# Appariement d'images d'oeuvres d'Art avec descriptions textuelles variées

Raphaëlle Karine Lemaire, Alexis Lechervy, Youssef Chahir

# Appariement d'images d'oeuvres d'Art avec descriptions textuelles variées

Raphaelle Lemaire[1] Alexis Lechervy[1] Youssef Chahir[1]

[1] Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000, Caen, France
raphaelle.lemaire@unicaen.fr alexis.lechervy@unicaen.fr youssef.chahir@unicaen.fr

## Résumé

*L'appariement texte-image se concentre généralement sur des descriptions textuelles [2]. Cependant, ces méthodes ne prennent pas en compte les scénarios complexes où le texte contient des informations non visuelles. Dans cette étude, nous comparons les performances de CLIP [9] sur des textes visuels, contextuels et mixtes. Nous montrons que le type de texte influence les performances des tâches de recherche multimodales. Un texte visuel est plus efficace pour la recherche d'images, tandis qu'un texte mixte est plus facile à retrouver dans le cas d'une recherche de texte.*

## Mots Clef

Appariement de modalités, Recherche multimodale, Apprentissage profond

## Abstract

*Text-image matching generally focuses on textual descriptions [2]. These methods do not take into account complex scenarios where the text contains non-visual information. In this study, we compare the performance of CLIP [9] on visual, contextual, and mixed texts. We show that the type of text influences the performance of multimodal search tasks. Visual text yields better results in image retrieval, while mixed text is more effective for text retrieval.*

## Keywords

Modality matching, Multimodal retrieval, Deep learning

# 1 Introduction

La compréhension des performances des modèles d'apprentissage profonds est essentielle pour leur application dans divers domaines tels que le médical, l'artistique et le social. Aligner les modalités visuelles et textuelles est une stratégie pour améliorer cette compréhension [1]. Dans cet article, nous nous intéressons à la recherche d'images à partir de textes et vice versa. Actuellement, cette tâche est principalement réalisée sur des textes courts et descriptifs [2], ce qui ne correspond pas toujours à la réalité. L'Art met en évidence les limites de ces approches avec des textes décrivant l'information visuelle et culturelle [8]. Notre étude propose d'évaluer les capacités d'un modèle d'appariement multimodal, CLIP [9], sur la recherche d'images et de textes en utilisant la base de données Artpedia [7] dont les textes sont classés en deux catégories : contextuelle et visuelle. Nous proposons une nouvelle base de texte construite à partir d'Artpedia composée de textes de nature variée. Nous ferons une analyse détaillée des performances en fonction de la nature des textes en section 3.

# 2 Méthode

CLIP [9] (Contrastive Language-Image Pretraining), un modèle développé par OpenAI, est entraîné de manière contrastive. L'architecture que nous utilisons, disponible sur la bibliothèque LAVIS [11], est basée sur un modèle VIT-L-14-336 [14] pré-entraîné sur ImageNet [13] pour la partie Image et BERT[15] pour la partie Texte qui sont ensuite alignés dans un espace commun. Cette approche d'entraînement permet à CLIP de développer des représentations multimodales cohérentes et discriminatoires. Durant nos expériences, nous avons comparé des modèles avec ou sans fine-tuning. Lorsque les modèles ont été ré-entraînés, nous avons optimisé la tâche cible à l'aide d'un algorithme de descente de gradient de type ADAM [16] sur 6 époques.

# 3 Expérimentations

## 3.1 Présentation du jeu de données

| | Train | Val | Test | Total |
|---|---|---|---|---|
| **img** | 2190 (77%) | 327 (11%) | 336 (12%) | 2853 |
| **vsl** | 6917 (77%) | 996 (11%) | 1022 (12%) | 8935 |
| **ctx** | 14382 (78%) | 2058 (11%) | 2069 (11%) | 18509 |

TABLE 1 – Descriptif de la base de donnée Artpedia

Artpedia est une base de données composée de tableaux d'art (img) associés à des phrases visuelles (vsl) et contextuelles (ctx). La part de phrases visuelles est de 32% (train : 32%, val : 33% et test : 33%). Certains liens de la base originelle n'étant plus inaccessibles, nous avons dû réduire le nombre d'images disponibles (cf Tabl.1). Nous proposons trois variantes de la base, composées soit des phrases visuelles (vsl), soit des phrases contextuelles (ctx), soit d'un mixte des deux (mix). Nous créons également une autre base reprenant les images de Artpedia et composée de dix nouveaux textes. Comme précédemment, nous proposons trois variantes de ces textes (vsl, ctx et mix). Notamment, les textes mixtes sont construits en garantissant la présence d'au moins une phrase de chaque. Les phrases sont issues

de la base Artpedia, dans la moitié des cas à l'identique, et dans l'autre moitié paraphrasée par BART-L [12].

