



HAL
open science

Améliorer l'Accessibilité des Sites Web grâce à l'Intelligence Artificielle : Focus sur la Génération de Descriptions d'Images

Linda Stefany Acosta Salgado, Jean-David Daviet, Lisa Jeanson

► To cite this version:

Linda Stefany Acosta Salgado, Jean-David Daviet, Lisa Jeanson. Améliorer l'Accessibilité des Sites Web grâce à l'Intelligence Artificielle : Focus sur la Génération de Descriptions d'Images. 2024. hal-04487261

HAL Id: hal-04487261

<https://hal.science/hal-04487261>

Preprint submitted on 3 Mar 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Améliorer l'Accessibilité des Sites Web grâce à l'Intelligence Artificielle : Focus sur la Génération de Descriptions d'Images

Improving Web Accessibility through Artificial Intelligence: A Focus on Image Description Generation

Linda, Acosta-Salgado,

Adeliom, linda.acosta@adeliom.com

Jean-David, Daviet

Adeliom, jean-david.daviet@adeliom.com

Lisa, Jeanson

Adeliom, lisa.jeanson@adeliom.com

The accessibility of websites has become imperative in an increasingly connected world. However, many websites remain inaccessible to certain audiences, particularly the visually impaired. One often overlooked aspect is the description of non-textual elements, especially images. This study explores how artificial intelligence can be utilized to enhance image accessibility on websites. Initially, we identified generative models for alternative image text descriptions and explored the last version of ChatGPT-4V. These models were tested on fifty images, and the generated descriptions were evaluated by a web content generation expert. The results of this experiment emphasize the importance of these models in reducing the time required for description generation. However, we observed that the obtained results did not meet the quality requirements for descriptive texts that could render images more accessible. Regarding Chat GPT-4V, the descriptions are of better quality, though its current efficiency is hampered by its slowness.

CCS CONCEPTS • Human-centered computing • Accessibility • Empirical studies in accessibility

Additional Keywords and Phrases: Web accessibility, image description, alternative text, generative models

L'accessibilité des sites Web est devenue un impératif dans un monde de plus en plus connecté. Cependant, de nombreux sites web restent inaccessibles à certains publics, et en particulier aux mal voyants. Un aspect souvent négligé est la description des éléments non-textuels, notamment des images. Cette étude explore comment l'intelligence artificielle peut être utilisée pour améliorer l'accessibilité des images sur un site web. Tout d'abord, nous avons identifié des modèles génératifs de textes alternatifs aux images et nous avons exploré la dernière version de ChatGPT-4 Vision. Ces modèles ont été testés sur cinquante images et les descriptions générées ont été évaluées par un expert en production de contenus pour les sites web. Les résultats de cette expérimentation soulignent l'importance de ces modèles pour réduire le temps nécessaire à la génération de descriptions. En revanche, nous avons constaté que les résultats obtenus ne répondent pas aux exigences de qualité de textes descriptifs susceptibles de rendre les images plus accessibles. Concernant ChatGPT-4V, les descriptions sont de meilleure qualité, bien que son efficacité actuelle soit limitée par sa lenteur.

Mots-clés additionnels : Accessibilité du Web, description d'image, texte alternatif, modèles génératifs,

Reference:

Linda Acosta-Salgado, Jean-David Daviet et Lisa Jeanson. 2024. Améliorer l'Accessibilité des Sites Web grâce à l'Intelligence Artificielle : Focus sur la Génération de Descriptions d'Images.

This is the author's draft version of the work. It is posted here for your personal use. Not for distribution. The definitive version of the record is published in IHM '24: Adjunct Proceedings of the 35th Conference on l'Interaction Humain-Machine.

1 INTRODUCTION

L'avènement de l'ère numérique est marqué par une croissance exponentielle du nombre de sites web étant utilisés par des utilisateurs de plus en plus nombreux et aux profils tout aussi variés. En effet, le nombre total d'utilisateurs de l'internet dans le monde a augmenté de 1,9 % au cours des 12 derniers mois, passant de 4,95 milliards en 2022 [22] à 5,16 milliards au début de l'année 2023 [23]. Plus qu'un simple outil, le Web est devenu indispensable pour mener à bien de nombreuses activités dans tous les domaines tant professionnels que personnels [25]. C'est pourquoi l'accès aux sites web est aujourd'hui, plus que jamais un enjeu de taille pour les concepteurs et les prescripteurs.

L'accessibilité web peut être définie comme le principe fondamental garantissant un accès universel au contenu en ligne, indépendamment des spécificités matérielles ou logicielles, de la diversité linguistique, culturelle, ou des capacités

physiques et mentales des utilisateurs [36,38]. Rendre une interface plus accessible consiste à mettre à disposition l'ensemble des informations et des services proposés par les sites web à un public aussi vaste que possible, favorisant ainsi une contribution égale et sans obstacles sur le web [14]. Les enjeux de l'accessibilité web sont multiples. Tout d'abord, il s'agit de favoriser l'autonomie en permettant aux individus en situation de handicap de naviguer sur les sites web et d'utiliser les solutions numériques sans assistance extérieure. L'accessibilité participe donc à garantir l'égalité d'accès à l'information présente sur le web et à participer à l'inclusion numérique. L'accessibilité profite également à l'ensemble des utilisateurs comme le soulignent Aizpurua et al. (2016) en améliorant la convivialité et la qualité générale des interfaces numériques [3]. Enfin, elle ouvre de nouvelles opportunités aux entreprises dont la présence sur le web se trouve renforcée et l'audience diversifiée. Ainsi, l'accessibilité web s'impose comme un impératif tant éthique que stratégique dans un monde de plus en plus connecté. Pourtant, et malgré l'existence de normes telles que les Web Content Accessibility Guidelines (WCAG) [44] et le Référentiel Général d'Accessibilité pour les Administrations (RGAA) [15], ainsi que des outils visant à évaluer l'accessibilité des sites Web (Stark Lab, Wave Accessibility Insight), de nombreux sites restent partiellement ou totalement inaccessibles pour certaines catégories d'utilisateurs [1,41]. Un récent rapport de WebAIM (Web Accessibility In Mind) [46], a révélé que 96,3 % des 1 million de sites évalués présentaient des erreurs d'accessibilité selon les critères d'accessibilité des WCAG. Bien que le nombre d'erreurs ait diminué de 1,5% par rapport aux années précédentes, ces chiffres restent alarmants.

