Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Vers une pédagogie inclusive : une classification multimodale des illustrations de manuels scolaires pour des environnements d'apprentissage adaptés

Saumya Yadav[1], Élise Lincker[2], Caroline Huron[3], Stéphanie Martin[3], Camille Guinaudeau[4, 5], Shin'ichi Satoh[5] and Jainendra Shukla[1]
(1) HMI Lab, IIIT-Delhi, India
(2) Cedric, CNAM, Paris, France
(3) Le Cartable Fantastique, Paris, France
(4) Japanese French Laboratory for Informatics, CNRS, Japan
(5) National Institute of Informatics, Tokyo, Japan
saumya@iiitd.ac.in

RÉSUMÉ

Afin de favoriser une éducation inclusive, des systèmes automatiques capables d'adapter les manuels scolaires pour les rendre accessibles aux enfants en situation de handicap sont nécessaires. Dans ce contexte, nous proposons de classifier les images associées aux exercices selon trois classes (*Essentielle*, *Informative* et *Inutile*) afin de décider de leur intégration ou non dans la version accessible du manuel pour les enfants malvoyants. Sur un ensemble de données composé de 652 paires (texte, image), nous utilisons des approches monomodales et multimodales à l'état de l'art et montrons que les approches fondées sur le texte obtiennent les meilleurs résultats. Le modèle CamemBERT atteint ainsi une exactitude de 85,25 % lorsqu'il est combiné avec des stratégies de gestion de données déséquilibrées. Pour mieux comprendre la relation entre le texte et l'image dans les exercices des manuels, nous effectuons également une analyse qualitative des résultats obtenus avec et sans la modalité image et utilisons la méthode LIME pour expliquer la décision de nos modèles.

ABSTRACT

Towards Inclusive Pedagogy : A multimodal Classification of Textbook Illustrations for Adaptive Learning Environments

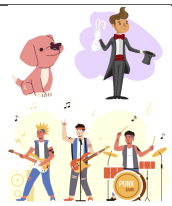
To foster inclusive education, automatic systems that can adapt textbooks to make them accessible to children with disabilities are necessary. In this context, we propose a task to classify the images associated with the exercises according to three classes (*Essential*, *Informative*, and *Useless*) to decide whether to integrate them into the accessible version of the textbook for visually impaired children. On a dataset composed of 652 (text, image) pairs, we use state-of-the-art monomodal and multimodal approaches and show that text-based approaches achieve better results. The CamemBERT model achieves an accuracy of 85.25% when combined with unbalanced data management strategies. To better understand the relationship between text and image in textbooks' exercises, we also perform a qualitative analysis of the results obtained with and without the image modality and use the LIME method to explain the decision of our models.

MOTS-CLÉS : Classification multimodale · Éducation inclusive · Explicabilité des modèles.

KEYWORDS: Multimodal Classification · Inclusive Education · Model explainability.

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 708–719.

TABLE 1 – Exemples de classification de paires (texte, image)

Classe	Essentielle	Informative	Inutile
Images			
Texte	Écris le son commun aux 3 objets représentés par les dessins.	Texte : À la préhistoire, les hommes dessinaient des peintures rupestres sur les murs de leur caverne - Q : Trouve le verbe. À quel temps est il conjugué ?	Recopie les phrases si tu reconnais le verbe “aller”. (a) Je quitte la maison à la même heure tous les matins. (b) Samedi, je me suis baladé dans le parc.

1 Introduction

Le droit à l’éducation est universel, transcendant les limitations imposées par des handicaps physiques ou cognitifs. Cependant, les ressources éducatives conventionnelles, en particulier les manuels scolaires, ne sont pas intrinsèquement conçues pour répondre aux besoins divers des apprenants, surtout ceux en situation de handicap. Cette disparité dans l’accessibilité des matériaux éducatifs entrave considérablement le processus d’apprentissage pour les enfants ayant des besoins spéciaux, amplifiant ainsi l’écart éducatif.

L’évolution des techniques d’intelligence artificielle a ouvert de nouvelles voies pour l’apprentissage adapté, pourtant l’intégration de ces technologies dans le soutien des enfants en situation de handicap reste peu explorée. Spécifiquement, les apprenants malvoyants rencontrent des barrières découlant de la nature visuelle des matériaux éducatifs standards, qui sont principalement basés sur le texte et l’image. Cela limite leur capacité à accéder à l’information et affecte leur engagement et leur motivation. Des associations commencent à produire des manuels numériques adaptés aux enfants en situation de handicaps, en effectuant les transformations à la main. Malheureusement, étant donné la grande diversité des collections et le renouvellement des programmes d’enseignement, ces adaptations artisanales ne permettent pas de répondre aux besoins. Dans ce contexte, l’utilisation d’approches automatiques est indispensable pour rendre accessible les matériaux éducatifs au plus grand nombre.

