An ecologically motivated approach to prompt engineering for zero-shot image classification with CLIP

Shawn Manuel, Abbas Guennoun

Introduction

Depuis la création du perceptron, les parallèles entre les réseaux de neurones humains et artificiels fournissent des informations fructueuses sur l'organisation structurelle et fonctionnelle de chacunes. De nos jours, plusieurs réseaux neuronaux profonds (RNP) présentent des caractéristiques réminiscentes de l'organisation hiérarchique du système visuel (Cichy et al., 2016). On constate chez ces modèles des performances qui atteignent, et dépassent parfois, le niveau des humains dans une variété de tâches visuelles (Yu et al., 2022). En effet, l'apprentissage statistique dans le contexte de la vision par ordinateur a rapidement révélé que la détection de bordures est une stratégie de bas niveau accessible, miroitant l'activité des neurones sensibles à l'orientation du cortex visuel primaire (V1) (Hubel & Wiesel, 1977; LeCun et al., 2015) . Mais qu'en est-il des représentations plus complexes que la simple orientation des bordures?

Chez l'Humain, le traitement hiérarchique de l'information dans le cortex visuel signifie que de V1 à V5, il y a une complexification graduelle des représentations à l'aide de celles des aires précédentes. En d'autres termes, les bordures détectées par V1 servent à « construire » des représentations d'objets plus sophistiquées, incluant notamment la forme et la taille, dans les aires suivantes. Une organisation hiérarchique analogue est présente chez plusieurs RNPs modernes, un point auquel nous reviendrons. Plus globalement, les humains bénéficient de l'apport d'autres régions cérébrales, comme le cortex préfrontal ventromédian et l'hippocampe, qui intègrent les connaissances préalables, pour créer des liens entre multiples représentations d'un phénomène ou d'un objet (Gilboa & Marlatte, 2017; van Kesteren et al., 2012). Les représentations résultantes sont maintenant étudiées sous le nom de « schéma », compris comme étant des structures d'informations possédant quatre caractéristiques déterminantes (Ghosh & Gilboa, 2014):

I. Une structure associative en réseau, c'est-à-dire des unités ayant des relations entre elles;

II. Une constitution basée sur plusieurs expositions;

III. Des unités « floues », c'est-à-dire qu'elles ne correspondent pas uniquement à un seul type de représentation;

IV. Être flexibles, c'est-à-dire qu'ils peuvent à la fois assimiler de nouvelles informations sans être modifiés et accommoder de telles informations en changeant leur structure.

Depuis l'avènement de réseaux multimodaux, comme CLIP, développé par OpenAI, il est possible de franchir une étape importante dans l'applicabilité des RNPs en liant des images à du texte (Radford et al., 2021). D'intérêt particulier est le fait que CLIP contient des représentations de 'haut-niveau', construites à l'aide de représentations de plus bas niveau, qui

adhèrent aux quatre postulats d'un schéma (voir Figure 1 et 2) (Goh et al., 2021). Toutefois, le développement et la correspondance de ces ensembles de représentations ou de schémas, c'est-à-dire celles des humains et celles des réseaux comme CLIP, est encore mal compris.