Parmi les erreurs d'accessibilité les plus courantes figure l'absence de texte alternatif appelé également "label" associé à chaque image dont le nombre a augmenté de façon surprenante de 9,5 % en 2023 [46]. Or, l'étude d'Acosta-Vargas et son équipe montre que les problèmes d'accessibilité les plus fréquents sur les sites web universitaires sont liés à l'absence de textes alternatifs aux images [2]. En effet, les images jouent un rôle essentiel dans la communication visuelle sur les sites web. Alors que par le passé les informations présentées étaient essentiellement textuelles, aujourd'hui, une multitude d'acteurs, tels que les institutions publiques, les entreprises et les individus, ont recours à des supports visuels variés, notamment des images et des vidéos, pour la diffusion de leurs contenus sur le web [12]. Ainsi, les personnes atteintes de déficiences visuelles, qui dépendent principalement de l'ouïe pour comprendre le contenu en ligne, ont de plus en plus de difficultés à accéder à ces informations visuelles, d'autant qu'elles sont rarement accompagnées de descriptions appropriées [9,15,44]. Cependant, l'inaccessibilité des images ne résulte pas systématiquement de l'absence de descriptions, mais aussi parfois du manque de clarté et/ou d'exactitude de ces descriptions. En effet, celles-ci ne permettent pas toujours aux personnes malvoyantes, qui s'appuient sur des lecteurs d'écran, de comprendre le contenu affiché [4,41]. Les travaux de Shrestha et son équipe ont par exemple établi que la plupart des descriptions d'images fournies ont tendance à être génériques, inadéquates et souvent peu fiables, ce qui les rend inefficaces [40].

Dans ce contexte, les récents progrès des technologies de l'intelligence artificielle (IA), en particulier dans le domaine du traitement du langage naturel (NLP), offrent des opportunités prometteuses. Les modèles linguistiques de type GPT (Generative Pre-trained Transformer), notamment ChatGPT [29], se distinguent par leurs applications directes dans le domaine du traitement du langage naturel. Cette étude, menée par les équipes d'Adeliom, une agence spécialisée dans la conception de solutions digitales visant à améliorer l'expérience utilisateur, explore la manière dont les modèles linguistiques, en particulier ChatGPT-4 Vision, peuvent être exploités pour améliorer l'accessibilité des images lors de la création d'un site web. Plus précisément, nous nous concentrons sur la génération automatique de textes alternatifs, visant à décrire efficacement et systématiquement le contenu visuel des sites web. L'état de l'art, le déroulement, les résultats de cette étude ainsi que leurs implications sont présentés dans cet article.

2 ÉTAT DE L'ART

2.1 L'accessibilité des contenus visuels

Afin d'améliorer les descriptions de contenus visuels, il est crucial de s'intéresser à la manière dont ces dernières sont générées. Les descriptions des images peuvent être élaborées par des rédacteurs humains ou de manière automatique par des modèles de langage. On peut également adopter une approche hybride, combinant la génération automatique et la validation humaine. La génération manuelle de ces descriptions est une tâche exigeante et chronophage, qui nécessite des efforts importants, ce qui peut altérer leur qualité [25]. De plus, Edwards et ses collègues ont étudié la complexité du processus de création manuelle de textes alternatifs. Ils ont conclu que la qualité des descriptions peut être influencée par l'expérience personnelle et le niveau d'expertise des rédacteurs, ainsi que par les contraintes logistiques, sociales et techniques du contexte dans lequel ils travaillent [16].

Les descriptions d'images hybrides ont démontré leur capacité à améliorer la qualité des descriptions, comme le mettent en évidence les travaux de Mangiatordi et Lazzari. Ces derniers ont généré des descriptions d'images intégrées

dans des sites web à l'aide de l'IA, puis au crowdsourcing pour affiner ces descriptions [30]. Bernardi et ses collègues ont démontré que l'évaluation des descriptions par des êtres humains est plus fiable que les méthodes d'évaluation automatique [6]. Cependant, il est important de souligner que cette approche hybride comporte certaines limitations. En effet, il s'agit d'une méthodologie chronophage, qui comporte des coûts additionnels, et nécessite une organisation spécifique pour mobiliser la contribution humaine [30].

Compte tenu du nombre de facteurs altérant la description des images et de l'influence du contexte, il est nécessaire de considérer des solutions automatiques capables de produire systématiquement des descriptions d'images de qualité.

2.2 L'apport des modèles de langage pour générer des descriptions de contenus

Un certain nombre d'outils de description automatique des images ont été développés ces dernières années [13,16,17]. Parmi ces outils, certains ont été testés pour générer des descriptions d'images sur les sites web [21,43]. Ces outils comportent des modèles visant à identifier les images non labellisées avant de générer des descriptions appropriées. Néanmoins, les recherches menées par Leotta et son équipe ont révélé que ces outils ont l'inconvénient de consommer beaucoup d'énergie. En effet, la génération de descriptions des images non labellisées nécessite des appels répétés à des API à distance, ce qui est très énergivore [25]. Notons que cette limite pourrait être dépassée en effectuant un seul appel, côté serveur, au moment de la création du site web.

Peu d'études ont été réalisées pour évaluer la qualité des descriptions générées à l'aide de ces outils et leur applicabilité dans la conception de sites web [8,34,48]. Leotta et son équipe ont cependant évalué l'exactitude des descriptions générées par quatre outils disponibles sur le marché en les comparant à des descriptions générées manuellement sur les mêmes images. Leurs résultats démontrent que les descriptions générées par ces outils sont jugées moins exactes que celles créées par des rédacteurs humains [25]. De plus, les auteurs recommandent d'intégrer les descriptions d'images dès la phase de développement de sites web. Toutefois, les outils examinés par les auteurs ne sont pas tous adaptés à cette étape de développement, et leur utilisation coûtent souvent trop chers aux développeurs dont les budgets de production sont restreints.

D'autres modèles sont plus accessibles et reconnus pour leur performance, leur reproductibilité et leur popularité ont été identifiées et évaluées par Bianco et son équipe. Les auteurs ont entrepris d'identifier les deux meilleures descriptions et de les fusionner à l'aide d'un modèle de langage. Le résultat de cette fusion génère une description plus complète, mais similaire aux modèles qui l'intègrent [8]. Notons que s'il existe des erreurs dans la description, elles sont reprises dans la description générée par la fusion.

