Université de Paris

UFR Mathématiques et Informatique

Les modèles pré-entrainés de traitement de données multimodales (image-texte) et leurs applications

Etat de l'art

31 octobre 2022

Master 2 Vision Machine Intelligente

Année universitaire 2022 – 2023

Table des matières

1. Introduction

L'un des premiers but de l'intelligence artificielle est d'imiter l'intelligence de l'homme, et ce par le biais d'algorithmes exécutés dans un environnement informatique. Une des approches dans ce domaine est de s'inspirer du fonctionnement du cerveau humain, plus particulièrement du neurone qui constitue l'unité fonctionnelle de la base du système nerveux., c'est le biomimétisme.

Avec l'essor de l'apprentissage profond, depuis le début des années 2010, l'IA s'oriente vers la conception de réseaux de neurones afin d'atteindre cet objectif. Initialement élaborés avec un unique type source de données (texte, image, voix), des données unimodales, ces algorithmes arrivaient à résoudre des problèmes complexes et obtenir de résultats satisfaisant : reconnaissance d'images, reconnaissance vocales e.g. Toutefois, nos expériences en tant qu'être humain sont multimodales, les données que nous recevons sont issues des sources multiples. Nous obtenons ces données, les combinons et y procédons afin d'en retirer une information et une représentation : nous apprenons du monde réel.

Malgré les progrès considérables réalisés avec les modèles unimodaux, ils ne sont pas suffisants pour couvrir l'ensemble des aspects de l'apprentissage humain. Très vite , nous apercevons la nécessité d'aborder un nouveau paradigme dans l'IA : passer d'une approche unimodale à une approche multimodale.

De ce fait , élaborer des modèles d'apprentissage profond à partir de données multimodales nous permet de réaliser davantage de tâches. A titre d'exemple, générer une image à partir d'un texte grâce à un réseau de neurones , celui-ci apprend un concept à partir de plusieurs types de données.

Ce changement de paradigme s'est manifesté ces dernières années, plus particulier dans le domaine de la vison par ordinateur, avec l'avènement de modèles fondateurs que nous aborderons dans cet état de l'art ainsi que leurs applications.

2. Contexte

Dans cet état de l'art nous nous focaliserons sur le traitement des données multimodales image-texte , *vision-language* (VL) par des modèles pré-entrainés.

Le pré-entrainement , ou *pre-training*, est une stratégie d'apprentissage qui permet d'entraîner la totalité d'un réseau (ou une partie) sur une tâche et des données différents de son utilisation finale. Ainsi, avec un modèle obtenu sur une tache donnée, ses paramètres (ou la totalité du modèle) sont utilisés afin de réaliser une autre tâche (*downstream taks*, nous utiliserons le terme tâche cible) : il s'agit de l'apprentissage par transfert, *transfer learning* .Par exemple utiliser un classificateur d'oiseaux pour classifier les différentes espèces de fleurs.

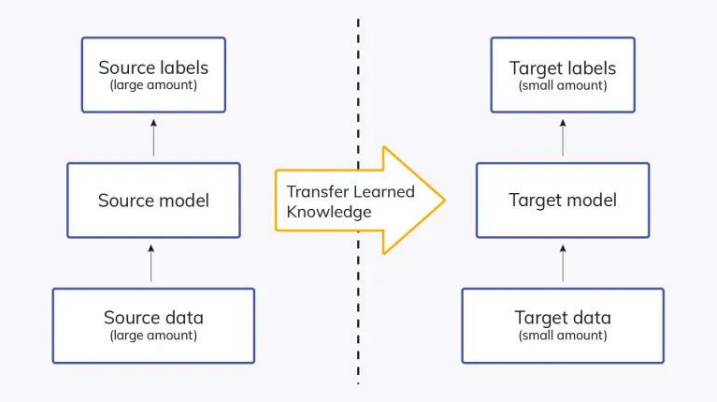

Figure 1 : Illustration du concept de transfert learning [1]

Il existe plusieurs configuration possible pour décliner le modèle de base : *fine-tunning, few—shot learning* , ou encore *zero-shot learning (ZSL).*

Ce pré-entraînement est généralement réalisé sur de très grandes bases de données type ImageNet, CIFAR-10, Oxford-IIIT Pet Images Dataset e.g. Grâce à l'accès en open-source de ce type de données , et l'amélioration des puissances de calculs avec l'arrivée des GPGPU, ces modèles pré-entrainés ont permis à l'apprentissage profond de prendre un nouveau tournant et espérer se rapprocher davantage vers une meilleure généralisation des modèles qui apprennent de concepts communs.

Il faut également souligner le fait que les modèles VL nécessitent des données avec un pré-traitement plus important : nettoyage de données , annotation humaine, analyse sémantique. L'entrainement de ces modèles nécessitent des infrastructure importantes et couteuses. Ainsi des modèles tels que CLIP, ALIGN ou encore Florence, ont été élaborés par des grandes entreprises de technologie (respectivement Open Ai, Google, Microsoft) présentent les meilleurs niveau de performance actuels.