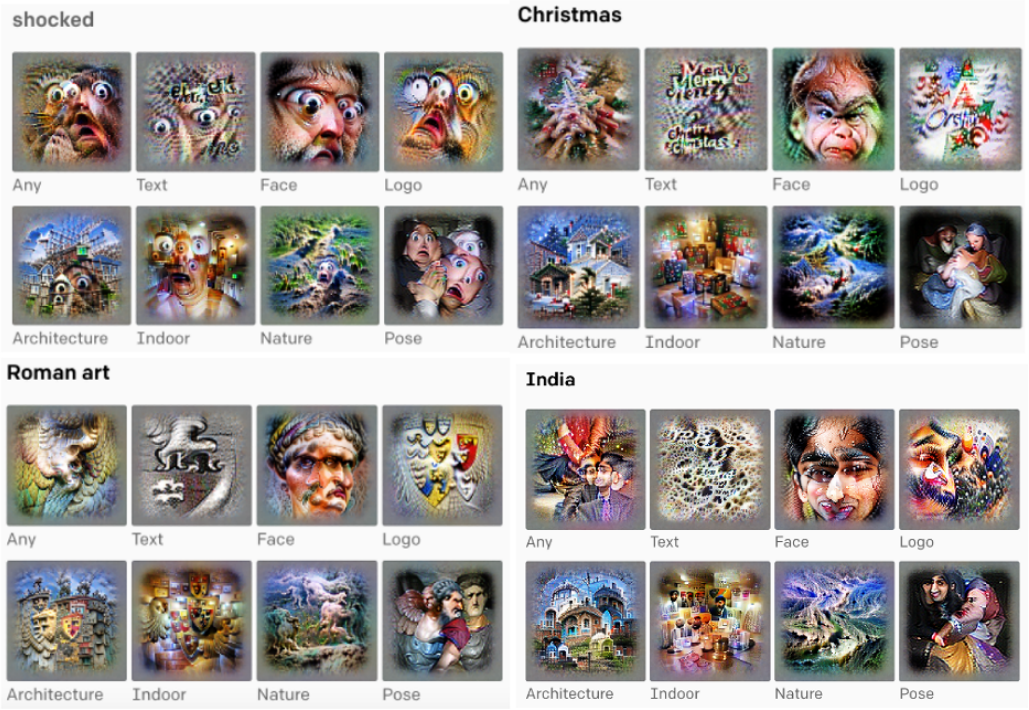

Figure 1. Visualisations de quatre neurones de CLIP répondant aux schémas du choc, de Noël, de l'art romain et de l'Inde. Ces images sont disponibles sur le blog de OpenAI et peuvent être analysées plus en détail à l'aide de leur « Microscope » en ligne. Chaque image correspond à la visualisation des images qui contribuent le plus à l'activité d'un neurone donnée.

Un défi central pour l'étude des schémas concerne le développement d'outils de mesure appropriés. Dans le cas des humains, cela se traduit par la création de questionnaires; tandis que dans le cas des RNPs comme CLIP, on parle plutôt de *prompt engineering*. Le but de ce projet est d'employer une banque de stimuli visuels accompagnée de variables psycholinguistiques pour comparer les schémas humains aux représentations de haut niveau développées par CLIP. L'utilisation de telles variables limite les décisions arbitraires et permet d'informer le *prompt engineering* avec des données écologiquement valides. Spécifiquement, nous voulons savoir si un réseau multimodal, comme CLIP, peut être lié à des données humaines pour aller plus loin que la simple reconnaissance d'objet et interroger des schémas complexes, notamment le dégoût, de manière fiable.

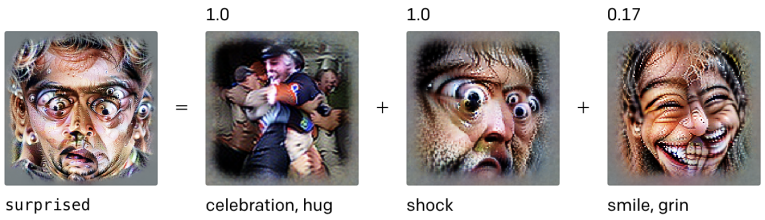

Figure 2. Visualisations des unités contribuant aux représentations d'images montrant la surprise (pris dans https://openai.com/blog/multimodal-neurons/)

Méthode

Pour ce faire, nous avons programmé un *pipeline* d'analyse avec Python (v3.9.12), implémenté sur deux machines : un MacBook Air (M2) avec mac OS Monterey (v12.5) ainsi qu'un Alienware Aurora Ryzen Edition R14 avec Ubuntu 22.04.1 LTS. Le code est disponible sur GitHub (https://github.com/AbbasGuennoun/MMD6020-A22-projet).

Réseau de neurones profond sélectionné

CLIP est un réseau de neurones artificiels multimodal dont l'apprentissage est basé sur les concepts du *zero-shot learning* et du *natural language supervision* (Radford et al., 2021; Zhang et al., 2017). Chaque composante des paires image-texte est traitée par un encodeur spécifique (encodeur de texte et encodeur d'image), pour ensuite créer un espace latent à haute dimension où il existe une forte similarité entre l'image et son texte correspondant (voir Figure 3). L'ensemble d'entraînement est constitué de 400 millions de paires image-texte, puisées de différentes sources disponibles sur Internet, comme Flickr (Radford et al., 2021).