Les avancées récentes dans le domaine du traitement du langage naturel (NLP) ont donné naissance à des modèles de langage particulièrement puissants, parmi lesquels se distingue ChatGPT [29]. Ce modèle suscite un intérêt significatif et croissant principalement en raison de ses applications directes dans le domaine du traitement du langage naturel. Il se distingue par sa capacité à fournir une interface linguistique dotée de compétences conversationnelles et de capacités de raisonnement couvrant un large éventail de domaines d'application [29,47]. L'efficacité de ChatGPT a déjà été explorée dans la rédaction de contenu textuel, comme le soulignent les travaux de Lentz & Mager [24]. ChatGPT a notamment permis d'améliorer la vitesse, la cohérence, la personnalisation et la qualité globale du contenu créé. D'autres recherches ont été menées pour évaluer l'utilité de ChatGPT dans la transformation de textes en images [5]. Certaines études ont même tenté d'intégrer des modèles d'analyse d'images à ces nouveaux modèles de langage [37,50]. Récemment, une quatrième version du ChatGPT a été rendue accessible, comportant une fonctionnalité permettant l'analyse d'images : ChatGPT-4V (ChatGPT-4 Vision).

Toutefois, à ce jour, aucune étude n'a proposé d'évaluer la capacité de ces modèles de langage à décrire les éléments composant une image en vue de les rendre accessibles. En outre, nous n'avons pas trouvé de recherche proposant de comparer de manière systématique les descriptions générées par ces modèles de langage aux descriptions fournies par d'autres modèles. Une telle étude nous paraît pourtant essentielle pour comprendre les avantages et les limites de ces nouveaux outils dans la génération automatique de descriptions d'images, et donc dans la conception de sites web et de solutions numériques plus accessibles.

3 MÉTHODOLOGIE

Dans cette étude, une démarche structurée en plusieurs étapes a été mise en place (Annexe A1).

3.1 Choix des modèles

La première étape a consisté à sélectionner les modèles de traitement de langage naturel appropriés pour cette étude sur la plateforme HuggingFace. Nous avons sélectionné les modèles selon les critères suivants : la disponibilité de leur code source et leur notoriété au sein de la communauté de recherche en apprentissage profond. De plus, ces modèles fournissent des exemples et des explications détaillées, simplifiant ainsi leur mise en œuvre et leur réutilisation.

Les trois modèles retenus sont les suivants :

- Blip : Bootstrapping Language-Image Pre-training [27] : ce modèle présente l'avantage d'une mise en place relativement simple, facilitant la création d'interfaces graphiques pour la conversion d'images en texte. Il a été sélectionné pour son potentiel en termes d'accessibilité et de facilité d'utilisation.
- GIT : Generative Image-to-text Transformer [45] : bien que le modèle GIT soit de taille réduite, il offre une implémentation aisée. Il représente un candidat prometteur pour notre étude et mérite d'être évalué en comparaison avec Blip.
- ViT-GPT2 [35] : une combinaison d'un encodeur d'image (ViT) et d'un décodeur de langage : GPT2 d'OpenAI.

Ainsi, les trois modèles open source précédents ont été comparés à la dernière version de ChatGPT, appelée ChatGPT-4 Vision (ChatGPT-4V), en raison de sa capacité à traiter et analyser des données visuelles. Le choix de ChatGPT-4V dans notre étude se justifie par sa capacité à générer des descriptions d'images.

3.2 Choix des Images

Dans le but d'évaluer la pertinence de ces modèles pour la description d'images sur un site web donné, nous avons pris comme exemple un site développé par l'agence Adeliom pour l'un de ses clients. Ce site a été choisi, car il est passé de 5,000 à 15,000 images lors de sa dernière refonte, avec 1,500 pages contenant chacune plusieurs images. Parmi l'actuelle collection de 15,000 images, nous avons choisi un échantillon de 50 images de manière aléatoire. Cette approche nous a permis de couvrir un échantillon diversifié de contenus visuels présents sur le site.

3.3 Description des Images

Chacune des 50 images sélectionnées a été décrite en utilisant les différents modèles. Pour les modèles open source, le code source associé à ces modèles a été utilisé pour générer les descriptions.

3.3.1 Mise en place du code

Pour chacun des 3 modèles trouvés via la plateforme HuggingFace, nous avons mis en place un code utilisant le langage Python. Le choix de ce langage était sans équivoque pour trois raisons. Tout d'abord, Python est largement reconnu comme la référence pour l'implémentation de scripts liés à l'intelligence artificielle, offrant une multitude de bibliothèques et de ressources adaptées à ce domaine. De plus, sa simplicité d'apprentissage et d'utilisation en font un outil utilisable y compris par des personnes qui ne possèdent pas une connaissance approfondie du langage. Enfin, Python permet la création aisée de serveurs web, ce qui était requis pour nos tests.

Nous avons utilisé des fragments de code disponibles sur la plateforme Huggingface pour nous aider à mettre en place ces modèles [42]. Nous avons dû les adapter à nos besoins spécifiques et installer les dépendances nécessaires au bon fonctionnement des modèles (annexe A2, A3 et A4).

3.3.2 Test du code

Après avoir personnalisé le code, nous avons effectué des tests manuels pour nous assurer que les descriptions étaient générées correctement. Ensuite, nous avons intégré le code dans un serveur web local, choisi dans le but de simplifier la comparaison des temps de génération entre les modèles (annexe A5).

3.3.3 Génération des descriptions

La mise en place de ce serveur web local via Python nous a permis très rapidement de générer des descriptions d'images et de faire des tests automatisés. Nous avons ainsi créé un serveur web simple capable de récupérer une image via une requête HTTP et de générer un label pour cette image en utilisant le modèle choisi au préalable. En entrée, nous fournissons une image, et en sortie, nous obtenons une description. En annexe A6, un graphique explique le fonctionnement général de ces modèles pour la transformation d'une image en texte.

3.3.4 Traduction des descriptions

Il est à noter que les modèles utilisés génèrent des libellés en langue anglaise, ce qui a nécessité l'intégration d'une étape de traduction intermédiaire après la récupération de la description, pour obtenir des labels en français. Pour ce faire, nous avons opté pour l'utilisation du modèle "opus-mt-en-fr" [19]. Ce choix s'est justifié par le fait que ce modèle est le plus téléchargé et le plus utilisé, et qu'il est directement répertorié sur la première page des modèles de traduction sur la plateforme Huggingface (Annexe A7). Ensuite, nous avons mis en place un script permettant d'envoyer les 50 images de test au serveur et d'en récupérer la description. Nous avons relevé et comparé les temps de génération de chaque description pour chacun des modèles utilisés.

Pour ChatGPT-4V, n'ayant pas accès à l'API, nous avons dû envoyer manuellement les 50 images en utilisant l'interface de ChatGPT depuis leur site <https://chat.openai.com> grâce au prompt suivant :

“Tu es un assistant pour la création de contenu sur un site spécialisé dans la vente de fenêtres, portes, garages, pergolas, vérandas. Voici une image utilisée sur le site. Décris-moi ce qu'elle représente afin que je puisse insérer cette description dans une balise alt. Elle doit donc décrire l'image et être brève, c'est-à-dire une dizaine de mots maximum. Réponds-moi seulement avec le texte à insérer. Décris-moi l'image en faisant une phrase.”