L’automatisation de l’adaptation de manuels scolaires a été peu étudiée. Lincker *et al.* (2023b) a été pionnier dans la classification des exercices des manuels en fonction de leurs objectifs d’apprentissage, facilitant l’adaptation automatique des manuels pour les enfants ayant des problèmes de coordination motrice (dyspraxie). Cette adaptation atténue le besoin d’écriture manuelle et rationalise les interactions avec les manuels tout en préservant l’intégrité éducative. Cependant, les auteurs se concentrent sur la mise en page et le contenu textuel des manuels, ce qui ne correspond pas aux besoins des élèves malvoyants. Afin de combler cette lacune, notre travail propose donc un nouveau cadre de classification des images accompagnant les exercices des manuels afin de déterminer leur caractère facultatif ou non. La classification de ces images dans trois classes (*Essentielle*, *Informative* et *Inutile*), voir exemples dans le Tableau 1, est cruciale pour l’adaptation des manuels, guidant les décisions sur l’inclusion d’une image dans l’interface utilisateur (avec du texte alternatif généré) ou son omission pour un document plus clair et plus accessible, adapté à la consommation auditive via un assistant vocal. Pour cela, nous annotons un ensemble de données de plus de 600 paires (texte, image) avec ces 3 catégories et utilisons des algorithmes de classification monomodaux et multimodaux à l’état de l’art. Notre motivation pour combiner les deux modalités, image et texte repose sur l’intuition qu’une

image *Inutile* présente un chevauchement sémantique significatif avec le texte (l'image ne fournit pas d'informations supplémentaires), tandis qu'une image *Essentielle* sera sémantiquement très différente du texte de l'exercice qu'elle illustre. Nous comparons également cette approche multimodale avec des méthodes monomodales pour analyser l'impact des différentes modalités indépendamment. Les principales contributions de ce travail sont triples : (1) une comparaison des approches multi et monomodales pour la classification (texte, image) ; (2) une comparaison qualitative des résultats pour mieux comprendre l'impact de chaque modalité ; (3) une analyse des fonctionnalités utilisées par le modèle à travers la méthode d'explicabilité LIME (Ribeiro *et al.*, 2016).

2 État de l'art

Le travail présenté dans cet article est lié à différents domaines : le Traitement Automatique des Langues appliqué aux manuels scolaires et la similarité texte-image.

La recherche appliquée aux manuels scolaires est relativement rare dans le domaine du traitement automatique des langues. La plupart des études existantes se concentrent soit sur l'analyse du contenu linguistique des manuels ((Green, 2019; Lucy *et al.*, 2020)), sur la génération de questions à partir de ceux-ci ((Ch & Saha, 2022; Gerald *et al.*, 2022)) ou la création de ressources lexicales qui pourraient être utilisées pour la classification et la représentation des manuels (Manulex (Lété *et al.*, 2004), ReSyf (Billami *et al.*, 2018) ou Alector (Gala *et al.*, 2020)). Semblable à notre objectif d'adapter les manuels pour les enfants en situation de handicap, des études récentes se sont concentrées sur la modélisation et l'extraction de contenu à partir de manuels (Lincker *et al.*, 2023b) ou la classification d'exercices basée sur leur objectif éducatif (Lincker *et al.*, 2023a). Cependant, ces travaux se concentrent uniquement sur l'analyse de la mise en page et du texte et non sur les images présentes dans le manuel.

L'analyse de la similarité ou de la nature de la relation entre un texte et une image associée est souvent fondée sur des *transformers* vision-langage pré-entraînés qui se basent typiquement sur des ensembles de données de légendes d'images tels que MS COCO (Lin *et al.*, 2014) ou Flickr30k (Young *et al.*, 2014) pour évaluer leurs modèles (par exemple, (Rao *et al.*, 2022) sur les tâches de recherche d'information ou (Huang *et al.*, 2019) sur les tâches de légendage d'images). Ces ensembles de données comprennent des images complexes représentant de multiples objets dans des arrière-plans riches. Malgré la richesse des domaines visuels dans ces ensembles de données, leurs légendes tendent à être des descriptions en une seule phrase, alors que dans notre ensemble de données, les relations entre le texte et l'image sont plus variées, où l'image et le texte peuvent être soit redondants, soit complémentaires. L'analyse comparative des modalités image et texte a été révolutionnée par l'introduction du modèle Contrastive Language–Image Pre-training (CLIP) (Radford *et al.*, 2021) qui apprend les concepts visuels à partir des descriptions textuelles, facilitant une association plus nuancée entre le texte et les images que les modèles traditionnels. Ce modèle offre la capacité d'estimer la similarité entre un texte et une image, largement utilisée dans la recherche d'informations cross-modale ou les système de questions-réponses visuels (*Visual Question Answering* (VQA)). Plus étroitement liés à notre travail, deux articles récents analysent la relation entre le texte et l'image dans le contexte de la recherche d'information image-texte (Qu *et al.*, 2021) et de la classification (Otto *et al.*, 2020). Otto *et al.* (2020) présentent un cadre de classification analysant les relations sémantiques entre les images et les descriptions textuelles. Ils définissent huit classes, s'appuyant sur trois concepts : l'Information Mutuelle Cross-Modale, la Corrélation Sémantique, et le Statut (qui décrivent la relation hiérarchique entre le texte et l'image). L'étude implique la création automatique d'ensemble de