### 3.2 Protocole expérimental

Nous étudions deux tâches, la recherche d'image à partir d'un texte et la recherche de textes à partir d'image. Pour évaluer ces tâches, nous mesurons la capacité du modèle à retrouver un élément pertinent parmi les 1, 5 et 10 premiers éléments renvoyés, en utilisant le rappel comme métrique (R@1,R@5 et R@10). Les écarts-types sont calculés sur 10 entraînements lorsque le modèle est ré-entraîné. Ils sont calculés à partir des performances sur les ensembles de test et de validation dans les autres cas.

### 3.3 Résultats

Les performances en recherche de texte (cf Tab. 2) sont meilleures qu'en recherche d'image (cf Tab. 4). Cette supériorité s'explique par la différence de difficulté entre la tâche de recherche de texte et celle de recherche d'image. Dans la première, il s'agit de retrouver au moins un texte parmi plusieurs associés à l'image, tandis que dans la seconde, l'objectif est de trouver l'unique image associée au texte.

| | | | **Recherche de texte** | | |
|---|---|---|---|---|---|
| | | | **R@1** | **R@5** | **R@10** |
| sans finetuning | phrases | **vsl** | 63.30±3.93 | 85.02±0.56 | 90.21±1.23 |
| | | **ctx** | 64.53±1.08 | 82.26±3.70 | 85.93±4.05 |
| | | **mix** | **74.70**±0.36 | **93.75**±0.028 | **97.02**±1.13 |
| | textes | **vsl** | 67.56±1.06 | 79.76±0.61 | 84.52±0.12 |
| | | **ctx** | 67.56±3.88 | 81.85±2.30 | 88.39±3.25 |
| | | **mix** | **79.17**±3.18 | **94.05**±2.06 | **96.73**±2.23 |
| avec finetuning | phrases | **vsl**$_v$ | 63.10 ±0.56 | 86.90±0.21 | 92.26±0.51 |
| | | **vsl**$_m$ | 63.10±0.44 | 86.31±0.85 | 92.86±0.39 |
| | | **ctx**$_c$ | 68.45±0.22 | 87.20±0.52 | 91.37±0.24 |
| | | **ctx**$_m$ | 67.86±1.04 | 88.39±1.12 | 91.37±0.36 |
| | | **mix**$_m$ | **76.19**±0.58 | **94.64**±0.55 | **97.92**±0.51 |
| | textes | **vsl**$_v$ | 71.13±1.02 | 82.14±0.52 | 88.99±0.71 |
| | | **ctx**$_c$ | 70.24±1.72 | 85.71±0.59 | 89.88±1.14 |
| | | **mix**$_m$ | **81.55**±0.74 | **94.94**±0.29 | **97.32**±0.30 |

TABLE 2 – Résultats sur la tâche de recherche de texte

En recherche de texte, les performances sur les protocoles visuels et contextuels sont équivalentes. Cependant, elles sont meilleures lorsque les deux caractéristiques sont mélangées (**mix**). Pour comprendre cette observation, nous avons analysé le type de texte récupéré (cf Tab. 3). Le modèle utilise de manière complémentaire les deux types de textes, avec une tendance à mieux retrouver les textes visuels (**vsl**) dans l'ensemble de test.En recherche d'image, le modèle est meilleur sur les phrases visuelles (cf Tab. 4), qui sont plus proches de la modalité que l'on recherche. Les performances s'avèrent significativement supérieures pour la recherche à partir de textes longs par rapport aux phrases seules. L'application d'un finetuning améliore les performances mais ne modifie pas ces observations.

| | **phrases visuelles** | **phrases contextuelles** |
|---|---|---|
| **R@1** | 29.17 | 45.54 |
| **R@5** | 41.43 | 40.36 |
| **R@10** | 45.96 | 36.34 |

TABLE 3 – Rappel sur les phrases mixtes de l'ensemble de test, sans finetuning

| | | | **Recherche d'image** | | |
|---|---|---|---|---|---|
| | | | **R@1** | **R@5** | **R@10** |
| sans finetuning | phrases | **vsl** | **45.08**±1.78 | **64.46**±1.81 | **73.29**±1.72 |
| | | **ctx** | 34.26±0.88 | 54.03±1.40 | 61.47±1.29 |
| | | **mix** | 36.20±1.11 | 56.94±0.34 | 64.96±0.25 |
| | textes | **vsl** | **59.46**±1.20 | **83.18**±1.44 | **89.46**±1.15 |
| | | **ctx** | 43.66±1.16 | 63.07±1.78 | 69.79±1.37 |
| | | **mix** | 51.46±0.67 | 71.61±0.04 | 77.08±0.28 |
| avec finetuning | phrases | **vsl**$_v$ | **45.60**±0.60 | **70.94**±0.21 | 78.47±0.23 |
| | | **vsl**$_m$ | 44.81±0.28 | 69.96±0.44 | **79.06**±0.28 |
| | | **ctx**$_c$ | 33.16±0.23 | 53.99±0.33 | 61.58±0.27 |
| | | **ctx**$_m$ | 34.22±0.74 | 52.63±0.25 | 61.09±0.49 |
| | | **mix**$_m$ | 38.05±0.59 | 58.85±0.22 | 67.26±0.38 |
| | textes | **vsl**$_v$ | **65.06**±0.52 | **87.62**±0.27 | **92.59**±0.23 |
| | | **ctx**$_c$ | 45.09±0.42 | 65.74±0.54 | 71.88±0.22 |
| | | **mix**$_m$ | 53.07±0.15 | 72.44±0.13 | 78.87±0.12 |