Puis, après chaque envoi d'image :

“Décris-moi l'image en faisant une phrase.”

À chaque étape, nous avons relevé le temps de génération au travers du temps de réponse au prompt accessible dans l'outil de développement du navigateur web.

4 ÉVALUATION DES DESCRIPTIONS

Pour chacune des images sélectionnées, quatre descriptions ont donc été générées à l'aide des modèles choisis. Un exemple des descriptions obtenues pour une image est présenté dans l'annexe A8.

Les descriptions générées ont été évaluées en prenant en compte deux critères : le temps de génération et la qualité descriptive du label. L'évaluation du temps de génération a été obtenue en mesurant la durée d'exécution du code pour chaque modèle, ce qui est essentiel pour évaluer l'efficacité opérationnelle des modèles [39].

Afin d'évaluer la qualité des descriptions, nous avons privilégié l'avis d'experts, lequel est, rappelons-le, plus pertinent pour saisir des problèmes d'accessibilité plus subtils [25]. Nous avons conçu une grille d'évaluation composée d'échelles de Likert de cinq niveaux d'appréciation (tout à fait d'accord, d'accord, ni d'accord ni pas d'accord, pas d'accord, pas du tout d'accord) pour chacun des 6 critères d'évaluation des labels identifiés par Bernardi et son équipe [6] :

- **Description correcte** : une description correcte transmet de manière précise et objective le contenu de l'image en termes d'objets, d'actions et de scènes [49].
- **Grammaire correcte** : une description grammaticalement correcte est exempte d'erreurs grammaticales, offrant ainsi une structure linguistique précise et compréhensible [26].
- **Information correcte** : une description qualifiée d'information correcte se distingue par son aptitude à exclure toute information superflue ou erronée [33].
- **Description pertinente** : une description pertinente est une représentation qui s'ajuste de manière à l'image, correspondant significativement à son contenu [26].
- **Description créative** : une description créative se caractérise par l'utilisation surprenante de mots, en particulier de verbes inhabituels, et par la capacité à présenter les éléments de manière poétique, ajoutant ainsi une dimension artistique à la description [26].
- **Description rédigée par l'humain** : une description qualifiée comme étant écrite par des humains évoque une narration qui semble avoir été créée par une personne, déployant une personnalité propre, plutôt que générée automatiquement ou rédigée de manière impersonnelle [33].

L'évaluation qualitative a été réalisée en sollicitant l'expertise d'un rédacteur de contenu UX chez Adeliom, spécialiste dans le domaine. À cet effet, nous lui avons fourni la grille d'évaluation contenant les liens vers les images sélectionnées et les différentes descriptions générées, sans révéler quel modèle avait généré chacune d'entre elles (Annexe A9). Les réponses obtenues ont été converties de l'échelle de Likert en scores (1 = pas du tout d'accord, 5 = tout à fait d'accord), avant d'être intégrées dans une feuille de calcul dédiée à l'analyse systématique et objective des données recueillies.

5 RÉSULTATS

Les valeurs moyennes pour chaque modèle au regard des six critères ont été calculées. La répartition des évaluations est présentée ci-dessous :

Table 1: Moyenne et écart type des résultats

		Description correcte	Grammaire correcte	Information correcte	Description pertinente	Description créative	Rédigée par l'humain	Temps (secondes)
Blip	μ	3,26	3,00	3,00	2,42	1,08	1,32	15,33
	σ	1,68	1,46	1,76	1,43	0,27	1,10	0,74
GIT	μ	2,92	4,32	3,00	2,16	1,00	1,90	18,33
	σ	1,69	1,24	1,76	1,20	0,00	1,11	5,40
VitGPT2	μ	1,00	3,46	1,00	1,00	1,00	1,00	16,83
	σ	0,00	1,84	0,00	0,00	0,00	0,00	1,50
ChatGPT-4V	μ	3,84	4,32	4,08	3,82	2,58	2,74	19,38
	σ	0,68	1,17	0,85	0,69	0,64	1,01	3,54

- **BLIP** : Blip se distingue par une description correcte des images avec une moyenne de 3,26, ce qui montre qu'il a une bonne capacité de base pour la tâche. Cependant, il présente une dispersion notable dans ses performances, comme le suggère son écart type relativement élevé. C'est également le modèle qui génère des labels le plus rapidement (15,33 secondes en moyenne). Cependant, lorsqu'il s'agit de créativité et de caractère humain dans les descriptions, le modèle présente des lacunes, comme en témoignent les moyennes de 1,08 et 1,32 respectivement. Cela suggère que les descriptions générées peuvent souvent sembler génériques ou robotiques.
- **GIT** : GIT est particulièrement efficace en termes de grammaire, avec une moyenne de 4,32 dans ce critère. Cependant, il présente une certaine variabilité dans ses performances, comme indiqué par son écart type. Tout comme BLIP, GIT est moins performant dans les critères de créativité et le caractère humain des descriptions. Cela indique que, bien que grammaticalement correctes, les descriptions peuvent manquer de nuance et de personnalisation. De plus, ce modèle est moins rapide que BLIP pour générer un label (18,33 secondes en moyenne).
- **VitGPT2** : Parmi les modèles évalués, VitGPT2 présente les descriptions de la plus faible qualité. Les moyennes sont les plus basses dans presque tous les critères. Son écart type bas suggère une certaine stabilité dans ses performances, bien que celles-ci soient globalement moins satisfaisantes.
- **ChatGPT-4V** : Le modèle ChatGPT-4V présente des performances remarquables, caractérisées par des évaluations élevées dans la majorité des critères, et par des écarts types relativement faibles. Avec une valeur moyenne de la grammaire de 4,32, ce modèle peut fournir, d'après notre expert, des descriptions claires et structurées. De plus, sa capacité à fournir des informations exactes (4,08) et pertinentes (3,82) le positionne comme un outil fiable pour interpréter le contenu visuel des images. Cependant, la moyenne de son temps de génération pour 1 label est de 19,33 secondes, ce qui en fait l'outil le plus lent pour générer des labels.

De manière générale, les évaluations ont révélé que la créativité était le critère le moins bien noté parmi les modèles évalués, tandis que la grammaire a reçu les évaluations les plus favorables, démontrant la capacité des modèles à produire des descriptions linguistiquement correctes. Il est intéressant de noter qu'aucune des descriptions générées, y compris celles de ChatGPT-4V, n'a été considérée comme rédigée par un être humain. Ce résultat souligne la difficulté pour les modèles de générer des descriptions "naturelles" ou "proches du naturel" des contenus visuels.

