Projet HAI923
« Réalisation d'un modèle CLIP Image-Texte »

Date limite pour rendre le travail : la date de remise sera précisée par mail sur Moodle.

Objectif :
Les premiers modèles d'IA géraient une seule modalité à la fois : des modèles pour l'image d'un côté, d'autres pour le texte. Récemment, des modèles multimodaux comme CLIP [Radford et al., 2021][1] sont apparus : ils prennent en compte plusieurs modalités, c'est-à-dire, par exemple, le texte et l'image. Ils permettent de retrouver, à partir d'une image, les textes qui la décrivent, et inversement, à partir d'un texte, l'image la plus pertinente. Ce principe est utilisé dans des systèmes comme DALL·E pour générer des images à partir d'une description. Outre la recherche "image-texte", CLIP peut classer des images en les comparant directement à des étiquettes textuelles, sans ré-entraîner de modèle (c'est ce qu'on appelle du "zéro-shot"). Il peut aussi aider aussi à mieux ordonner les résultats et à nettoyer un jeu de données (dédoublonnage, hors-sujet).

L'objectif du projet est de réaliser un petit modèle CLIP à partir d'un jeu de données Flickr fourni. À l'issue du projet, vous proposerez une image et le modèle renverra une description correspondante et inversement, en entrant une description textuelle, il retrouvera l'image la plus adaptée. Par exemple en donnant comme texte : « A big dog in the woods », vous devrez obtenir quelque chose comme l'illustre la Figure 1.

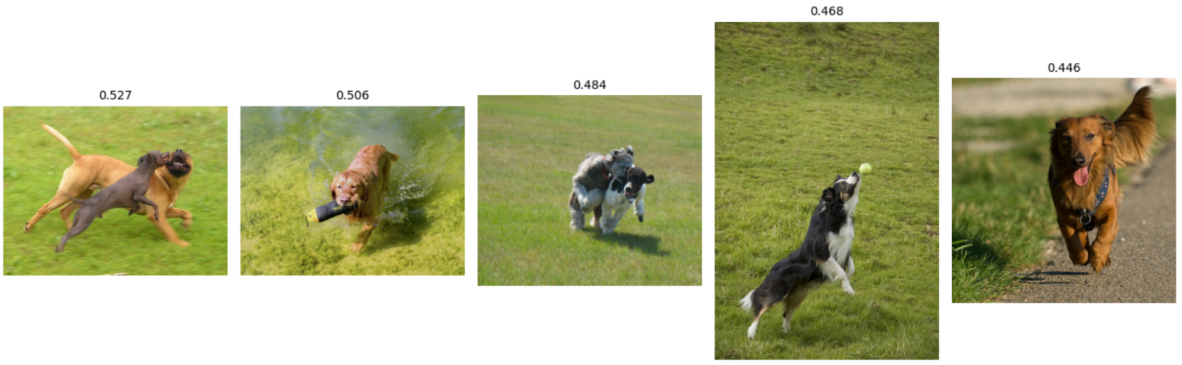

Figure 1 - Un exemple de texte et d'images retournées

Remarques :

- Pour rappel différentes ressources et guides avec les codes associés sont à votre disposition ici : https://gite.lirmm.fr/poncelet/deeplearning/
- Pour réaliser votre projet, il est important d'avoir bien compris les notions d'espace latent (voir « Guide pratique de la vision par apprentissage profond », les notions de transformer (voir TP précédent et « Guide pratique de l'apprentissage profond de données textuelles »).

[1] A. Radford et al. Learning Transferable Visual Models From Natural Language Supervision (CLIP). ICML 2021. - https://arxiv.org/pdf/2103.00020.

- Outre, le fait de vous permettre d'accéder aux données, le fichier « ProjetClip.ipynb » vous offre de nombreuses astuces.
- **De l'usage de l'IA dans votre projet.** Le projet a été contraint à plusieurs endroits pour éviter que vous n'utilisiez une réponse issue de l'IA. Nous avons également fait faire par plusieurs IA le même projet et nous savons ce qu'elles génèrent. Lors de la correction, si nous avons le moindre doute qu'une partie des travaux (rapport ou code) a été réalisée avec un usage abusif d'IA, un oral aura lieu pour l'équipe concernée. En ce qui concerne le rapport, nous autorisons l'utilisation de l'IA pour vérifier les erreurs de syntaxes ou de mauvaises formulations mais en aucun cas nous n'autorisons l'utilisation de l'IA pour rédiger des parties. N'oubliez pas que nous avons l'habitude et que nous travaillons dans le domaine donc nous repérons vite.

Le jeu de données que nous allons utiliser pour le projet a été créé à partir de données de la plateforme de partage de photos Flickr[2]. Nous avons retenu 4 catégories (« bike », « ball », « water », « dog »). Chacune des catégories contient 150 images et textes associés.

N'oubliez pas que pour vous aider le notebook « ProjetClip.ipynb » contient de nombreux codes et conseils. Il faut le lire attentivement.