données à partir de MSCOCO, VIST (Malakan *et al.*, 2023) et ImageNET (Deng *et al.*, 2009) et se base sur deux classifieurs d'apprentissage profond, un classique et un en cascade, pour évaluer la difficulté de la tâche. Bien que ces ensembles de données soient publiquement disponibles, ils diffèrent significativement de notre objectif. Les parties textuelles associées aux images dans ces ensembles de données consistant principalement en une légende d'image d'une seule phrase (MSCOCO) ou une étiquette d'un seul mot (ImageNet). Le texte dans nos exercices, destiné à poser une question, peut en effet être constitué de plusieurs phrases et servir un but légèrement différent.

3 Données

Les images des activités et leçons de manuels scolaires jouent différents rôles, et reconnaître l'importance de ces éléments visuels est crucial pour l'adaptation automatique des manuels. Dans ce travail, les exercices avec images ont été extraits de trois manuels scolaires français pour le primaire. Pour ce faire, chaque manuel au format PDF est converti en fichier XML au format ALTO en utilisant les outils pdfalto [1] et MuPDF [2]. Cette approche permet l'extraction des mots dans une représentation structurée et organisée du contenu tout en fournissant des informations sur la mise en page et le style de police. Suivant la méthode employée par Lincker *et al.* (2023b), les mots extraits sont ensuite regroupés en segments de texte, qui à leur tour sont regroupés en blocs d'activités en fonction de la mise en page, du style de police et des caractéristiques d'espacement. Les images sont associées aux blocs selon leur position sur la page. Finalement, deux experts ont manuellement annoté les images avec leur texte respectif en trois classes différentes :

- `Images Essentielles` : Ces images sont indispensables pour comprendre ou résoudre une activité.
- `Images Informatives` : Elles contribuent à la compréhension du texte et fournissent des informations supplémentaires sans être essentielles pour résoudre l'exercice (ajout d'indices pour résoudre l'exercice ou explication sur un concept inconnu des élèves).
- `Images Inutiles` : Elles peuvent être exclues lors de l'adaptation pour les enfants en situation de handicap afin de simplifier l'interface adaptée.

Pour simplifier au maximum l'interface adaptée pour les enfants malvoyants, il est essentiel d'exclure les images de la classe *Inutile* des manuels adaptés. Un exemple de la classification des images avec leur texte respectif est présenté dans le Tableau 1. Dans la classe *Essentielle*, l'image est obligatoire pour résoudre l'exercice, tandis que le but de l'image de la classe *Informative* est de donner des informations supplémentaires à l'élève, qui peut ne pas savoir de qu'est exactement une peinture rupestre. Enfin, dans le dernier exemple, l'image associée au texte « Copie les phrases si tu... » n'est pas utile pour résoudre l'exercice et n'a qu'un but décoratif.

Trois manuels scolaires français ont été utilisés dans notre étude ; deux provenant du même éditeur et un troisième venant d'un éditeur différent. Pour maintenir une évaluation rigoureuse, nous avons combiné les manuels du même éditeur et les avons partitionnés en ensembles d'entraînement et de validation en utilisant un ratio de 80/20. L'ensemble de test est, quant à lui, constitué à partir des paires (texte, image) extraites du manuel d'un éditeur différent afin de garantir que les modèles sont évalués sur des données non vues lors de l'entraînement. Le Tableau 2 présente le nombre de paires (texte, image) ainsi que le nombre de mots correspondants pour chaque classe des ensembles d'entraînement / validation et de test. Le Tableau 3 illustre, quand à lui, les 5 mots les plus fréquents dans chacune des 3 classes annotées, révélant des motifs distincts dans les occurrences de mots à

1. https://github.com/kermitt2/pdfalto
2. https://github.com/ArtifexSoftware/mupdf

TABLE 2 – Nombre de paires (texte, image) et nombre de mots différents dans chaque classe

	Essentielle		Informative		Inutile	
	# mots	# paires	# mots	# paires	# mots	# paires
Entraînement+Validation	16.0	258	21.5	96	11.2	58
Test	13.5	131	32.4	75	11.2	42

TABLE 3 – Mots les plus fréquents dans chaque classe

Test			Entraînement+Validation		
Essentielle	Informative	Inutile	Essentielle	Informative	Inutile
nom : 29	mot : 77	texte : 14	écris : 160	mot : 93	verbe : 21
dessin : 27	texte : 46	verbe : 11	dessin : 103	texte : 47	texte : 20
écris : 24	verbe : 41	phrase : 10	mot : 101	verbe : 42	phrase : 16
trouver : 23	observe : 34	recopie : 8	nom : 89	observe : 24	mot : 15
donne : 17	phrase : 25	combine : 7	utilise : 75	écris : 23	écris : 15

travers les catégories. En effet, nous pouvons remarquer que les mots « nom », « dessin » et « écris » sont assez caractéristiques de la classe *Essentielle*, alors que les classe *Informative* et *Inutile* tendent à partager davantage les mots qui apparaissent les plus fréquemment (« texte », « verbe » ou « phrase »).