6 DISCUSSION

Les modèles testés ont montré une capacité variable à générer des descriptions d'images de qualité, et ce, dans un laps de temps relativement court. Toutefois, il est important de noter qu'aucun modèle n'a obtenu une évaluation entièrement favorable. Cette observation met en évidence la complexité inhérente à la génération automatique de descriptions d'images. En dépit des avancées technologiques, l'IA ne parvient pas toujours à saisir de manière suffisamment précise le contexte ni les subtilités d'une image. Il est important de souligner que la pertinence d'une

description dépend fortement du contexte d'utilisation, ce qui varie selon les sites web voire les pages individuelles. La notion d'adaptation contextuelle n'a pas été pleinement explorée dans notre étude, mais elle soulève des questions importantes sur la capacité des modèles d'IA à comprendre et à intégrer les nuances contextuelles dans la génération automatique de descriptions d'images.

Les modèles open source que nous avons évalué se distinguent par leur facilité d'utilisation. Cependant, ces derniers utilisent des datasets génériques disponibles sur la plateforme HuggingFace pour générer les descriptions. Le contexte s'est révélé être un facteur significatif influençant la pertinence des descriptions générées par les modèles d'IA. Or, la nature générale et peu spécialisée de ces datasets a eu un impact négatif sur la qualité des libellés générés. Pour obtenir des résultats de meilleure qualité, il serait nécessaire d'envisager l'utilisation de jeux de données spécifiques au domaine du client, avec un entraînement du modèle sur ces données afin d'aligner les réponses sur les besoins spécifiques de ce contexte. Dans l'ensemble, les performances des modèles et des datasets génériques se sont révélées insuffisantes pour notre application, en raison de leur spécificité limitée. Toutefois, il est envisageable que ces modèles puissent donner de meilleurs résultats pour des images plus courantes et généralistes que celles que nous avons testées. L'idée de former des bases de données spécifiques au contexte mérite également d'être explorée pour améliorer la qualité des descriptions sur des images plus complexes.

Ensuite, même si ChatGPT-4V produit des descriptions de meilleure qualité, son utilisation est plus contraignante par sa lenteur et son interface. L'utilisation d'une API nous permettrait toutefois de dépasser ces problématiques. Néanmoins, le coût associé à l'utilisation de l'API pourrait s'avérer prohibitif pour de nombreuses entreprises et potentiels utilisateurs. On peut anticiper qu'à l'avenir, des solutions open source similaires seront développées et mises à disposition des communautés de développeurs et de chercheurs. Cependant, malgré les avancées des modèles d'IA, des considérations éthiques, la protection de la vie privée des données et la nécessité d'une surveillance demeurent des défis à relever, comme l'ont souligné Lentz et Mager [24].

Notons que le délai nécessaire à la génération de descriptions lors de l'utilisation d'un serveur local ne peut être directement comparé à celui des serveurs OpenAI, qui traitent un volume bien plus important de requêtes dans des centres de données aux États-Unis. Le choix du serveur local visait principalement à permettre une comparaison rapide des temps de génération entre différents modèles. Il convient de souligner que les temps de génération sont étroitement liés à la taille du modèle [10]. Ainsi, si l'adoption d'un modèle open source était envisagée pour une utilisation plus étendue que de simples tests locaux, la possibilité d'une migration vers un serveur dédié à l'utilisation de l'apprentissage automatique serait à considérer.

En ce qui concerne les limites de cette étude et les perspectives, notons que l'évaluation des descriptions a été réalisée par un seul expert. Or pour une validation plus robuste, des évaluations multiples effectuées par plusieurs experts seraient souhaitables [16]. Parallèlement, il est impératif d'inclure les utilisateurs finaux dans le processus d'évaluation des descriptions, notamment les personnes malvoyantes.

Il est important de noter que notre étude s'est orientée sur des sites Web français, avec un expert français, justifiant ainsi le recours à la traduction des descriptions. Cependant, la traduction des descriptions de l'anglais au français a introduit des incertitudes et a eu un impact sur le temps de génération des labels. Pour garantir une évaluation précise, l'évaluation de l'expert devrait être réalisée dans la langue de génération du modèle, impliquant des experts natifs de cette langue. Dans notre cas, l'utilisation de bases de données en français pourrait significativement améliorer la qualité des descriptions. Cela permettrait d'éviter les erreurs induites par l'utilisation de traducteurs et de mieux répondre aux besoins spécifiques des utilisateurs francophones.

L'évaluation manuelle des descriptions produites par les modèles s'est révélée être une tâche complexe. D'autres recherches ont mis en avant des outils d'évaluation automatique, tels que la métrique "InfoMetIC" proposée par Hu et son équipe [20], qui offre la capacité de détecter les mots incorrects et les régions d'image non mentionnées dans une description, tout en fournissant un score de qualité global du texte [39]. Parallèlement, Shrestha [40] a proposé un cadre et des modèles basés sur un réseau neuronal pour une évaluation automatique des descriptions d'images en termes de conformité aux directives NCAM [18]. Bien que nous reconnaissons que l'évaluation automatique seule ne soit pas suffisante pour évaluer la qualité d'une description, nous estimons que dans des études ultérieures, il serait judicieux d'utiliser ces outils en association avec l'expertise humaine, ce qui faciliterait considérablement la tâche d'évaluation tout en la rendant plus objective.

Par rapport aux descriptions générées par les différents modèles, la pertinence d'une description dépend souvent du contexte spécifique de l'image. Dans certains cas, les descriptions ont mis l'accent sur des détails qui ne semblaient pas être les plus pertinents par rapport au message global de l'image. Bien que grammaticalement correctes, certaines descriptions peuvent être formulées de manière lourde ou complexe, rendant la compréhension moins aisée pour

l'utilisateur final (d'autant que les labels seraient en réalité lus par un assistant vocal). Nous suggérons l'utilisation potentielle de phrases nominales (sans verbe) pour plus de simplicité et de concision.

7 CONCLUSION

L'accessibilité du web, dont l'importance a été largement reconnue, représente un défi significatif, notamment en ce qui concerne l'accessibilité des contenus non textuels, tels que les images. Cette étude a exploré la manière dont l'IA peut contribuer à pallier cette lacune. Nous avons évalué quatre modèles génératifs de descriptions d'images, en mettant l'accent sur le temps de génération et la qualité des descriptions. L'objectif était de déterminer leur applicabilité potentielle dans la phase de conception de sites web. Les résultats de cette expérimentation ont mis en lumière que les modèles open source, bien que prometteurs, pourraient nécessiter une base de données spécifiquement adaptée au contexte pour garantir des descriptions plus pertinentes et en français, répondant ainsi aux besoins d'un public francophone. Le modèle ChatGPT-4V se distingue actuellement en termes de qualité de description. Cependant, il est essentiel de noter que cette qualité supérieure s'accompagne d'un temps de génération substantiel.