3. Les modèles fondations basées sur l'apprentissage auto-supervisé contrasté

Un modèle fondation correspond à un large modèle d'intelligence artificielle entrainé sur une grande quantité de donné. Le modèle qui en résulte est utilisé pour une variété de tâche. Bien que ces modèles ne présentent pas de nouvelles techniques (souvent basées sur techniques d'apprentissage auto-supervisée), l HAI[1] et le CRFM[2] [2] estiment la nécessité de leur attribuer un terme spécifique du fait de leur impact et leur haut potentiel dans le domaine de l'IA. Ainsi ces deux entités décrivent les modèles fondations de la manière suivante : « un nouveau paradigme pour construire des systèmes d'IA […] entrainer un modèle sur une quantité colossale de données, et l'adapter pour d'autres applications ».

Dans le cas de modèle texte-image, il existe actuellement trois modèles fondations, à savoir : CLIP, ALLIGN, et Florence. Ces trois modèles utilisent une méthode d'apprentissage auto-supervisé ,

[1] HAI : The Stanford Institute for Human-Centered Artificial Intelligence's
[2] CRFM : Center for Research on Foundation Models

appelé apprentissage auto-supervisé contrasté, *contrastive sefl-supervised learning* (SSL contrasté).

Le SSL contrasté est une technique qui améliore les performances des tâches de vision. Elle utilise le principe du contraste des échantillons entre eux pour apprendre les attributs communs aux classes de données et les attributs qui distinguent une classe de données d'une autre. Ce mode d'apprentissage, qui imite la façon dont les humains apprennent le monde qui les entoure, a donné des résultats prometteurs dans la littérature sur l'apprentissage profond, mais aussi sur la vision par ordinateur. [3]

Chacun de ses modèles fondations ont été déclinés de différentes manières lors du transfert d'apprentissage. Dans le cadre de cet état de l'art nous allons uniquement aborder la configuration ZSL. Cette configuration permet d'associer des classes observées (lors de la phase pré-entrainement) et non observées (lors de la phase de test) : le modèle peut reconnaitre une classe sans avoir été entrainé dessus au préalable.

En effet, les humains peuvent effectuer un apprentissage semblable au ZSL grâce à leurs base de connaissances linguistiques existante . Cette base fournit alors une description d'un objet qu'ils n'ont jamais observé et ainsi établir un lien entre les objets qu'ils ont déjà observé et leurs base de connaissance [4]. Par exemple, si une personne a déjà observé un cheval, pour lui donner une description d'un zèbre, nous pourrons lui préciser qu'il existe un animal semblable au cheval qui présente des rayures noires et blanches sur son pelage. Lorsque la personne verra un zèbre sur une image, il saura le reconnaître, mais aussi savoir que c'est la première fois qu'il en observe un. Le ZSL est une voie possible pour reproduire cette puissante capacité humaine

3.1 CLIP : le premier modèle fondation

CLIP , pour ***C**ontrastive **L**anguage **I**mage **P**retaining* est un réseau de neurones développé par l'entreprise Open AI et présenté en janvier 2021. Ce modèle a été entrainé dans le but de comprendre les similarités entre les images et leurs légendes. [5]

L'équipe a tenté plusieurs approches au niveau du pré-entrainement .Initialement, le but était de prédire depuis uniquement l'image , une légende en joignant un réseau de neurones convolutif et un Transformer Lagunage Model (TLM) ou un Bag Of Words (BoW) en se basant sur des travaux de Joulin et al.[6](2016) et Mahajan et al. [7] (2018).

Afin de tester et comparer les résultats de ces différentes stratégie de pré-entrainement , ils ont utilisé la méthode ZSL.

En effet avec un TLM (courbe bleu) les mots ainsi que leur ordre sont à prendre en considération pour déterminer la précision du texte. Tandis qu'avec BoW (courbe jaune) , l'ordre ne compte pas , il suffit d'obtenir les bons termes : BoP est trois plus efficace que TLM dans une configuration ZSL.

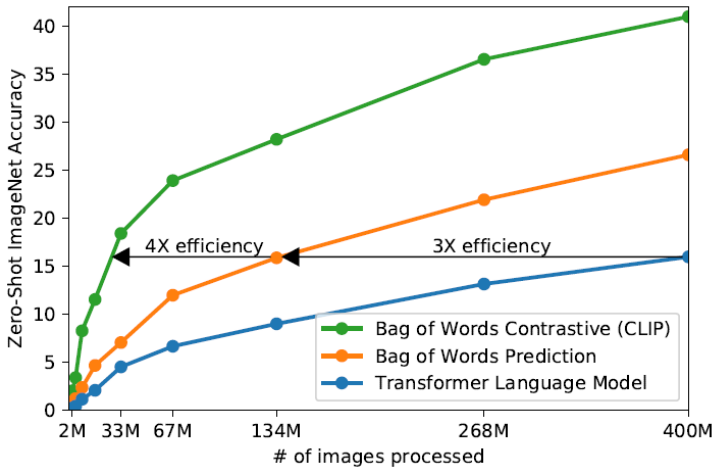

Figure 2 Illustration des différentes stratégies de pré-entrainement[5]

De plus, ces approches de prédictions nécessitent des ressources informatiques importantes. Le TLM avec 63 millions de paramètre mettaient trois plus de temps à reconnaitre les classes d'imageNet qu'un simple BoP. Pour y remédier, les chercheurs ont décidé de réduire les possibilités de sortie de texte, c'est le Bag of Words Contrastive (BoC). Avec une image donnée et 32768 légendes aléatoire, le modèle doit indiquer pour chaque texte , le taux de probabilité qu'elle soit correcte pour l'image en question.