4 Approches

Pour la classification des illustrations, nous avons utilisé différentes modalités. Pour cela, nous avons utilisé des approches multimodales, qui considèrent le texte et l'image simultanément ainsi que des méthodes monomodales qui se fondent uniquement sur le texte ou sur l'image.

4.1 Approche multi-modale

CLIP (Radford *et al.*, 2021) est une architecture permettant d'apprendre des représentations visuelles à partir d'une faible supervision textuelle. Nos données textuelles étant extraites de livres en français, nous avons tout d'abord traduit les textes en anglais en utilisant la bibliothèque de traduction hors ligne open source Argos Translate [3], qui repose sur OpenNMT (Klein *et al.*, 2017), pour répondre à l'exigence d'entrée de texte en anglais de CLIP. Nous avons utilisé la variante RN101 de CLIP car elle donnait de meilleurs résultats sur notre ensemble de données de validation.

Dans cette première approche multi-modale, nous avons calculé la similarité entre les images et les textes à partir du modèle CLIP afin d'analyser la relation entre l'image et le texte, à partir de l'équation suivante :

$$CLIPScore(I,T) = max(100 * cos(E_I, E_T), 0) \quad (1)$$

où E_I et E_T correspondent aux plongements de l'image et du texte, respectivement. Nos données textuelles pouvant dépasser la longueur par défaut de 77 *tokens* définie par CLIP, nous avons utilisé deux stratégies différentes pour gérer cette limitation : la troncature et la segmentation. Dans le

3. https://github.com/argosopentech/argos-translate

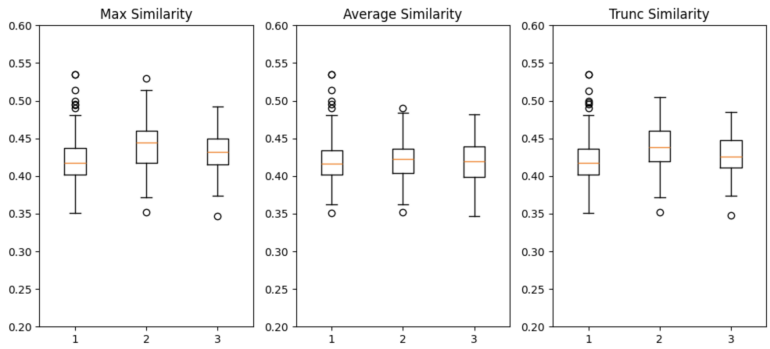


FIGURE 1 – Similarité texte-image calculée à partir du modèle CLIP pour chaque classe (1 : *Essentielle*, 2 : *Informative*, 3 : *Inutile*)

premier cas, les textes sont tronqués à la longueur par défaut alors qu'ils sont divisés en plusieurs segments dans le second cas (la similarité globale entre le texte et l'image correspond alors à la moyenne ou à la similarité maximale de tous les segments textuels). La Figure 1 présente les valeurs de similarité obtenues avec ces deux stratégies. Nous pouvons constater que, contrairement à notre intuition initiale, il n'existe que peu de différences de similarité entre texte et image en fonction des 3 classes. Par conséquent, nous définissons une deuxième approche multimodale, consistant à combiner les plongements des images obtenues grâce à la même variante RN101 du modèle CLIP et les plongements des textes, calculés grâce aux modèles de langue présentés dans la section suivante, pour les fournir à un Perceptron Multi-Couches (*Multi-Layers Perceptron* - MLP).

4.2 Approches mono-modales

Fondées sur le texte Pour l'encodage des plongements basés sur le texte dans l'approche monomodale, nous avons utilisé le modèle de langue BERT (Kenton & Toutanova, 2019), appliqué sur les données traduites en anglais, pour une comparaison plus directe avec les résultats fournis par CLIP, et le modèle de langue CamemBERT (Martin *et al.*, 2020) appliqué directement sur les exercices en français. Nous avons utilisé la même approche de classification que celle utilisée dans le cadre multi-modal. Les textes des exercices ont d'abord été tokenisés avant d'être fournis aux modèles `bert-base-uncased`[4] pour l'anglais et `camembert-base`[4] pour le français, permettant l'extraction de plongements contextualisées à partir de la dernière couche cachée de l'architecture BERT et CamemBERT. Un pooling moyen adaptatif a été appliqué à travers la dimension de la longueur de la séquence afin d'obtenir une représentation de taille fixe, encapsulant les informations essentielles du texte d'entrée. Les plongements textuels obtenus à travers ces modèles sont ensuite soumis à un MLP pour la classification.