Suite à cette étude, notre prochaine étape consistera à entreprendre une évaluation directe avec les utilisateurs finaux. L'objectif de cette évaluation sera de comprendre non seulement la pertinence des descriptions générées, mais surtout leur utilité en tant qu'éléments d'orientation pour permettre aux utilisateurs d'atteindre des objectifs dans un contexte spécifique.

Parallèlement, de nouvelles recherches seront engagées sur la génération automatique de descriptions pour d'autres éléments visuels d'un site, couvrant des formats émergents tels que les GIFs, les images tridimensionnelles et les micro-interactions. Des études récentes, telles que celles de Bhatnagar et Chen (2022), ont cherché à améliorer l'accessibilité des images GIF en utilisant des techniques d'apprentissage automatique [7]. De même, Marques et Ferreira ont exploré des études visant à mieux comprendre la description des images dynamiques, notamment dans le contexte de l'accessibilité pour les malvoyants [31]. Cependant, leurs recherches ont révélé un manque d'études spécifiques dédiées au développement de technologies pour la description d'images dynamiques, suggérant ainsi un besoin de recherche continue dans ce domaine.

Dans un second temps, d'autres recherches pourront se faire sur la description des contenus visuels appliquée à de nouvelles formes d'interaction. Bien que des recherches aient déjà traité de la description du contenu visuel dans les interfaces mobiles [11,28,32], peu de recherches ont été identifiées dans des contextes de réalité augmentée ou virtuelle. Les résultats d'études récentes, tels que ceux de Cho et Kim (2023) portant sur la description d'images multidimensionnelles, indiquent que l'intégration de descriptions tridimensionnelles en couches (premier plan, milieu et arrière-plan) peut significativement améliorer la compréhension spatiale des utilisateurs [12]. Cette approche pourrait servir de base pour améliorer l'accessibilité dans un contexte immersif, ouvrant ainsi de nouvelles perspectives dans le domaine de l'accessibilité numérique.

RÉFÉRENCES

- [1] Hayfa.Y. Abuaddous, Mohd Zalisham, and Nurlida Basir. 2016. Web Accessibility Challenges. *Int. J. Adv. Comput. Sci. Appl.* 7, 10 (2016). <https://doi.org/10.14569/IJACSA.2016.071023>
- [2] Patricia Acosta-Vargas, Mario González, and Sergio Luján-Mora. 2020. Dataset for evaluating the accessibility of the websites of selected Latin American universities. *Data Brief* 28, (February 2020), 105013. <https://doi.org/10.1016/j.dib.2019.105013>
- [3] Amaia Aizpurua, Simon Harper, and Markel Vigo. 2016. Exploring the relationship between web accessibility and user experience. *Int. J. Hum.-Comput. Stud.* 91, (July 2016), 13–23. <https://doi.org/10.1016/j.ijhcs.2016.03.008>
- [4] Tahani Alahmadi and Steve Drew. 2018. Evaluation of image accessibility for visually impaired users. *J. Access. Des. All* 8, 2 (November 2018), 125–160. <https://doi.org/10.17411/jaccess.v8i2.167>
- [5] Mourad Bahani, Aziza El Ouazizi, and Khalil Maalmi. 2023. The effectiveness of T5, GPT-2, and BERT on text-to-image generation task. *Pattern Recognit. Lett.* 173, (September 2023), 57–63. <https://doi.org/10.1016/j.patrec.2023.08.001>
- [6] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. *J. Artif. Intell. Res.* 55, (February 2016), 409–442. <https://doi.org/10.1613/jair.4900>
- [7] Apoorva Bhatnagar and Min Chen. 2022. Improving GIF Image Accessibility on Web. In *2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI)*, August 09, 2022, San Diego, CA, USA. IEEE Press, San Diego, CA, USA, 190–191. . <https://doi.org/10.1109/IRI54793.2022.00050>
- [8] Simone Bianco, Luigi Celona, Marco Donzella, and Paolo Napoletano. 2023. Improving Image Captioning Descriptiveness by Ranking and LLM-based Fusion. Retrieved October 25, 2023 from <http://arxiv.org/abs/2306.11593>
- [9] Jeffrey P. Bigham, Ryan S. Kaminsky, Richard E. Ladner, Oscar M. Danielsson, and Gordon L. Hempton. 2006. WebInSight:: making web images accessible. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, October 23, 2006, Portland Oregon USA. ACM, Portland Oregon USA, 181–188. . <https://doi.org/10.1145/1168987.1169018>
- [10] Filipe Calegario, Vanilson Burégio, Francisco Erivaldo, Daniel Moraes Costa Andrade, Kailane Felix, Nathalia Barbosa, Pedro Lucas da Silva Lucena, and César França. 2023. Exploring the intersection of Generative AI and Software Development. Retrieved January 3, 2024 from