Cette stratégie de pré-entrainement appelé apprentissage auto supervisé contrasté, fut conservé par l'équipe de recherche. CLIP. Ceci a permis de rendre le modèle quatre fois plus efficace (courbe verte sur la figure 2).

Description des données :

Le modèle est entraîné sur un data set de 400 millions de paires images-texte issues d'internet. Un pré-traitement conséquent a eu lieu afin de retirer les doublons, images non appropriés e.g : chaque paire est unique , chaque image a une légende. L'apprentissage est donc supervisé.

Pré-entrainement :

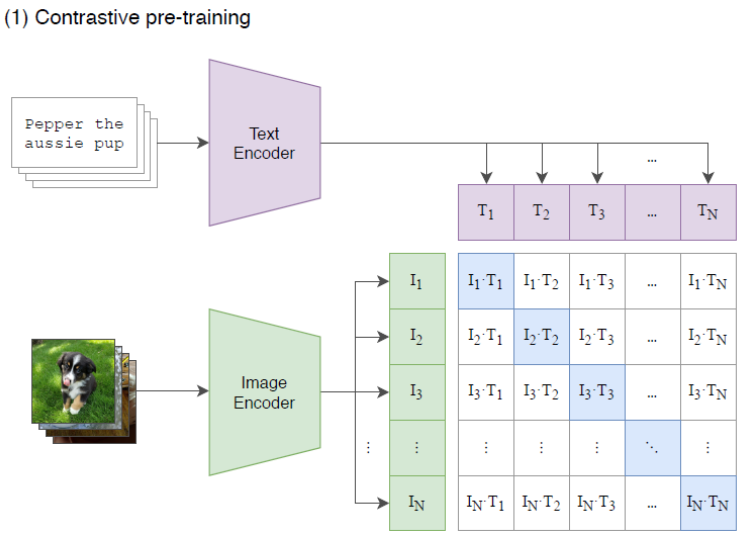

Figure 3 Illustration de l'architecture de CLIP [5]

La paire image-texte passe par deux encodeurs simultanément . Un premier pour l'image (ResNet50)[8] et un second pour le texte (ViT , Vision Transformer) [9]. Ces deux encodeurs permettent de coder ces informations dans un espace d'intégration multimodale. Le modèle doit alors maximiser le score similarité cosinus (cases bleues dans la figure X), tout en minimisant la même métrique pour les images et textes non associés (cases grises blanches). Ceci est réalisé pour la totalité des paires de données. Cette étape du pré-entrainement fut assez long : il a fallu 30 jours (18 jours pour Resnet et 12 pour ViT) d'entrainement sur 592 GPU v100.

Evaluation :

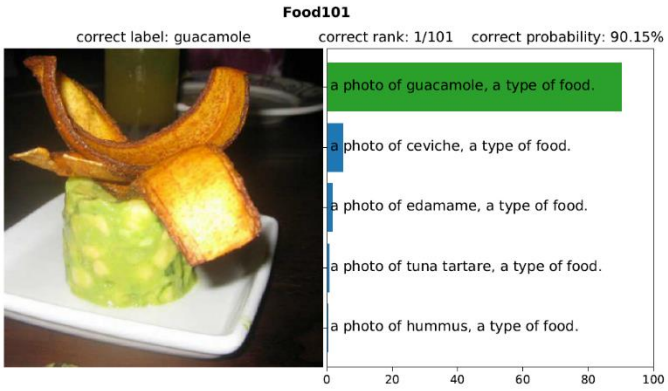

Figure 4 Exemple de proposition de classification pour une image donnée. [5]

Le modèle est ensuite utilisé dans une configuration ZSL lors de certaines évaluation.

Le label correspond à la classe, la vérité terrain est indiqué en vert si elle est correcte (elle est orange dans le cas contraire) et son taux de probabilité est également calculé. Nous pouvons remarquer un *template* au niveau du texte : « a photo of [xxxxx],» . [xxxx] correspond au label que doit déterminer CLIP. Le restant de la phrase est déterminé selon le jeu de donné testé.

CLIP représente des résultats remarquable pour la classification d'images naturels . Ceci s'explique en grande partie par le fait que ce type d'image est plus présente sur le web, et donc également dans le jeu de données utilisé lors du pré-entrainement. CLIP a pu dépasser certaines modèles considérés comme état de l'art.

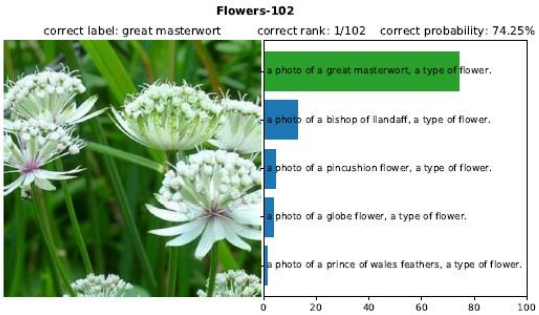

Figure 5 Exemple de classification d'image naturelle (dataset Flowers-102)

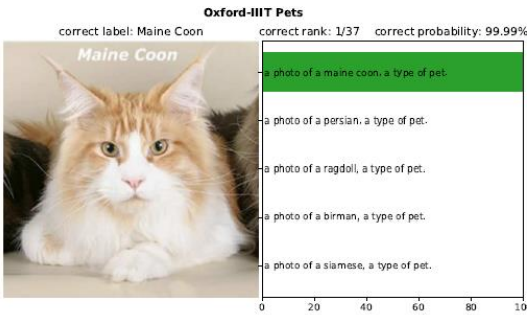

Figure 6 Exemple de classification d'image naturelle (dataset Oxford III Pets)