Le projet est à réaliser par équipe de 4 personnes[3]. Avant de commencer vous devez inscrire votre groupe sur le lien suivant :
https://docs.google.com/spreadsheets/d/1y7EP1ev29xr7UxKpD5HD4IFhTQkzEpL1R3RuSuP8tfA/edit?usp=sharing
Attention :

- La personne qui apparaît sur la première colonne sera chargée de déposer le projet pour tous les membres du groupe. A partir du moment où vous serez inscrit vous aurez un numéro de projet qui correspond à celui du numéro indiqué sur le fichier. C'est ce numéro qui vous permettra d'identifier les fichiers à rendre.
- Toute personne non inscrite sur le fichier 15 jours après le démarrage du projet aura 0 à l'UE. Donc pensez tout de suite à vous inscrire pour ne pas laisser passer le temps.

Travail demandé :

1) **Réaliser un classifieur d'images pour les 4 classes**

L'objectif ici est de mettre en place un modèle de CNN qui soit capable de faire de la classification d'images.
Attention : l'objectif n'est pas d'obtenir le meilleur classifieur mais plutôt d'avoir un premier modèle qui pourra être utilisé par la suite. Donc ne perdez pas de temps à essayer d'optimiser votre classifieur. Par la suite nous enlèverons justement la partie « classification » de votre modèle.

2) **Réaliser un classifieur de textes pour les 4 classes.**

---

[2] https://www.flickr.com (dernière consultation octobre 2025).
[3] Il peut être autorisé exceptionnellement d'avoir un nombre différent mais pour cela il est impératif d'avoir l'aval de vos encadrants au préalable.

L'objectif est d'utiliser le modèle « SmallBert » qui est disponible dans le guide « Guide Pratique de l'Apprentissage Profond de Données Textuelles »[4] afin de créer un classifieur mais cette fois-ci sur les textes.

Attention :

- Comme précédemment l'objectif n'est pas de perdre de temps à rechercher le meilleur classifieur mais d'avoir un premier modèle qui servira par la suite.
- Contrairement à BERT, SmallBert ne contient pas de token spécial <CLS> il faudra donc faire attention pour savoir comment « résumer » une phrase pour votre classifieur.
- Réaliser un classifieur de textes pour les 4 classes.

3) Réaliser un modèle CLIP

La Figure 2 illustre ce que vous devez réaliser. Voilà ce qu'il faut réaliser. Les images et les textes vont passer via un encoder afin d'obtenir des embeddings et le principe de CLIP est de faire une matrice où « les images qui correspondent à un texte » se retrouve proches dans l'espace latent. Comme ça quand on propose une image on peut trouver les textes.

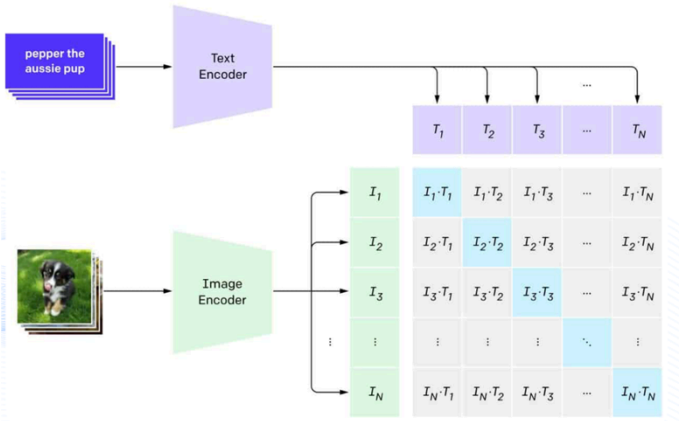

Figure 2 - Le principe d'un modèle CLIP (source : [A. Radford et al.])

A ce niveau, vous avez déjà réalisée une bonne partie pour l'image et une bonne partie pour le texte.

- Pour l'image : définissez un nouveau modèle inspiré du précédent dans lequel vous enlevez toutes les couches de classification (i.e. toutes les dernières couches jusqu'à la couche flatten comprise) pour que le modèle retourne uniquement un vecteur d'embeddings (en d'autres termes la sortie devient simplement une couche dense de la taille des embeddings que vous voulez mais sans fonction d'activation).
- Pour le texte : comme précédemment il faut définir un nouveau modèle inspiré de celui que vous avez fait dans la partie classification. Ici également la sortie du modèle doit produire uniquement un vecteur d'embeddings (donc la sortie devient simplement une couche dense que la taille des embeddings que vous voulez sans fonction d'activation.
  Attention, vous voyez sur la figure que les dimensions des embeddings doivent être similaires. En outre, pour pouvoir être assemblés il faut que leurs sorties soient normalisées.
- Création du modèle CLIP : il vous faut créer un modèle CLIP qui combine les deux modèles précédents.

[4] Voir https://gite.lirmm.fr/poncelet/deeplearning/ (dernière consultation octobre 2025).