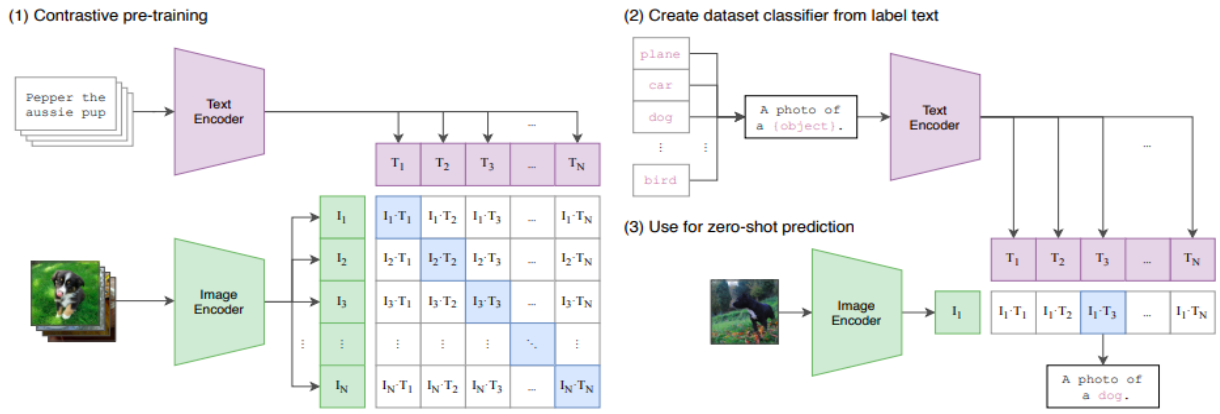

Figure 3. Architecture de CLIP (pris dans https://openai.com/blog/clip/).

Stimuli visuels sélectionnés

La base de données DIRTI (Disgust-Related-Images) consiste en un ensemble d'images standardisé et validé qui permet l'étude du dégoût en psychologie (Haberkamp et al., 2017). Cette base de données rassemble 240 images déclenchant le dégoût et 60 images neutres organisées en six catégories : nourriture, animaux, produits corporels, blessures/infections, mort et hygiène (voir Figure 4). Elles ont ensuite été notées selon une échelle de un jusqu'à neuf mesurant le peur, le dégoût ainsi que la valence et l'intensité de l'émotion éveillée (où *1 = aucune* et *9 = très fortement*). Plusieurs de ces images ont été prises par les auteurs, d'autres en ligne sur des sites comme Flickr.

Pipeline d'analyse

1. *Embedding 'baseline' avec les 6 catégories d'images DIRTI*

Pour un survol du pipeline d'analyse, voir la figure 5. D'abord, nous avons illustré les capacités de classification *zero-shot* de CLIP en lui fournissant les 300 images DIRTI accompagnées de 6 étiquettes textuelles correspondant aux catégories d'images. Les étiquettes ont été codées sous la forme suivante : « *This is a photo of …* », suivi de chaque catégorie. Après avoir encodé les images et les étiquettes, 512 caractéristiques ont été extraites par image. Une réduction de dimensionnalité *UMAP (Uniform Manifold Approximation and Projection)* a été appliquée pour projeter les images dans un espace latent à deux dimensions et faciliter la visualisation de leur *clustering* (voir Figure 6). Puis, à l'aide des prédictions d'étiquettes de CLIP, nous avons calculé ses performances prédictives (accuracy, precision, recall, F1 score). En parallèle, nous avons entraîné un modèle de classification K plus-proches-voisins (*k nearest neighbors, KNN;* k=6) sur les caractéristiques d'images extraites avec une validation croisée à 10 plis pour déterminer si les six catégories d'images DIRTI sont mieux reconnues par CLIP et avoir une base de comparaison.

Figure 4. Exemples d'images DIRTI pour chaque catégorie, incluant neutre, selon un spectre de dégoût (pris dans Haberkamp et al., 2017).

2. *Embedding psycholinguistique du dégoût*

Pour interroger le schéma du dégoût, il a été nécessaire d'ajouter plus d'informations à chaque étiquette. Ainsi, nous avons modifié les étiquettes de la façon suivante: « *This is a **disgusting** photo of …* », suivi de chaque catégorie. Pour déterminer quelles images étaient « dégoûtantes », nous avons choisi celles ayant un score de dégoût supérieur ou égal à 5. Nous avons choisi ce seuil, car il correspond à un écart-type au-dessus du score de dégoût moyen pour l'ensemble des images. Après avoir adapté les étiquettes textuelles fournies à CLIP et les étiquettes correspondant aux images dégoûtantes, nous les avons encodés à nouveau pour avoir de nouvelles prédictions et avons entraîné un modèle KNN (k=12) comme à l'étape précédente.