Néanmoins pour des images plus complexes type imagerie médicales ou satellitaires , CLIP présentent de faibles résultats. Le modèle a également présenté des limites pour des taches plus complexes telles que compter le nombre d'objets sur une image, faire la différence entre deux

objets semblables (par exemple reconnaitre le modèle exacte d'une voiture). A titre d'exemple , sur des la base de données MNSIT, un jeu de données de chiffres écrits à la main, CLIP a atteint uniquement 88% , le meilleur résultat atteint dessus étant 99.82% avec un RMDL[10]

3.2 ALIGN : vers un changement d'échelle

ALIGN , *A Large-scale Image and Noisy-Text Embedding,* est une architecture de réseau de neurones créé par une équipe de recherche de Google et présenté en mai 2021 [11]. Tout comme CLIP, il utilise un système de double encodage et une stratégie de SSL contrasté à l'étape du pré-entrainement. Néanmoins, comme il est précisé dans l'article, l'étape de traitement couteux des donnée s pour CLIP est l'une de ses limites

Pour y remédier, l'équipe de recherche propose d'augmenter le nombre d'images dans le jeu de données (1,8 milliard paires au lieu de 400 millions pour CLIP). Ceci afin de démontrer que malgré la présence de bruit (dû à un pré-traitement moins conséquent) , augmenter la quantité de donnés suffit pour améliorer les performances du modèle et présenter de meilleurs résultats que des systèmes considérés comme SOTA, dont CLIP. Prouvant par la même occasion que leur modèle pré-entrainé est *« scalable »,* c'est-à-dire qu'il est capable de s'adapter à une quantité de données croissante.

Description du jeu de données :

Figure 7 : Exemple de paires image-texte issue de l'article [11]

Les images sont issues d'internet, et leurs textes correspondent à l'attribut *alt.* Toutefois, toutes les descriptions ne sont pas pertinentes (représenté en italique dans la figure 7), d'où la présence de bruit car elles ont été conservées.

Ils ont retiré les images inappropriés, également les doublons lors de la phase de tests afin d'éviter des répétitions. Au niveau du texte , ils ont retiré les *alt* partagées par plus de 100 images, ceux qu'ils considèrent trop court (moins de trois caractères) ou trop long (plus de vingt caractères)

Pré-entrainement :

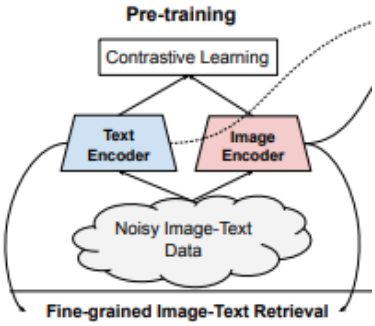

Figure 8 : Schéma des étapes de pré-entrainement d'ALIGN [11]

Tout comme CLIP, ALIGN utilise une approche SSL contrasté. Les encodeurs sont entrainés avec une fonction de perte d'apprentissage contrasté (un *softmax* normalisé) : au sein d'un même batch , on minimise la distance entre les paires texte-images positives tout en maximisant les paires négatives (i.e. pas correctement labelisés).

$$L_{i2t} = -\frac{1}{N}\sum_{i}^{N} log\frac{exp(x_i^T y_i/\sigma))}{\sum_{j=1}^{N} exp(x_i^T y_i/\sigma))}$$

$$L_{t2i} = -\frac{1}{N}\sum_{i}^{N} log\frac{exp(y_i^T x_i/\sigma))}{\sum_{j=1}^{N} exp(y_i^T x_i/\sigma))}$$

Figure 9 Fonction de perte utilisée pour ALIGN [11]

L'encodeur image est l'EfficientNet-L2 [12] , et l'encodeur texte est le BERT-Large [13]. Le modèle pré-entrainé présente 800 millions de paramètres. La durée de l'entrainement et le nombre de GPU nécessaire n'ont pas été mentionné dans l'article.

Evaluation :

Tout comme CLIP , ALIGN a procédé à une configuration ZSL pour évaluer son modèle. Sur la figure 10 nous avons les résultats dans le cas d'une tache recherche texte-image multiple *(multiple image-text retrieval)*

		Flickr30K (1K test set)					
		image → text			text → image		
		R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot	ImageBERT	70.7	90.2	94.0	54.3	79.6	87.5
	UNITER	83.6	95.7	97.7	68.7	89.2	93.9
	CLIP	88.0	98.7	99.4	68.7	90.6	95.2
	ALIGN	**88.6**	**98.7**	**99.7**	**75.7**	**93.8**	**96.8**

Figure 10 Résultats sur la tache de recherche image-texte sur les base de données Flicker30k et MS-COCO [11]

ALIGN a pu battre des modèles considérés comme SOTA sur certaines base de données, telle que Flicker30K.