Finalement, pour améliorer davantage notre représentation sémantique des exercices, nous utilisons le modèle de langage CamemBERT affiné sur des données éducatives : leçons et activités de quatre manuels (deux manuels de la collection utilisée pour l'entraînement, excluant les exercices utilisés pour construire notre jeu de données, et deux autres manuels non vus), 1293 Fantastiques Exercices fournis par l'association Le Cartable Fantastique[5], et les 79 textes de lecture originaux d'Alector. Ce modèle de langue *CamemBERT-education*, proposé dans Lincker *et al.* (2023a), est utilisé de façon similaire à CamemBERT pour calculer les plongements des textes des exercices.

4. L'utilisation de modèles plus large sur notre ensemble de validation a fourni de moins bons résultats.
5. https://www.cartablefantastique.fr/

Fondées sur l'image Pour l'extraction des plongements des images des exercices, nous avons utilisé les modèles ResNET (He *et al.*, 2016), VGG16 (Simonyan & Zisserman, 2015) et Inception-v3 (Szegedy *et al.*, 2016), qui sont connus pour leur efficacité sur les tâches de reconnaissance d'image. L'étape initiale consiste à charger un modèle `ResNet-50` (ou `VGG16` ou `Inception-v3`) pré-entraîné puis de personnaliser la dernière couche du modèle afin de produire des plongements de taille 512, identiques à ceux du modèle CLIP. Toutes les couches, à l'exception de la dernière couche modifiée, sont gelées pour préserver les connaissances encodées dans les couches antérieures. Par ailleurs, les images ont été préalablement redimensionnées pour garantir la compatibilité avec les attentes d'entrée du modèle ResNET.

Finalement, suivant la même méthode que pour le modèle CLIP ou la classification monomodale basée sur le texte, les plongements extraits sont fournis à un MLP pour notre tâche de classification.

4.3 Gestion des données déséquilibrées

Les données annotées dans le cadre de ce travail sont fortement déséquilibrées, comme le montre le Tableau 2. Le déséquilibre des classes pouvant affecter la performance du modèle, en favorisant particulièrement la classe majoritaire, nous avons utilisé deux stratégies pour gérer ce problème de déséquilibre, conjointement ou séparément :

— `Stratégie de pondération des classes` **:** Cette stratégie consiste à attribuer différents poids aux classes afin d'accorder plus d'importance à la classe minoritaire. Les pondérations sont calculées à l'aide de la fonction `compute_class_weight` de la bibliothèque Python `scikit-learn`, avec la stratégie *balanced* qui ajuste dynamiquement le poids des classes en fonction de leur distribution dans les données d'entraînement, fournissant des poids plus élevés aux classes sous-représentées.

— `Génération de données` **:** Dans cette stratégie, l'ensemble d'entraînement initial est complété par 150 instances de la classe *Inutile*. Ces nouvelles instances correspondent à des exercices sans image extraits de nos manuels d'entraînement auxquels nous avons associé des images aléatoires provenant de la classe *Inutile* au sein des mêmes manuels. Par la suite, nous avons fusionné ces données augmentées avec les données d'entraînement originales et suivi la même procédure d'extraction de plongements et de passage à travers le MLP que nous l'avions fait pour les modèles précédents.

5 Expérimentations

5.1 Configuration

Le modèle CLIP propose plusieurs variantes, et nous avons sélectionné RN101 pour sa performance supérieure sur la partie validation de notre jeu de données, avec un plongement de dimension 512. Nous avons ensuite extrait les plongements des données monomodales, en maintenant la même dimension et la même procédure qu'avec le modèle CLIP. Nous avons utilisé diverses techniques pour fusionner les modalités, incluant l'extraction du maximum et du minimum de deux plongements et leur addition. Les résultats optimaux, sur notre ensemble de validation, ont été obtenus en concaténant les deux plongements. Ainsi, nous les avons concaténé en un vecteur de taille 1024 que nous avons fourni en entrée du MLP. L'architecture MLP utilisée se compose d'une couche d'entrée, de deux couches cachées avec activation ReLU, et d'une couche de sortie pour les tâches de classification.

