **A**, ainsi qu'une comparaison de ces performances sur les modèles *ad-hoc* et *pre-trained*.

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, datasets, métriques, etc vous revient, ceci dit nous vous conseillons de vous restreindre pour les classifieurs à des modèles appris via des loss de type softmax. Nous vous proposons ces quelques idées:

Pour les classifieurs:

- **DeepFace**: vous pouvez, en particulier, utiliser ce blogpost, qui fournit des pointeurs de code en keras pour charger l'architecture et les poids du modèle.
- **Arcface**: ou bien, vous pouvez utiliser ArcFace et cette implémentation pytorch réalisée par les auteurs.

Pour l'évaluation:

- Qualité du classifieur **C** : à minima, l'accuracy du classifieur.
- Qualité de l'attaquant **A** : à minima, l'accuracy du classifier attaqué.

Pour les datasets:

- LFW: que vous avez déjà vu en TP,
- Megaface: qui contient plusieurs millions de paires labellées.

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 (white-box, adversarial, ou autre), 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).