Même constat pour le jeu de données d'ImageNet au niveau de la classification (toujours dans une configuration ZSL) :

Model	ImageNet	ImageNet-R	ImageNet-A	ImageNet-V2
CLIP	76.2	88.9	**77.2**	**70.1**
ALIGN	**76.4**	**92.2**	75.8	**70.1**

Figure 11 ALIGN présente de meilleurs résultats que CLIP sur les variantes d'ImageNet [11]

3.2 FLORENCE : vers une généralisation plus poussée.

Florence est un modèle fondation élaborée par Microsoft présenté en 2021[14]. Il constitue le modèle fondation le plus récent abordé dans cet état de l'art.

La motivation de l'équipe de recherche est de d'élaborer un modèle capable de s'adapter sur une variété de tâches qu'ils ont catégorisé en trois dimensions : espace *(space*), modalité *(modality)* et temps (*time)*.

Nous pouvons trouver une variété de taches , par exemple :

- Espace : de la classification d'image , à la segmentation sémantique
- Temp : du statique (image) à la dynamique (vidéo)
- Modalité : du simple RGB à la profondeur RGB

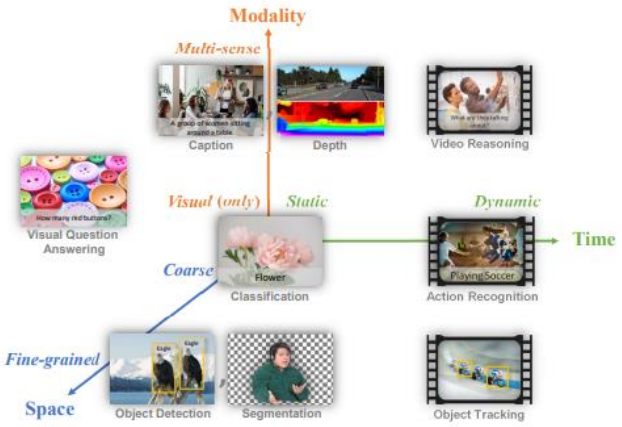

Figure 12 : Les tâches courantes de vision réparties en trois dimensions [14]

Ainsi, le modèle peut être facilement adapté pour différentes tâches relatives à la vision par ordinateur, mais plus particulièrement des taches demandant des traitements de données multimodales : la recherche d'image par le contenu, classification d'image, détection d'objet, reconnaissance d'action dans les vidéos, réponse à des questions visuelles e.g. Dans le cas de CLIP et ALIGN, leurs applications étaient restreintes aux taches relatives à l'association image-texte (statiques).

Description de la base de données utilisée pour le pré-entrainement :

Les paires images-textes sont également issues d'internet : la base de données initiale comportait 3 milliards de paires. Après une série de pré-traitement pour retirer les doublons et les contenus indésirables, elle a été réduite à 900 millions de paires image-texte : cette base de données est intitulée FLD-900m.

Architecture et pré-entrainement

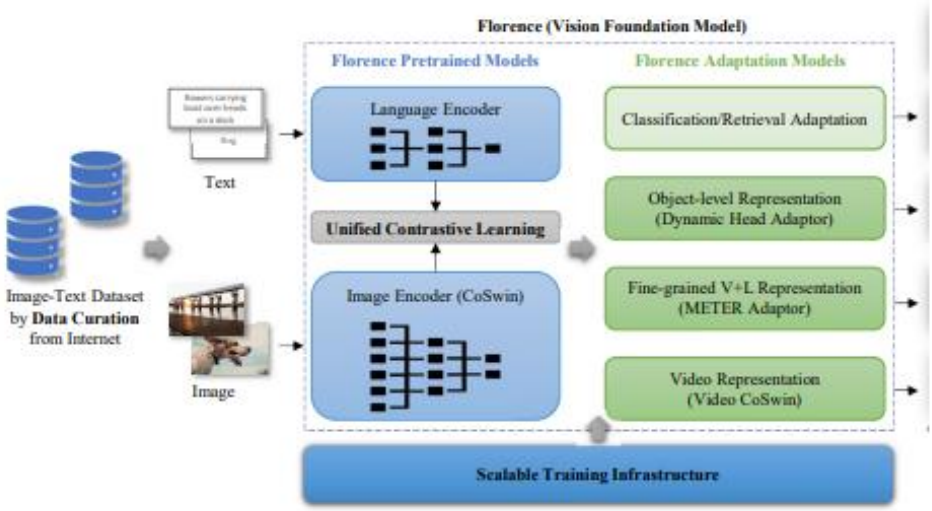

Figure 13 Architecture du modèle Florence [14]

Tout comme CLIP et ALIGN, l'équipe de recherche utilise un système double encodage . Au niveau du traitement des images ils utilisent un encodeur CoSwin, (un Swin transformer) [15]. Pour le texte, ils utilisent également un transformer (configuration proche de CLIP) composé de 12 couches. Le modèle présente 893 millions de paramètres, pour un temps d'entrainement de 10 jours sur 512 NVIDIA-A100 GPU.