TABLE 4 – Résultat de classification avec les modèles mono et multi-modaux

Modèles	Modalité	Exactitude
Classe majoritaire	-	0.7267
BERT	texte	0.8156
CamemBERT	texte	0.8361
CamemBERT-educational	texte	0.80
ResNET	image	0.5246
Inception-v3	image	0.5041
VGG16	image	0.50
BERT+ResNET	texte+image	0.7582
CamemBERT+ResNET	texte+image	0.8033
CLIP	texte+image	0.8074

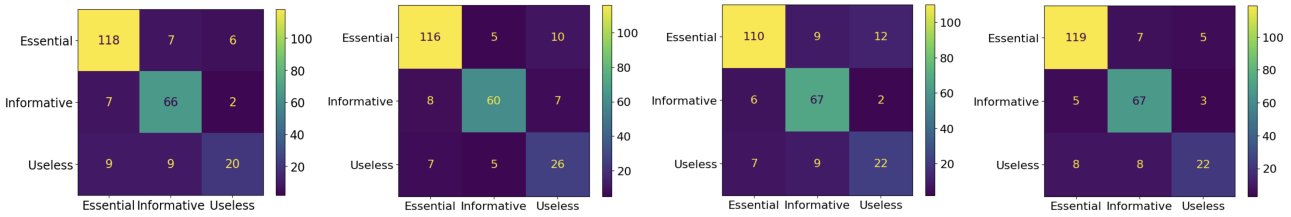
TABLE 5 – Résultats de classification avec CamemBERT et les stratégies de gestion de données déséquilibrées.

Class Weight	Data Augmentation	Exactitude
✗	✗	0.8361
✓	✗	0.8156
✗	✓	0.8279
✓	✓	**0.8525**

5.2 Résultats

Le Tableau 4 présente les résultats obtenus sur le jeu de données de test, pour les approches de classification mono-modales et multi-modales. Nous pouvons constater que les modèles basés sur le texte surpassent leurs homologues basés sur l'image. En effet, la classification basée sur l'image donne des résultats inférieurs à la classe majoritaire pour les trois modèles (ResNET, VGG16 et Inception-v3) suggérant que l'image seule n'est pas suffisante pour décider si une image est nécessaire, informative ou inutile dans le contexte d'un exercice. Pour la classification basée sur le texte, la meilleure performance est obtenue avec le modèle de langue CamemBERT sans affinage des données éducatives. CamemBERT-education étant supposé avoir une meilleure représentation sémantique des données éducatives en français (il offre de meilleurs résultats sur une tâche de classification des exercices en fonction de leurs objectifs pédagogiques (Lincker *et al.*, 2023a)), sa faible performance sur notre tâche de classification suggère que les plongements sémantiques de la partie textuelle des exercices ne constituent pas un facteur décisif dans le processus de classification des images.

Par ailleurs, le Tableau 4 montre que les approches multimodales n'améliorent pas les performances par rapport à la classification basée sur le texte seulement. Dans ce cas, les meilleurs résultats sont obtenus grâce au modèle CLIP, légèrement au-dessus de ceux obtenus avec la concaténation des plongements dérivés du texte et de l'image, en utilisant respectivement CamemBERT et ResNet. Ces faibles résultats s'expliquent par le fait que, comme montré précédemment sur la Figure 1, les valeurs de similarités entre le texte et l'image calculées avec le modèle CLIP ont des valeurs presque similaires pour les trois classes, contrairement à notre intuition qui était que les images inutiles avaient une similarité sémantique plus élevée avec le texte (redondance) tandis que les images nécessaires avaient une similarité sémantique plus faible (complémentarité).



(a) (b) (c) (d)

FIGURE 2 – Matrices de confusion pour le modèle CamemBERT seul (a), avec une augmentation de données (b), avec une pondération des classes (c) et avec une combinaison des deux stratégies (d) (axe des x = classe de référence, axe des y = classe prédite)

TABLE 6 – Comparaison des résultats obtenus avec les approches mono- et multi-modales. ✓(resp. ✗) correspond à la classification correcte (resp. incorrecte) de la paire (texte, image)

	Exercice 1	Exercice 2	Exercice 3
Texte	Résouds ce rébus.	Choisis les bons adjectifs pour décrire la princesse.	Quel type d'art est-ce ?
Image			
Monomodal (texte)	✗	✓	✓
Monomodal (image)	✓	✓	✗
Multimodal	✗	✗	✗

Finalement, le Tableau 5 présente les résultats obtenus avec les différentes stratégies utilisées pour traiter notre problème de déséquilibre des données. Les meilleurs résultats sont obtenus lorsque les stratégies de pondération des classes et d'augmentation des données sont utilisées conjointement, atteignant une exactitude de 85,25%. D'un point de vue qualitatif (cf. Figure 2), l'augmentation des données tend à classer plus d'exemples dans la classe *Inutile*, à la fois correctement et incorrectement, (2b), tandis que la stratégie de pondération tend à améliorer le nombre d'instances correctement classées pour les classes sous-représentées (*Inutile* et *Informative*) aux dépends de la classe *Essentielle*, (2c). Enfin, combiner les deux stratégies améliore globalement le nombre d'instances correctement classifiées pour toutes les classes, (2d).