- La loss pour des modèles CLIP est une loss contrastive. L'intuition de cette loss est qu'elle rapproche les couples image–texte qui vont ensemble et éloigne ceux qui ne correspondent pas. Ne vous inquiétez pas le code de la loss contrastive est donné. Par contre, il ne faut pas oublier de l'intégrer à votre modèle.
- Apprenez votre modèle et sauvegardez-le. Attention pensez à bien vérifier sur un premier test que votre chaîne pour sauvegarder est valide sinon vous ne pourrez pas recharger votre modèle.
- Faites de l'inférence avec votre modèle. Donnez-lui un texte et regardez les images obtenues. De la même manière donnez-lui une image et regardez les textes obtenus. Pour les deux tests vous retournerez 5 réponses (top-k=5) et afficherez le score de la réponse.

4) **Travail facultatif** :
    a. A faire que si toutes les étapes précédentes ont été réalisées. Vous pouvez si vous le souhaitez remplacer SmallBERT par DistilBERT (attention à vos dimensions et attention il faut utiliser le tokenizer de Bert DistilBertTokenizerFast sous peine d'avoir des problèmes d'alignement) et voir si cela améliore la qualité du résultat.
    b. Les réponses ne sont pas forcément toujours bonnes car les textes sont très courts. Vous pouvez tout à fait les enrichir. Là oui vous pouvez vous faire aider par un LLM pour étendre vos phrases en veillant à ce qu'elles conservent quand même la même sémantique. N'oubliez pas que les textes sont dans le fichier caption.csv mais aussi dans le répertoire caption.

Travail à rendre :

Attention : Le non-respect des consignes[5] (*nom des fichiers, format, fichier Excel des membres du groupe et de la personne qui doit remettre non rempli avant l'évaluation, fichier manquant – rapport, ipynb, pdf, video, nombre de pages du rapport, …*) sera pénalisé de **-4**. Vous êtes en dernière année, vous allez bientôt aller dans le monde professionnel. Il est vraiment impératif que vous soyez conscient qu'il faut respecter les consignes qui vous sont données.

La personne chargée de déposer le travail devra déposer sur Moodle :

- **Un fichier zippé** identifié par le numéro du groupe. Par exemple, si vous êtes le groupe 1, le fichier doit s'appeler : « 1.zip ».
  Ce fichier doit contenir :
    - Le rapport au format pdf dont le nom sera identifié par le numéro de groupe. Par exemple si vous êtes le groupe 1, le fichier devra s'appeler « 1.pdf ».

[5] La liste n'est pas exhaustive. Les éléments qui sont marqués concernent ce que nous avons constaté les années précédentes en M1 et M2. Mettez-vous simplement à notre place lorsque nous avons des rapports qui s'appellent « rapport.pdf » sans les numéros et nom d'étudiants, des fichiers compressés avec des formats exotiques qui ne se décompressent qu'avec un type de machine, des fichiers manquants et que nous passons des heures à essayer de trouver une solution ou d'essayer de savoir qui est l'auteur de rapport.pdf…. Ce n'est pas normal. **Vous devez vraiment apprendre à respecter les consignes !!** dans quelques mois si votre responsable vous demande un fichier en pdf et que vous lui donnez en .ipynb et qu'il se retrouve à devoir présenter devant des clients … Imaginez sa tête .. et ce qu'il vous dira après. Donc cette année nous avons décidé de vous pénaliser pour que vous en soyez bien conscient.

Le rapport est à rédiger en Latex en suivant le template disponible ici : https://www.lirmm.fr/~poncelet/Ressources/template_projet.zip.
Le template est classique, il impose juste la taille des marges.
Il doit faire 8 pages maximum. Il est possible d'utiliser des annexes pour un maximum de 2 pages pour toutes les annexes.
Remarque : il est conseillé d'utiliser pmllatex (https://plmlatex.math.cnrs.fr/login) une instance d'Overleaf gérée par le CNRS et qui offre beaucoup de liberté notamment pour partager le document. Il faut dans ce cas utiliser votre adresse institutionnelle.
Il y a plusieurs parties dans le projet, vous avez toute liberté pour présenter le mieux possible le travail réalisé. Surtout ne perdez pas de temps à décrire l'objectif ou les données car tout le monde les connaît. Focalisez-vous sur ce que vous avez fait pour le valoriser.

- Le ou les notebooks aux formats « .ipynb » et « .pdf ». Le notebook devra être préfixé par le numéro de votre groupe et le caractère « _ » et après vous pouvez l'appeler comme vous voulez. Par exemple, si vous êtes le groupe 1, les fichiers .ipynb et .pdf doivent forcément commencer par « 1_ » (e.g « 1_monprojet.ipynb », « 1_monprojet.pdf »).

Attention : tous les fichiers doivent contenir le nom, prénom et numéro de carte d'étudiants de chacun des membres du projet.

Quelques conseils :

- Le projet n'est pas forcément simple car vous allez devoir manipuler des modèles un peu complexes. Il est donc impératif de ne pas traîner. Les deux premières questions sont assez simples mais la création du modèle CLIP nécessite de comprendre pas mal de choses. Nous vous aidons sur des points difficiles mais aider ne veut pas dire « faire pour vous ».
- Pensez à bien lire le sujet pour ne rien oublier, pensez à bien lire tout ce qui est les compléments qui vous sont donnés, pensez à bien lire les parties sur les « guides » qui vous sont conseillés.
- Pensez à bien vous répartir le travail car les délais sont courts volontairement.