Toutefois , à la différence de CLIP et ALIGN où l'entraiment s'est fait sur des paires images-textes, ici nous avons des triplets image-label-description. De plus, les triplets ne sont pas uniques , une image peut avoir plusieurs descriptions dans FLDM900. Le triplet est présenté de la manière suivante : Triplet(x,y,z), où :

- x – l'image
- t – langage de description
- y – langage de label

Ce modèle est également entrainé avec une méthode SSL contrasté où toutes les paires de tests d'images associées au même label « *y* » sont considérées comme des instances positives. Toutefois, nous pouvons considérer leur approche comme étant un SSL contrasté bidirectionnel car la fonction de perte (*Contrastive Loss*) contient deux éléments :

Avec L_{i2t} la fonction de perte pour une approche image vers texte et L_{t2i} pour approche image texte.

$$\mathcal{L} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i}.$$

$$\mathcal{L}_{i2t} = -\sum_{i\in\mathcal{B}} \frac{1}{|\mathcal{P}(i)|} \sum_{k\in\mathcal{P}(i)} \log \frac{\exp(\tau \boldsymbol{u}_i \boldsymbol{v}_k)}{\sum_{j\in\mathcal{B}} \exp(\tau \boldsymbol{u}_i \boldsymbol{v}_j)}, \quad (2)$$

where $k \in \mathcal{P}(i) = \{k | k \in \mathcal{B}, y_k = y_i\}$, and the supervised language-to-image contrastive loss

$$\mathcal{L}_{t2i} = -\sum_{j\in\mathcal{B}} \frac{1}{|\mathcal{Q}(j)|} \sum_{k\in\mathcal{Q}(j)} \log \frac{\exp(\tau \boldsymbol{u}_k \boldsymbol{v}_j)}{\sum_{i\in\mathcal{B}} \exp(\tau \boldsymbol{u}_i \boldsymbol{v}_j)}, \quad (3)$$

where $k \in \mathcal{Q}(j) = \{k | k \in \mathcal{B}, y_k = y_j\}$.

Figure 14 Fonction de perte du modèle de Florence [14]

Ils combinent alors deux tâches : la mise en correspondance des images avec les étiquettes et l'attribution d'une description à une étiquette unique.

Evaluation :

L'équipe de recherche a procédé à plusieurs méthodes d'évaluation et de comparaisons, notamment avec une approche ZSL. Leur modèle a pu battre des modèles considérées comme état-de-l 'art sur certaines tâches et base de données .

Method		Flickr30K (1K test set)				MSCOCO (5K test set)			
		Image → Text		Text → Image		Image → Text		Text → Image	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
	ImageBERT (Qi et al., 2020)	70.7	90.2	54.3	79.6	44.0	71.2	32.3	59.0
	UNITER (Chen et al., 2020d)	83.6	95.7	68.7	89.2	-	-	-	-

		Aquarium	BCCD	Chess Pieces	Mask Wearing	Oxford Pets	Packages	Pistols	PKLot	Pothole	Thermal	Wildfire Smoke
	Images	638	364	292	149	3680	26	2986	12416	665	203	737
	Categories	7	3	12	2	37	1	1	2	1	2	1
Zero-shot	ZSD	16.0	1.2	0.1	0.6	0.3	58.3	31.5	0.2	2.4	37.4	0.002
	Florence	43.1	15.3	13.4	15.0	68.9	79.6	41.4	31.4	53.3	46.9	48.7

Figure 16 Comparaison pour la détection d'objet, Florence présente de meilleurs résultats [14]

La principale contribution de Florence se trouve dans la détection d'objet, notamment au niveau de la généralisation de cette tâche avec le ZSL . Le modèle a pu battre des modèles considérés précédemment comme état-de-l'art : DyHead (Dai et al., 2021) [16] et ZSD (Bansal et al., 2018) [17]

Tout comme CLIP , Florence présente une meilleure généralisation sur des images naturelles (68.9% sur l'Oxford Pets contre 15% sur BCCD (base de données d'images biologiques)).

Par ailleurs sur des images plus complexes , comme Widfire Smoke, nous sommes passés à un taux de reconnaissance de 0.002% à 48.7% , ce qui est assez impressionnant pour une configuration ZSL

4. Un exemple d'application : la génération visuelle

Parmi les taches existantes dans le domaine du traitement multimodales images-textes nous pouvons citer les taches génératrices. Il existe des sous catégories :

- Réponse automatique à des questions visuelle, *Visual Question Answering (VQA)*: système de question-réponse sur une image
- Description de scène, *Visual Captioning*: générer une légende/description depuis une image
- Génération visuelle, *Visual Generation* : générer une représentation visuelle depuis une entrée de texte

La dernière catégorie a suscité énormément d'intérêt ces dernières années, notamment avec la réalisation d'open AI , DALL-E 2 [18]. Ce système est considéré comme le système d'intelligence artificielle faisant le plus preuve de créativité à ce jour. Il est capable de créer des représentations visuelles existantes et réalistes (cf. figure 17) mais aussi la représentation d'objets qui n'existent pas dans la nature (cf figure 18)

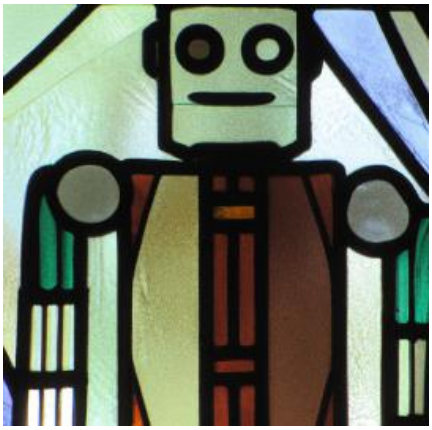

Figure 17 : image geéerée avec DALL-E 2 avec le texte : « a strained glass window depicting a robot »

Figure 18 : image generée avec DALL-E avec le texte « a tulipe made of white feathers"

L'architecture de DALL E est principalement composé de deux blocs important : CLIP que nous avons présenté plus haut , et l'utilisation de GLIDE[19] un modèle de diffusion qui fut une approche totalement nouvelle dans ce domaine.