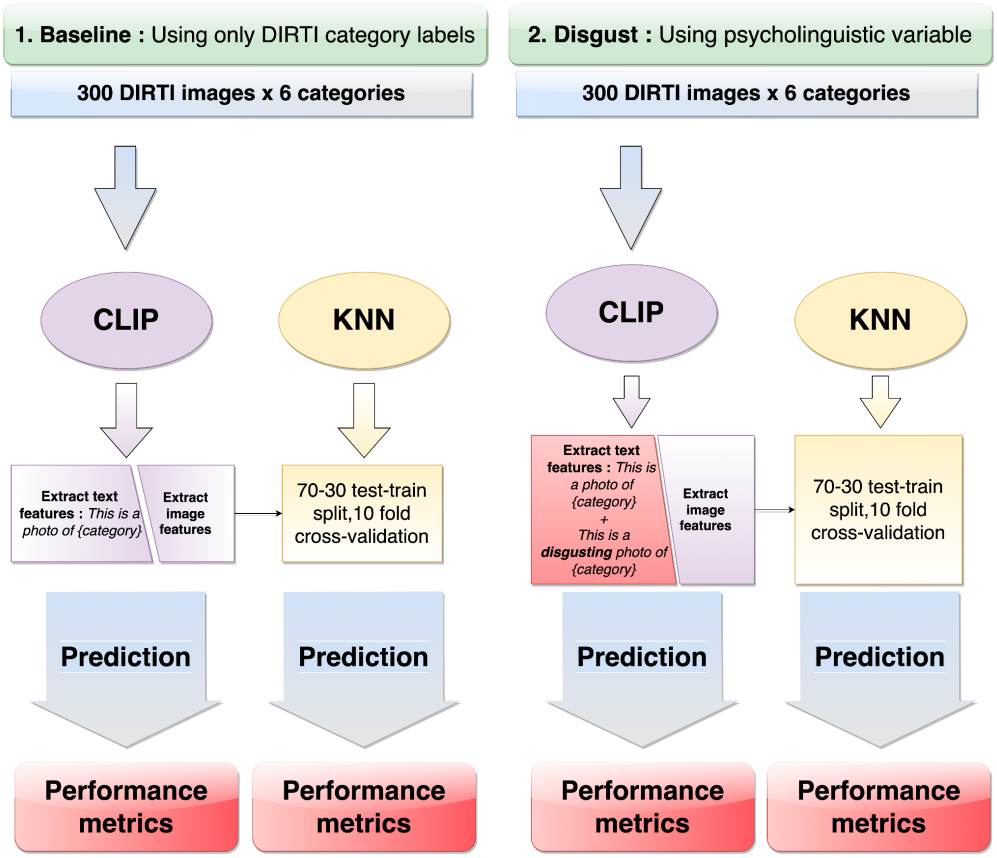

Figure 5. Schéma résumant le pipeline utilisé.

Dans les modèles baseline (1), CLIP est utilisé pour encoder les images ainsi que le texte. Ce dernier est ensuite testé sur la base de données DIRTI (modèle pré-entraîné). Un modèle KNN servant de base de comparaison a été entraîné et testé parallèlement sur les caractéristiques d'images extraites. Dans les modèles de la deuxième phase (2), une variable psycholinguistique a été ajouté pour le modèle CLIP et le KNN afin d'observer le comportement du réseau de neurones profonds lorsqu'on ajoute une telle variable. Les mêmes étapes de preprocessing (encodage avec CLIP) ont été effectués

Résultats

1. *Embedding 'baseline' avec les 6 catégories d'images DIRTI*

La figure 6 montre la projection 2D des caractéristiques d'images. On voit que la division en 6 catégories n'est pas bien représentée par CLIP. Les animaux (rouge) et les blessures/infections (mauve) sont bien séparés, tandis que la nourriture, les produits corporels, la mort et l'hygiène se superposent.

Le modèle CLIP a été testé sans ajout de la variable psycholinguistique et un modèle KNN a été entraîné pour servir comme comparateur. Le modèle KNN a obtenu la meilleure exactitude (accuracy) à 0,76, tandis que le modèle CLIP a obtenu une exactitude à 0,55. Les moyennes "macro" (chaque classe est traité de façon égale) ont été rapportés pour les deux modèles concernant les autres métriques (tableau 1) : CLIP a obtenu respectivement une précision à 0,55, un rappel à 0,50 et un F1-score à 0,49, contre le KNN qui a obtenu une

précision à 0,75, un rappel à 0,74 et un F1-score à 0,74. Le modèle KNN baseline a obtenu de meilleures performances que CLIP.