TABLE 4 – Résultats sur la tâche de recherche d'image

Nous avons également comparé un apprentissage uniquement sur les données de la catégorie cible (ex : apprentissage sur des phrases visuelles et évaluation sur du visuel (**vsl**$_v$)) à un apprentissage reposant sur des données mixtes (ex : apprentissage sur des phrases mixtes et évaluation sur du visuel (**vsl**$_m$)). Nous n'avons pas noté de différence de performance significative nous invitant à penser qu'il peut être préférable d'utiliser toutes les données lors de l'apprentissage. Notons que nous ne présentons pas les résultats de [7] dont le protocole est plus simple et les performances notablement plus faibles.

## 4 Conclusion et perspectives

Nous avons montré que la tâche de recherche d'image est intéressante pour analyser des textes longs comportant à la fois des phrases contextuelles et visuelles. De plus, nous observons de meilleures performances sur le contenu visuel que sur le contenu contextuel du fait d'une saillance ou distinction plus importante. Néanmoins, le déséquilibre des classes lié à une plus grande proportion de données contextuelles par rapport aux données visuelles, conduit à une préférence naturelle du système pour choisir le contenu contextuel lors de la recherche. Dans le prolongement de ces recherches, nous envisagerons d'ajuster l'apprentissage de la recherche texte-image afin de tenir compte du type de phrase présente dans un texte, en forçant le modèle à choisir des phrases à caractère visuel.

**Remerciements.** Cette publication a été réalisée à l'aide des ressources de calcul du CRIANN et le premier auteur est financé par le GIP Millénaire Caen 2025 (www.millenairecaen2025.fr) dans le cadre du projet Art Bien-être et Cerveau.

# Références


[1] J. She et E. Cetinic, Understanding and Creating Art with AI : Review and Outlook, *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 18, p. 1-22, 2022.

[2] D. B. Ebaid, M. M. Madbouly, et A. A. El-Zoghabi, Bi-directional Image–Text Matching Deep Learning-Based Approaches : Concepts, Methodologies, Benchmarks and Challenges, *International Journal of Computational Intelligence Systems*, vol. 16, p. 1-22, 2023.

[3] R. Guo et al., A Survey on Image-text Multimodal Models, *arXiv*, p. 1-22, 2023.

[4] M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, et A. Mian, Visual Attention Methods in Deep Learning : An In-Depth Survey, *arXiv*, p. 1-20, 2022.

[5] P. Xu, X. Zhu, et D. A. Clifton, Multimodal Learning With Transformers : A Survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1-20, 2023.

[6] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, et M. Shah, Transformers in Vision : A Survey, *ACM Computing Surveys*, vol. 54, p. 200 :1-200 :41, 2022.

[7] M. Stefanini, M. Cornia, L. Baraldi, M. Corsini, et R. Cucchiara, Artpedia : A New Visual-Semantic Dataset with Visual and Contextual Sentences in the Artistic Domain, *ICIAP Image Analysis and Processing*, p. 729-740, 2019.

[8] L. Baraldi, M. Cornia, C. Grana, et R. Cucchiara, Aligning Text and Document Illustrations : Towards Visually Explainable Digital Humanities, *ICPR International Conference on Pattern Recognition*, vol 24th, p. 1097-1102, 2018.

[9] A. Radford et al., Learning Transferable Visual Models From Natural Language Supervision, *PMLR International Conference on Machine Learning*, p. 8748–8763, 2021.

[10] J. Li, D. Li, C. Xiong, et S. Hoi, BLIP : Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, *ICML International Conference on Machine Learning*, 2022.

[11] D. Li et al., LAVIS : A Library for Language-Vision Intelligence, *arXiv*, 2022.

[12] M. Lewis et al., Bart : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv*, 2019.

[13] J. Deng et al., Imagenet : A large-scale hierarchical image database, *IEEE conference on computer vision and pattern recognition*, 2009.

[14] A. Dosovitskiy et al., An Image is Worth 16x16 Words : Transformers for Image Recognition at Scale. *ICLR International Conference on Learning Representations*, 2021.

[15] J. Devlin, M.-W. Chang, K. Lee, et K. Toutanova, BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding *arXiv*, 2019.

[16] D. P. Kingma et J. BA. Adam : A method for stochastic optimization. *arXiv*, 2014.