5.3 Analyse des résultats

Le Tableau 6 présente la comparaison des résultats obtenus avec des modèles mono-modaux (basés sur le texte ou sur l'image) ou multimodaux sur 3 exercices de la classe *Essentielle*. Sur ces exercices, le modèle basé sur le texte étiquette incorrectement l'exercice 1, ayant pour objectif l'analyse d'image (lecture d'un rébus), tandis que le modèle basé sur l'image fournit de bons résultats pour les exercices 1 et 2, c'est-à-dire dans des cas d'analyse et de description d'images. Cependant, la modalité visuelle ne permet pas de classifier correctement l'exercice 3, que ce soit dans une approche mono- ou multi-modale, alors que la modalité textuelle (sous la forme d'une question ouverte) est suffisante pour la classification. Finalement, l'approche multimodale fournit de mauvais résultats pour ces 3 exemples, mettant en évidence les défis dans l'intégration efficace des modalités texte et image.

Afin d'améliorer l'interprétabilité de nos différents modèles, nous utilisons la méthode LIME (pour Explications Interprétables Locales Agnostiques au Modèle - *Local Interpretable Model-agnostic*

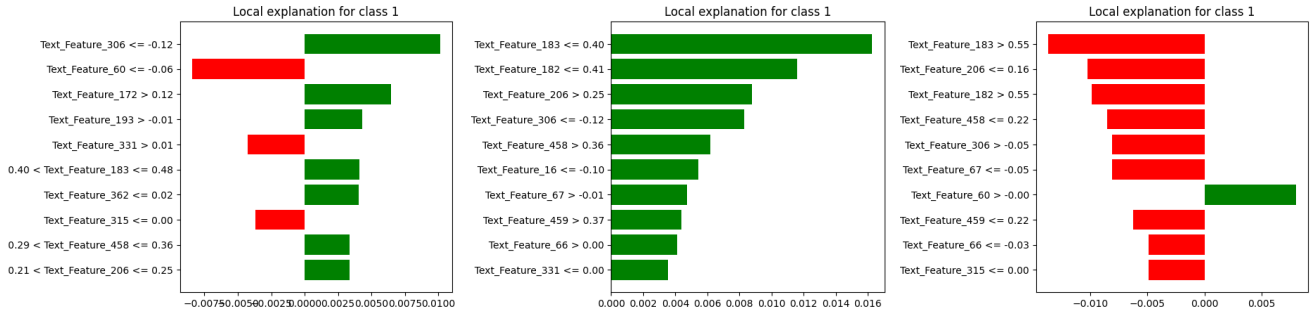


FIGURE 3 – Explications fournies par LIME pour 3 instances aléatoires des classes (a) Inutile (b) Informative et (c) Essentielle

Explanations) (Ribeiro *et al.*, 2016) qui fournit des explications localisées pour des prédictions individuelles et améliore la transparence. La Figure 3 présente le résultat de LIME pour trois instances aléatoires de trois classes différentes, où l'axe des y représente les 10 meilleures caractéristiques extraites, c'est-à-dire, les plongement produits par CamemBERT. Pour la classe *Inutile*, la prédiction semble être influencée par un équilibre délicat de contributions positives et négatives de diverses caractéristiques, suggérant que la frontière de décision pour la classe *Inutile* est nuancée, sans aucune caractéristique dominante orientant la prédiction. Au contraire, les classes *Informative* et *Essentielle* présentent des contributions de caractéristiques positives plus prononcées indiquant des influences plus fortes sur la décision du modèle. Par ailleurs, certaines caractéristiques, telles que les caractéristiques 183 ou 206, présentent des impacts variables à travers les classes indiquant leur pertinence contextuelle dans la différenciation entre les classes *Essentielle*, *Informative* et *Inutile*.

6 Conclusion et perspectives

Notre étude aborde le besoin impératif d'une éducation inclusive en proposant un système automatique de classification des images associées aux exercices de manuels scolaires en 3 classes (*Essentielle*, *Informative* et *Inutile*). Nous avons utilisé un ensemble de données composé de 652 paires (texte, image) et avons exploré les approches monomodales et multimodales pour la classification. Étonnamment, les méthodes monomodales basées sur le texte ont surpassé leurs homologues multimodales. Notre analyse des résultats montre que l'aspect sémantique n'est probalement pas le plus important pour la classification des images, et que des éléments surfaciques du texte de l'exercice jouent une part importante dans la classification. Finalement, nous avons utilisé la méthode d'explication LIME pour obtenir des aperçus du processus de prise de décision de nos modèles et ainsi montrer que les classes *Informative* et *Essentielle* étaient fortement caractérisées contrairement à la classe *Inutile*. Malgré ces résultats prometteurs, notre étude présente certaines limitations, la principale étant la quantité relativement faible de nos données, qui ne sont par ailleurs pas partageables avec la communauté, pour des raisons de propriété intellectuelle, ce qui entrave la reproductibilité de nos expériences. Cette limitation souligne la nécessité de disposer de jeux de données plus larges et disponibles publiquement. À l'avenir, nous prévoyons d'élargir notre ensemble de données en (1) extrayant plus de manuels scolaires (2) incorporant des données de Otto *et al.* (2020) afin d'améliorer la représentation de la classe *Informative*.