Les matrices de confusion (figure 7) montrent que le modèle KNN a eu de mauvaises prédictions concernant les classes 2 et 5, alors que le modèle CLIP a eu de mauvaises prédictions concernant les classes 2 et 4. Puisque CLIP est un modèle pré-entraîné, il existe plus d'exemples sur la matrice de confusion pour ce modèle, car il a été testé sur la totalité de la base de données. En revanche, le KNN a été entraîné sur un ensemble d'entraînement (70% des données) et testé sur un ensemble de test (30% des données).

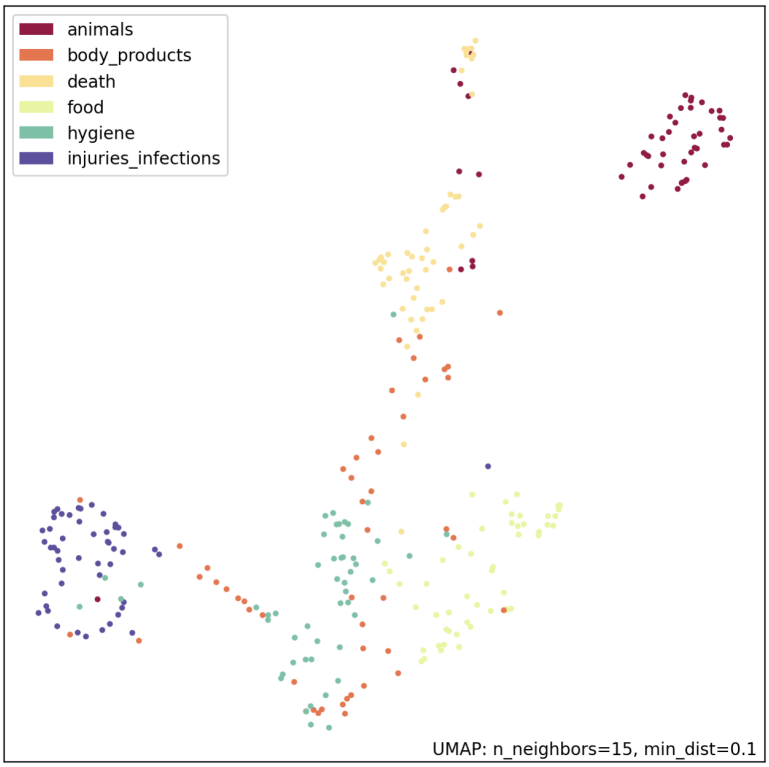

Figure 6. Projection 2D des caractéristiques d'images extraites par CLIP

Baseline		
Modèle	**KNN**	**CLIP**
Accuracy	0,76	0,55
Precision (macro avg)	0,75	0,55
Recall (macro avg)	0,74	0,50
F1-score (macro avg)	0,74	0,49

Tableau 1. Métriques de performances
Comparaison entre les performances de CLIP et KNN, entraînés sur les données, sans l'utilisation de la variable psycholinguistique