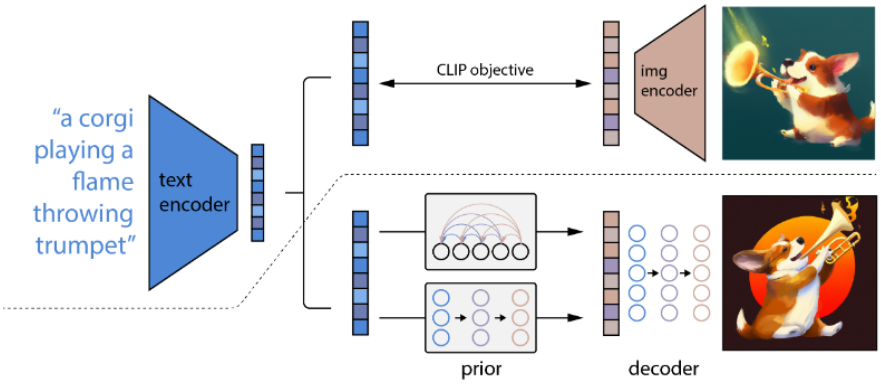

Figure 19 Architecture de DALL-E 2[20]

Les modèles de diffusion sont des modèles génératifs capables de synthétiser des images de haute qualité à partir d'une variable latente.

Dans cette architecture, CLIP fait office d'encodeur de texte. La description écrite en langage naturel est incorporée au sein du réseau qui servira d'entrée pour le modèle , ce sont les *prior :* chaque *prior* génère un mapping avec les incorporations d'images correspondantes.

Une fois le mapping terminé, un décodeur intervient pour générer l'image finale : c'est le modèle de diffusion GLIDE (*Guided Language to Image Diffusion for Generation and Editing).* Ce modèle prend en considération les informations textuelles.

Les modèles de diffusion sont des modèles génératifs basés sur des *Transformers*. Ils rajoutent graduellement du bruit (en général Gaussien) sur un type de donnée, en l'occurrence une image dans notre cas. Depuis cette image bruitée, le modèle doit reconstruire l'image originale : ils apprennent alors à générer des images. Un modèle entrainé sur un jeu de données de fleurs pourra générer des photos réalistes de fleurs. Toutefois , pour générer une espèce spécifique de fleur , cette information doit être précisée par une donnée textuelle. Le succès de GLIDE est dû à cet aspect , son entrainement a été renforcé grâce à l'incorporation des données textuelles, qui sont les *prior* décrits auparavant.

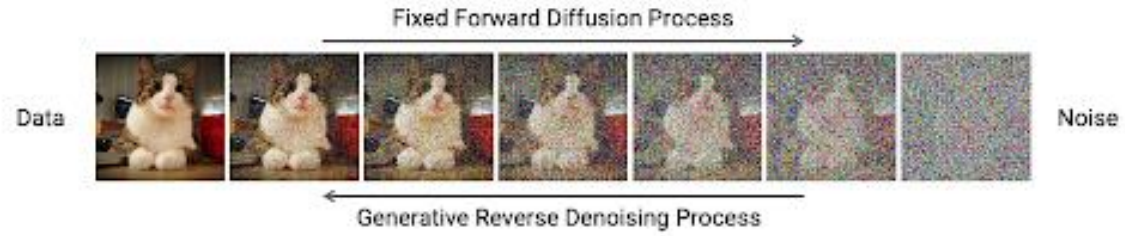

Figure 20 Démonstration des étapes du modèle de diffusion [21]

DALL-E présente tout de même des limites. Parmi celles listées par les chercheurs [18], celui de générer du texte cohérent

Figure 21 : Exemple d'image générée par DALL E 2 , « A signe that says Deep Learning » [18]

DALL E présente également des limites au niveau de la génération scènes complexes :

Nous pouvons remarquer sur cette figure 22 le manque de détails sur les affiches publicitaires.

Figure 5Figure 22 : Exemple d'image générée par DALL E 2 , « Times Square » [18]

5. Discussion

Bien que les modèles présentés dans l'état de l'art soient récents (avec le modèle le plus ancien CLIP qui a été présenté en janvier 2021), ils suscitent l'intérêt de nombreuses industries qui voient en ces systèmes un moyen de renforcer la capacité des systèmes à reproduire les capacités visuelles de l'homme .

Nous pouvons également relever le fait que ces modèles fondations soient élaborés par de grandes boites de technologie (Open AI, Microsoft ou en encore Google). La réalisation de ce genre de modèle demande des données importantes et un matériel ultra-performant et coûteux, ce qui n'est forcément à la portée de toute de la communauté de chercheurs dans ce domaine.

De plus, certaines types de données sont moins disponibles sur internet, comme les images médicales ou encore satellitaires : il est difficile de généraliser un modèle sur des images non naturelles. Florence présente toutefois des résultats prometteurs sur cet aspect.