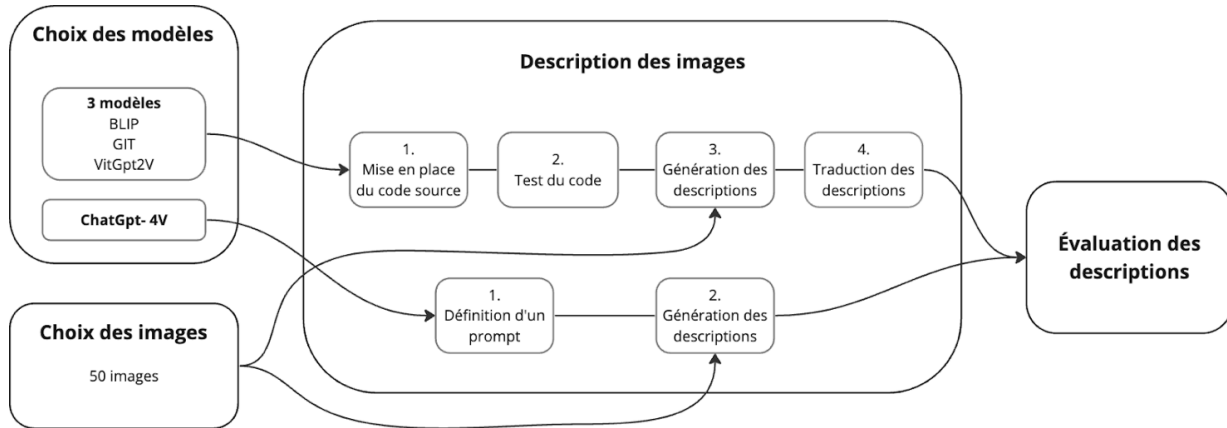
<http://arxiv.org/abs/2312.14262>

- [11] Jiешan Chen, Chunyang Chen, Zhenchang Xing, Xiwei Xu, Liming Zhu, Guoqiang Li, and Jinshui Wang. 2020. Unblind your apps: predicting natural-language labels for mobile GUI components by deep learning. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering (ICSE '20)*, October 01, 2020, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 322–334. . <https://doi.org/10.1145/3377811.3380327>
- [12] Jaemin Cho and Hee Jae Kim. 2023. Dimensional alt text: Enhancing Spatial Understanding through Dimensional Layering of Image Descriptions for Screen Reader Users. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, April 19, 2023, Hamburg Germany. ACM, Hamburg Germany, 1–6. . <https://doi.org/10.1145/3544549.3585706>
- [13] Fida Hussain Dahri, Asghar Ali Chandio, Nisar Ahmed Dahri, and Muhammad Ali Soomro. 2023. Image Caption Generator Using Convolutional Recurrent Neural Network Feature Fusion. 19, 03 (2023).
- [14] João Dias, Diana Carvalho, Tânia Rocha, and João Barroso. 2022. Automated Evaluation Tools for Web and Mobile Accessibility: proposal of a new adaptive interface tool. *Procedia Comput. Sci.* 204, (2022), 297–304. <https://doi.org/10.1016/j.procs.2022.08.036>
- [15] Direction interministérielle du numérique. 2019. Référentiel général d'amélioration de l'accessibilité (RGAA 4.1.2). Retrieved October 25, 2023 from <https://accessibilite.numerique.gouv.fr/methode/criteres-et-tests/>
- [16] Emory J. Edwards, Michael Gilbert, Emily Blank, and Stacy M. Branham. 2023. How the Alt Text Gets Made: What Roles and Processes of Alt Text Creation Can Teach Us About Inclusive Imagery. *ACM Trans. Access. Comput.* 16, 2 (June 2023), 1–28. <https://doi.org/10.1145/3587469>
- [17] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M. Kitani, and Jeffrey P. Bigham. 2020. Twitter A11y: A Browser Extension to Make Twitter Images Accessible. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, April 21, 2020, Honolulu HI USA. ACM, Honolulu HI USA, 1–12. . <https://doi.org/10.1145/3313831.3376728>
- [18] Bryan Gould, Trisha O'Connell, and Geoff Freed. 2008. Effective Practices for Description of Science Content within Digital Talking Books. Retrieved October 27, 2023 from http://ncamftp.wgbh.org/ncam-old-site/experience_learn/educational_media/stemdx.html
- [19] Helsinki-NLP. 2020. opus-ml-en-fr. Retrieved October 26, 2023 from <https://huggingface.co/Helsinki-NLP/opus-ml-en-fr>
- [20] Anwen Hu, Shizhe Chen, Liang Zhang, and Qin Jin. 2023. InfoMetIC: An Informative Metric for Reference-free Image Caption Evaluation. Retrieved October 23, 2023 from <http://arxiv.org/abs/2305.06002>
- [21] Hyeonhak Jeong, Minki Chun, Hyunmin Lee, Seung Young Oh, and Hyunggu Jung. 2023. WATAA: Web Alternative Text Authoring Assistant for Improving Web Content Accessibility. In *28th International Conference on Intelligent User Interfaces*, March 27, 2023, Sydney NSW Australia. ACM, Sydney NSW Australia, 41–45. . <https://doi.org/10.1145/3581754.3584127>
- [22] Simon Kemp. 2022. *Digital 2022: Global Overview Report*. Retrieved October 26, 2023 from <https://datareportal.com/reports/digital-2022-global-overview-report>
- [23] Simon Kemp. 2023. *Digital 2023: Global Overview Report*. Meltwater and We Are Social. Retrieved October 9, 2023 from <https://datareportal.com/reports/digital-2023-global-overview-report>
- [24] Amanda Lentez and Gabriela Mager. 2023. Exploring the Potential of ChatGPT in Enhancing User Experience (UX) Writing. In *Human Interaction & Emerging Technologies (IHET 2023): Artificial Intelligence & Future Applications*, 2023. AHFE Open Acces. . <https://doi.org/10.54941/ahfe1004030>
- [25] Maurizio Leotta, Fabrizio Mori, and Marina Ribauda. 2022. Evaluating the effectiveness of automatic image captioning for web accessibility. *Univers. Access Inf. Soc.* (August 2022). <https://doi.org/10.1007/s10209-022-00906-7>
- [26] C. Li, B. Cheng, J. Chen, P. Gu, N. Deng, and D. Li. 2011. A web service performance evaluation approach based on users experience. In *Proc. -IEEE Int. Conf. Web Serv., ICWS*, 2011. 734–735. . <https://doi.org/10.1109/ICWS.2011.29>
- [27] Junnan Li, Dongxu Li, Caimeing Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. Retrieved October 18, 2023 from <http://arxiv.org/abs/2201.12086>
- [28] Yang Li, Gang Li, Luheng He, Jingjie Zheng, Hong Li, and Zhiwei Guan. 2020. Widget Captioning: Generating Natural Language Description for Mobile User Interface Elements. Retrieved October 25, 2023 from <http://arxiv.org/abs/2010.04295>
- [29] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuan Yuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. 2023. Summary of ChatGPT-Related Research and Perspective Towards the Future of Large Language Models. *Meta-Radiol.* 1, 2 (September 2023), 100017. <https://doi.org/10.1016/j.metrad.2023.100017>
- [30] Andrea Mangiatordi and Marco Lazzari. 2018. Combined use of artificial intelligence and crowdsourcing to provide alternative content for images on websites. In *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, January 2018. 1–6. . <https://doi.org/10.1109/CCNC.2018.8319312>
- [31] João Marcelo Dos Santos Marques and Simone Bacellar Leal Ferreira. 2023. Description of dynamic images: a systematic mapping for the understanding and transparency of information that may be perceptible visually impaired citizens. In *Proceedings of the XIX Brazilian Symposium on Information Systems*, May 29, 2023, Maceió Brazil. ACM, Maceió Brazil, 229–236. . <https://doi.org/10.1145/3592813.3592910>
- [32] Forough Mehralian, Navid Salehnamadi, and Sam Malek. 2021. Data-driven accessibility repair revisited: on the effectiveness of generating labels for icons in Android apps. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, August 20, 2021, Athens Greece. ACM, Athens Greece, 107–118. . <https://doi.org/10.1145/3468264.3468604>
- [33] Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daume Iii. 2012. Midge: Generating Image Descriptions From Computer Vision Detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, Avignon, France. Avignon, France, 747–756. . <https://doi.org/10.5555/2380816.2380907>
- [34] Ab Shaqoor Nengroo and K. S. Kuppasamy. 2018. Accessible images (AIMS): a model to build self-describing images for assisting screen reader users. *Univers. Access Inf. Soc.* 17, 3 (August 2018), 607–619. <https://doi.org/10.1007/s10209-017-0607-z>
- [35] nlpconnect. 2022. vit-gpt2-image-captioning. *Hugging Face*. Retrieved October 18, 2023 from <https://huggingface.co/nlpconnect/vit-gpt2-image-captioning>
- [36] Lucas Prado Osco, Eduardo Lopes de Lemos, Wesley Nunes Gonçalves, Ana Paula Marques Ramos, and José Marcato Junior. 2023. The Potential of Visual ChatGPT for Remote Sensing. *Remote Sens.* 15, 13 (January 2023), 3232. <https://doi.org/10.3390/rs15133232>
- [37] Rohan Patankar, Harshik Sethi, Aditi Sadhukha, Nishant Banjade, and Ashwini Mathur. 2023. Image Captioning with Audio Reinforcement using RNN and CNN. In *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, June 2023. 591–596. . <https://doi.org/10.1109/ICSCSS57650.2023.10169692>
- [38] Helen Petrie, Andreas Savva, and Christopher Power. 2015. Towards a unified definition of web accessibility. In *Proceedings of the 12th International Web for All Conference*, May 18, 2015, Florence Italy. ACM, Florence Italy, 1–13. . <https://doi.org/10.1145/2745555.2746653>
- [39] S. Schwartz. 2019. Transformational content and relationships: Research, analytical tools, and big data in shaping the News User Experience (UX). In *Scholarly Ethics and Publishing: Breakthroughs in Res. and Pract.* IGI Global, 791–809. Retrieved from