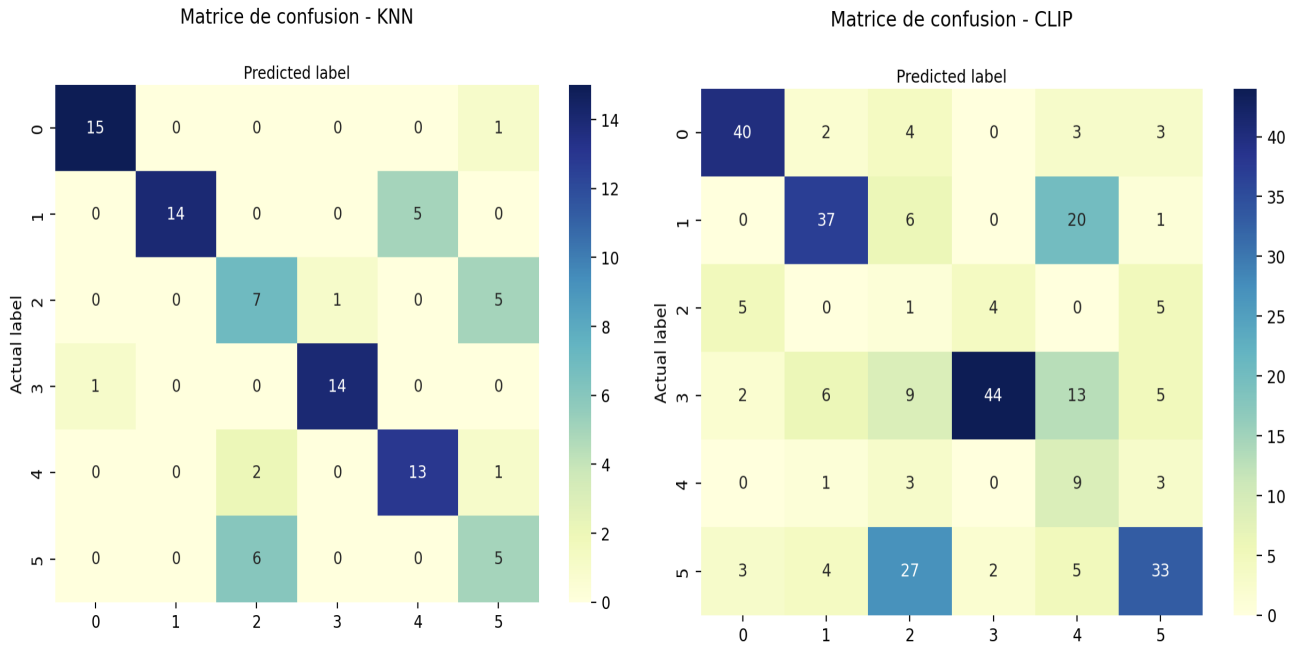

Figure 7. Matrices de confusion des modèles Baseline : KNN et CLIP

2. *Embedding psycholinguistique du dégoût*

Lors de la deuxième phase, six catégories ont été ajoutées, représentant des variables psycholinguistiques. Le même schéma que les baselines a été utilisé avec le test de CLIP, ainsi que l'entraînement d'un modèle KNN pour obtenir une base de comparaison. L'exactitude de CLIP est passée à 0,17 contre 0;69 pour le KNN. Concernant les autres métriques, CLIP a respectivement obtenu une précision à 0,23, un rappel à 0,17 et un F1-score à 0,13 (moyenne "macro"), contre une précision de 0,62, un rappel à 0,57, et un F1-score à 0,57 pour le modèle KNN. Les performances ont drastiquement baissé pour le modèle CLIP (Tableau 2). Les performances du KNN baissent mais restent très pauvres, et très proches de l'aléas (*à noter que deux catégories n'ont pas été générées dans le tableau de confusion pour le KNN, suggérant une erreur au niveau du code).*

Modèle entraîné en ajoutant une variable psycholinguistique		
Modèle	**KNN**	**CLIP**
Accuracy	0,69	0,17
Precision (macro avg)	0,62	0,23
Recall (macro avg)	0,57	0,17
F1-score (macro avg)	0,57	0,13

Tableau 2. Métriques de performances

Comparaison entre les performances de CLIP et KNN écologiquement motivé.