Par ailleurs, certains de ces modèles semblent présenter des bais mathématiques pouvant porter préjudice à certaines catégories de personnes en raison de leur genre ou de leur appartenance ethnique, ce fut le cas de CLIP[5]. Les chercheurs ont mené une expérience au cours de laquelle CLIP a été chargé de classer 10 000 images issus de *FairFace,* une collection de plus de 100 000 photos montrant des personnes d'ethnicité blanche, noire, indienne, d'Asie de l'Est, d'Asie du Sud-Est, du Moyen-Orient et d'Amérique Latine. Dans le but de vérifier les biais du modèle qui pourraient affecter certains groupes démographiques, les auteurs ont ajouté « animal », « gorille », « chimpanzé », « orang-outan », « voleur », « criminel » et « personne suspecte », aux catégories existantes dans *FairFace.*

A la suite de ces expérimentations, ils ont fait le constat suivant : *« 4,9% des images étaient mal classées dans l'une des classes non humaines que nous avons utilisées dans nos sondages « animal », « gorille », « chimpanzé », « orang-outan »*. Parmi celles-ci, les images de personnes noires présentaient le taux de classification erroné le plus élevé (environ 14%) tandis que tous les autres groupes démographiques présentaient des taux de classification erroné inférieurs à 8 %.

Ils ont également mentionné le risque que ces biais restent présents dans ces modèles malgré des modifications apportées au système : « *leurs effets se manifestent à la fois de manière visible et invisible* »

6. Conclusion

Le traitement de données multimodales image-texte est un domaine relativement récent , le premier modèle fondation a été présenté en 2021. Ces modèles ont permis de réduire la frontière entre intelligence-machine et intelligence humaine en utilisant des méthodes du type SSL contrasté. Ces modèles ont également en point commun la volonté de réaliser des modèles de la généralisation en augmentant la quantité de données entrainée (ALIGN) ou encore redéfinir un espace de dimension pour traiter une plus grande variété de tâches (Florence). Nous avons pu voir que la déclinaison de ces modèles ont permis de résoudre des problèmes qui reproduisent des compétences humaines surprenante, telle que la créativité avec DALL-E.

Néanmoins, ces modèles présentent des limites qui doivent être pris en considération, notamment le risque de porter préjudice à certaines catégories de personnes en raison de leurs appartenance ethniques, âge, sexe e.g.

7. Références

1 : Figure issue de l'article de Derrick Mwiti *Transfer Learning Guide: A Practical Tutorial With Examples for Images and Text in Keras* , publié sur le site Neptune en Juin 2022

2 : Bommasani, Rishi; Hudson et al. (Aout 2021). *On the Opportunities and Risks of Foundation Models (Rapport)*. arXiv:2108.07258.

3 : Hangwei Qian, Tian Tian, Chunyan Miao, Nanyang Technological University Singapore *What Makes Good Contrastive Learning on Small-Scale Wearable-based Tasks?* ACM SIGKDD(2022). arXiv : 2202.05998

4 : Li Zhang, Tao Xiang, Shaogang Gong *Learning a Deep Embedding Model for Zero-Shot Learning,* pubié dans IEEE Consumer Electronics Magazine en 2017 , page 3010, 3019 ,arXiv : 1611.05088

5 : Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever *Learning Transferable Visual Models From Natural Language Supervision* 38ème conference sur le Machine Learning, Janvier 2021, arXiv 2103.00020

6 : Ang Li, Allan Jabri, Armand Joulin, Laurens van der Maaten, Université de Maryland, ,*Learning Visual N-Grams from Web Data* , ICCV(2017), p. 4183 – 4192

7 : Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, Laurens van der Maaten , « *Exploring the Limits of Weakly Supervised Pretraining»,* ECCV 2018, pages 181 – 201. arXiv : 1805.00932

8: Kaiming He, Xiangyu Zhang, Shaoqing Ren Jian Sun, Microsoft, *Deep Residual Learning for Image Recognition*, IEEE(2015) arXiv : 1512.03385

9 : Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* , arXiv : 2010.11929, ICLR (2021)

10 : Li Deng, *The MNIST Database of Handwritten Digit Images for Machine Learning Research (Best of the Web)* , IEEE (2012) p. 141-142

11 : Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, Tom Duerig 3 *Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision* , ICML(Mai 2021) arXiv : 2102.05918

12 : Mingxing Tan, Quoc V. Le, *EfficientNetV2: Smaller Models and Faster Training*, ICML(2021): p. 10096-10106 . arXiv : 2104.00298

13 : Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding",*NAACL-HLT(2018): 4171-*4186* . arXiv:1810.04805

14 : Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, Pengchuan Zhang *Florence: A New Foundation Model for Computer Vision*, CoRR(Novembre 2021) arXiv : 2111.11432

15 : Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, *"Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*" CoRR(2021), arXiv : 2103.14030

16 : Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., and Zhang, *L. Dynamic head: Unifying object detection heads with attentions.* CVPR (2021), p. 7373–7382 . arXiv : 2106.08322

17 : Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, Ajay Divakaran , *Zero-Shot Object Detectio*n, ECCV(2018) p384–400, arXiv : 1804.04340

18 : Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen: *Hierarchical Text-Conditional Image Generation with CLIP Latents* CoRR(2022) arXiv : 2204.06125

19 : Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, Mark Chen, *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models* CoRR(2021) arXiv : 2112.10741

20 : Illustration issue du site NimbleBox.ai *DALL E 2: AI That Can Render Masterpieces from Text!*

21 : Illustration issue du site NVIDIA *Improving Diffusion Models as an Alternative To GANs,* Arash Vahdat et Karsten Kreis