Remerciements Ce travail a été financé par le *NII International Internship Program* et le projet MALIN (MAnuels scoLaires INclusifs / ANR-21-CE38-0014).

Références

BILLAMI M. B., FRANÇOIS T. & GALA N. (2018). ReSyf : a French lexicon with ranked synonyms. In *International Conference on Computational Linguistics*.

CH D. R. & SAHA S. K. (2022). Generation of multiple-choice questions from textbook contents of school-level subjects. *IEEE Transactions on Learning Technologies*.

DENG J., DONG W., SOCHER R. *et al.* (2009). Imagenet : A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*.

GALA N., TACK A., JAVOUREY-DREVET L. *et al.* (2020). Alector : A parallel corpus of simplified french texts with alignments of misreadings by poor and dyslexic readers. In *12th Language Resources and Evaluation for Language Technologies*.

GERALD T., ETTAYEB S., LE H. Q., VILNAT A., PAROUBEK P. & ILLOUZ G. (2022). An annotated corpus for abstractive question generation and extractive answer for education. In *Conférence sur le Traitement Automatique des Langues Naturelles*.

GREEN C. (2019). A multilevel description of textbook linguistic complexity across disciplines : Leveraging NLP to support disciplinary literacy. *Linguistics and Education*.

HE K., ZHANG X., REN S. & SUN J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*.

HUANG L., WANG W., CHEN J. & WEI X.-Y. (2019). Attention on attention for image captioning. In *IEEE/CVF international conference on computer vision*.

KENTON J. D. M.-W. C. & TOUTANOVA L. K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, p. 4171–4186.

KLEIN G., KIM Y., DENG Y. *et al.* (2017). OpenNMT : Open-source toolkit for neural machine translation. In *Association for Computational Linguistics - System Demonstrations*.

LÉTÉ B., SPRENGER-CHAROLLES L. & COLÉ P. (2004). MANULEX : A grade-level lexical database from french elementary school readers. *Behavior Research Methods, Instruments, & Computers*.

LIN T.-Y., MAIRE M., BELONGIE S. *et al.* (2014). Microsoft coco : Common objects in context. In *European Conference in Computer Vision*.

LINCKER É., GUINAUDEAU C., PONS O. *et al.* (2023a). Noisy and unbalanced multimodal document classification : Textbook exercises as a use case. In *20th International Conference on Content-based Multimedia Indexing*.

LINCKER E., PONS O., GUINAUDEAU C., BARBET I., DUPIRE J., HUDELOT C., MOUSSEAU V. & HURON C. (2023b). Layout-and activity-based textbook modeling for automatic pdf textbook extraction. In *Intelligent Textbooks 2023*.

LUCY L., DEMSZKY D., BROMLEY P. & JURAFSKY D. (2020). Content analysis of textbooks via natural language processing : Findings on gender, race, and ethnicity in texas US history textbooks. *AERA Open*.

MALAKAN Z. M., ANWAR S., HASSAN G. M. & MIAN A. (2023). Sequential vision to language as story : A storytelling dataset and benchmarking. *IEEE Access*.

MARTIN L., MULLER B., SUÁREZ P. J. O. *et al.* (2020). Camembert : a tasty french language model. In *Annual Meeting of the Association for Computational Linguistics*.

OTTO C., SPRINGSTEIN M., ANAND A. & EWERTH R. (2020). Characterization and classification of semantic image-text relations. *International Journal of Multimedia Information Retrieval*.

QU L., LIU M., WU J., GAO Z. & NIE L. (2021). Dynamic modality interaction modeling for image-text retrieval. In *44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

RADFORD A., KIM J. W., HALLACY C. *et al.* (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*.

RAO J., WANG F., DING L. *et al.* (2022). Where does the performance improvement come from ? -a reproducibility concern about image-text retrieval. In *ACM SIGIR Conference on Research and Development in Information Retrieval*.

RIBEIRO M. T., SINGH S. & GUESTRIN C. (2016). "why should I trust you ?" : Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining*.

SIMONYAN K. & ZISSERMAN A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)* : Computational and Biological Learning Society.

SZEGEDY C., VANHOUCKE V., IOFFE S., SHLENS J. & WOJNA Z. (2016). Rethinking the inception architecture for computer vision. In *IEEE conference on computer vision and pattern recognition*.

YOUNG P., LAI A., HODOSH M. & HOCKENMAIER J. (2014). From image descriptions to visual denotations : New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*.