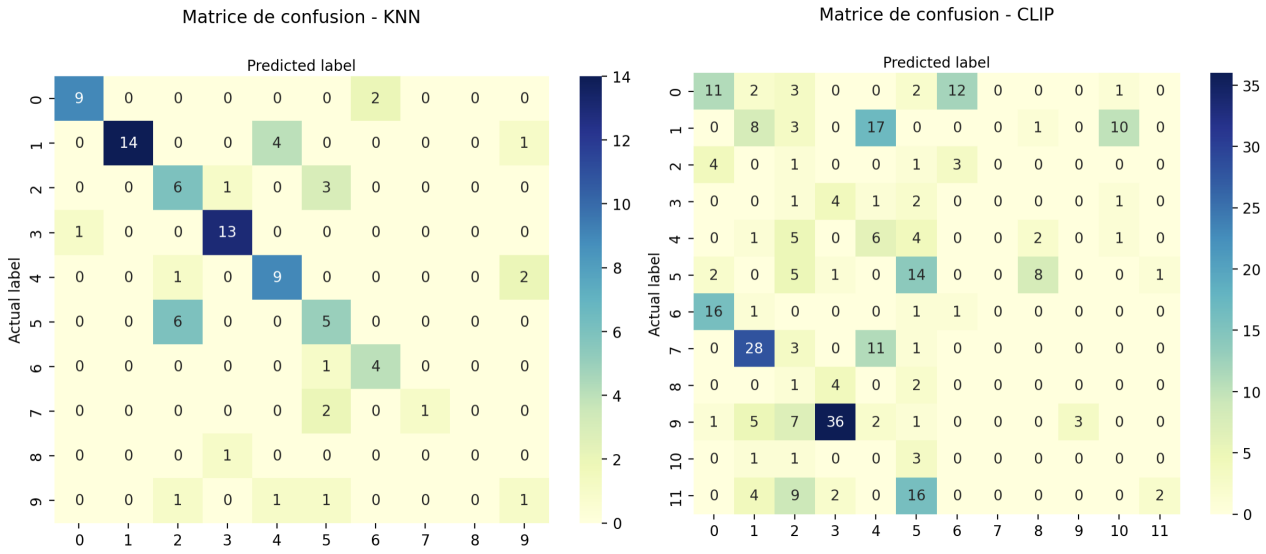

Figure 8. Matrices de confusion des modèles écologiquement motivés : KNN et CLIP

Discussion

L'objectif de ce projet a été d'employer une banque de stimuli visuels accompagnée de variables psycholinguistiques pour comparer les schémas humains aux représentations de haut niveau développées par le réseau multimodal CLIP. Dans un premier temps, nous avons utilisé CLIP pour classifier les images DIRTI selon leur catégorie respective, atteignant une exactitude de 0,55; tandis qu'un modèle KNN a obtenu une exactitude de 0,76. Dans un second temps, nous avons tenté d'améliorer les capacités prédictives de CLIP à l'aide de données psycholinguistiques sur le dégoût. Comme précédemment, nous avons entraîné un modèle KNN (k=12) comme comparateur. CLIP a atteint une exactitude de 0,17, tandis que le KPPV a atteint 0,69, soit un peu moins qu'au baseline. Dans les deux cas, un simple KNN a été plus performant que CLIP.

Comment se fait-il que les performances prédictives d'un simple KNN surpassent celles de CLIP? Pour les modèles « baseline », la faible performance de CLIP témoigne du fait que les catégories d'images décidées par les auteurs de DIRTI sont mal représentées par CLIP. Ou, à tout le moins, que le seul *prompt* sélectionné pour chaque catégorie (ex. *This is a photo of food*) est insuffisant pour permettre une séparation optimale. D'autres projets ont déjà exploré la possibilité d'utiliser plusieurs variantes de *prompt* pour chaque catégorie dans le but d'améliorer la classification, soit en variant la description (ex. *This is a painting of, This is a low quality picture of, etc.*) ou la catégorie en soi tout en restant dans le même champ lexical (ex. catégorie cible: camion; variante d'étiquette: *This is a photo of a car*) (Ben-Shaul, 2022). À la lumière de ses performances médiocres au baseline, il n'est guère surprenant de constater que CLIP ne performe pas mieux dans la condition écologiquement motivée. La mauvaise performance du KNN dans ce contexte témoigne une fois de plus du manque de représentation des catégories choisies par les auteurs.

Les limites de cette étude incluent notamment des erreurs au niveau du pipeline d'analyse pour la section sur le dégoût. En effet, deux catégories semblent avoir été omises de la classification et du calcul des performances prédictives, ce qui brouillent l'interprétation des résultats de cette section. La possibilité que certaines images de la base de données DIRTI aient été incluses dans les images d'entraînement de CLIP constitue une autre limite. Par la suite, il se peut que la base de données DIRTI elle-même ne soit pas un choix optimal pour interroger des schémas complexes. Des pistes futures seraient d'utiliser une autre base de données ou d'utiliser des participants humains pour fournir les descriptions à chaque image. Cette dernière solution permettrait d'éviter deux choix arbitraires supplémentaires: 1) comment formuler les descriptions et 2) comment choisir le seuil d'inclusion d'une image dans une catégorie donnée.

Conclusion

Malgré ses limites, cette étude pilote est la première à explorer l'existence de schémas complexes dans un réseau multimodal, ainsi que leur correspondance avec des schémas humains. L'importance de cette recherche pour la médecine computationnelle réside dans le fait qu'elle pourrait être adaptée pour étudier des schémas chez des populations psychiatriques et ainsi créer un lien entre les méthodes d'analyse objectives et projectives. En surplus, cette approche est aisément implémentable à grande échelle et à distance, si nécessaire. Finalement, bien que CLIP n'ait pu atteindre de meilleures performances lorsqu'augmenté par les données DIRTI, nos résultats suggèrent qu'une exploration plus approfondie est nécessaire pour bien cerner le potentiel des réseaux multimodaux.

Bibliographie :

Ben-Shaul, I. (2022, February 28). Having fun with CLIP features—Part I. *MLearning.Ai*. https://medium.com/mlearning-ai/having-fun-with-clip-features-part-i-29dff92bbbcd

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), Article 1. https://doi.org/10.1038/srep27755

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia*, *53*, 104–114. https://doi.org/10.1016/j.neuropsychologia.2013.11.010

Gilboa, A., & Marlatte, H. (2017). Neurobiology of Schemas and Schema-Mediated Memory. *Trends in Cognitive Sciences*, *21*(8), 618–631. https://doi.org/10.1016/j.tics.2017.04.013

Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., & Olah, C. (2021). Multimodal Neurons in Artificial Neural Networks. *Distill*, *6*(3). https://doi.org/10.23915/distill.00030

Haberkamp, A., Glombiewski, J. A., Schmidt, F., & Barke, A. (2017). The DIsgust-RelaTed-Images (DIRTI) database: Validation of a novel standardized set of disgust pictures. *Behaviour Research and Therapy*, *89*, 86–94. https://doi.org/10.1016/j.brat.2016.11.010

Hubel, D. H., & Wiesel, T. N. (1977). Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *198*(1130), 1–59. https://doi.org/10.1098/rspb.1977.0085

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), Article 7553. https://doi.org/10.1038/nature14539

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A.,

Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, *35*(4), 211–219. https://doi.org/10.1016/j.tins.2012.02.001

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., & Wu, Y. (2022). *CoCa: Contrastive Captioners are Image-Text Foundation Models* (arXiv:2205.01917; Version 2). arXiv. http://arxiv.org/abs/2205.01917

Zhang, L., Xiang, T., & Gong, S. (2017). Learning a Deep Embedding Model for Zero-Shot Learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3010–3019. https://doi.org/10.1109/CVPR.2017.321

Annexes :

	precision	recall	f1-score	support
0	0.80	0.77	0.78	52
1	0.74	0.58	0.65	64
2	0.02	0.07	0.03	15
3	0.88	0.56	0.68	79
4	0.18	0.56	0.27	16
5	0.66	0.45	0.53	74
accuracy			0.55	300
macro avg	0.55	0.50	0.49	300
weighted avg	0.70	0.55	0.60	300

(a)

	precision	recall	f1-score	support
0	0.94	0.94	0.94	16
1	1.00	0.79	0.88	19
2	0.50	0.38	0.43	13
3	0.93	0.93	0.93	15
4	0.78	0.88	0.82	16
5	0.44	0.64	0.52	11
accuracy			0.78	90
macro avg	0.76	0.76	0.76	90
weighted avg	0.80	0.78	0.78	90

(b)

Tableau 3. Métriques de performances détaillées (par classe) des deux modèles baseline. (a : CLIP, b : KNN)

	precision	recall	f1-score	support
0	0.32	0.35	0.34	31
1	0.16	0.21	0.18	39
2	0.03	0.11	0.04	9
3	0.09	0.44	0.14	9
4	0.16	0.32	0.21	19
5	0.30	0.45	0.36	31
6	0.06	0.05	0.06	19
7	0.00	0.00	0.00	43
8	0.00	0.00	0.00	7
9	1.00	0.05	0.10	55
10	0.00	0.00	0.00	5
11	0.67	0.06	0.11	33
accuracy			0.17	300
macro avg	0.23	0.17	0.13	300
weighted avg	0.36	0.17	0.15	300

	precision	recall	f1-score	support
0	0.90	0.82	0.86	11
1	1.00	0.74	0.85	19
2	0.43	0.60	0.50	10
3	0.87	0.93	0.90	14
4	0.64	0.75	0.69	12
5	0.42	0.45	0.43	11
6	0.67	0.80	0.73	5
8	1.00	0.33	0.50	3
9	0.00	0.00	0.00	1
10	0.25	0.25	0.25	4
accuracy			0.69	90
macro avg	0.62	0.57	0.57	90
weighted avg	0.72	0.69	0.69	90

Tableau 4. Métriques de performances détaillées (par classe) des modèles écologiquement motivés. (a : CLIP, b : KNN)

Liste des figures :

Liste des tableaux :