- <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85130653602&doi=10.4018%2f978-1-5225-8057-7.ch039&partnerID=40&md5=6323f877abeed4826b0ab4af39f1883>
- [40] Raju Shrestha. 2022. A transformer-based deep learning model for evaluation of accessibility of image descriptions. In *2022 14th International Conference on Machine Learning and Computing (ICMLC)*, February 18, 2022, Guangzhou China. ACM, Guangzhou China, 28–33. .
<https://doi.org/10.1145/3529836.3529856>
- [41] Shivam Singh, Anurag Bhandari, and Nishiith Pathak. 2018. *Accessify: An ML Powered Application to Provide Accessible Images on Web Sites*.
<https://doi.org/10.1145/3192714.3192830>
- [42] Perla Suresh and Johannes Kolbe. 2023. What is Image-to-Text? - Hugging Face. *Hugging Face*. Retrieved December 12, 2023 from <https://huggingface.co/tasks/image-to-text>
- [43] Tejal Tiwary and Rajendra Mahapatra. 2023. Enhancement in web accessibility for visually impaired people using hybrid deep belief network –bald eagle search. *Multimed. Tools Appl.* 82, (February 2023), 1–22. <https://doi.org/10.1007/s11042-023-14494-y>
- [44] W3C. 2023. Web Content Accessibility Guidelines (WCAG) 2.1. Retrieved October 25, 2023 from <https://www.w3.org/TR/WCAG21/>
- [45] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. GIT: A Generative Image-to-text Transformer for Vision and Language. Retrieved October 18, 2023 from <http://arxiv.org/abs/2205.14100>
- [46] WebAIM. 2023. *The WebAIM Million. The 2023 report on the accessibility of the top 1,000,000 home pages*. Retrieved October 11, 2023 from <https://webaim.org/projects/million/>
- [47] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. <https://doi.org/10.48550/arXiv.2303.04671>
- [48] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, February 25, 2017, Portland Oregon USA. ACM, Portland Oregon USA, 1180–1192. . <https://doi.org/10.1145/2998181.2998364>
- [49] Yezhou Yang, Ching Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-Guided Sentence Generation of Natural Images. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, July 2011, Edinburgh, Scotland, UK. Association for Computational Linguistics, Edinburgh, Scotland, UK, 444–454. . Retrieved January 19, 2024 from <https://aclanthology.org/D11-1041>
- [50] Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. 2023. ChatGPT Asks, BLIP-2 Answers: Automatic Questioning Towards Enriched Visual Descriptions. <https://doi.org/10.48550/arXiv.2303.06594>

À APPENDICES

A.1 DESCRIPTION DE LA DÉMARCHE MISE EN PLACE



A.2 SCRIPT BLIP

```
from cgi import FieldStorage
from PIL import Image
from transformers import BlipProcessor, BlipForConditionalGeneration
class Labeliser:
    def generate(
        uploaded_file: FieldStorage,
        model: str = "Salesforce/blip-image-captioning-large",
    ) -> str:
        processor = BlipProcessor.from_pretrained(model)
        model = BlipForConditionalGeneration.from_pretrained(model)
        raw_image = Image.open(uploaded_file).convert("RGB")
        # unconditional image captioning
        inputs = processor(raw_image, return_tensors="pt")
        out = model.generate(**inputs)
        response = processor.decode(out[0], skip_special_tokens=True)
        print(response)
        return response
```

A.3 SCRIPT GIT

```
from cgi import FieldStorage
from PIL import Image
from transformers import GitProcessor, GitForCausalLM
import torch
class Labeliser:
    def generate(
        uploaded_file: FieldStorage, model: str = "microsoft/git-large-coco"
    ) -> str:
        processor = GitProcessor.from_pretrained(model)
        model = GitForCausalLM.from_pretrained(model)
```

```

raw_image = Image.open(uploaded_file).convert("RGB")
device = "cuda" if torch.cuda.is_available() else "cpu"
model.to(device)
inputs = processor(images=raw_image, return_tensors="pt")
pixel_values = inputs.pixel_values.to(device)
generated_ids = model.generate(pixel_values=pixel_values, max_length=50)
responses = processor.batch_decode(generated_ids, skip_special_tokens=True)
response = responses[0]
print(response)
return response

```

A.4 SCRIPT VIT-GPT2

```

from transformers import pipeline
class LabeliserURL:
    def generate(url: str, model: str = "nlpconnect/vit-gpt2-image-captioning") -> str:
        image_to_text = pipeline("image-to-text", model)
        response = image_to_text(url)[0]["generated_text"]
        return response

```

A.5 SCRIPT SERVEUR

```

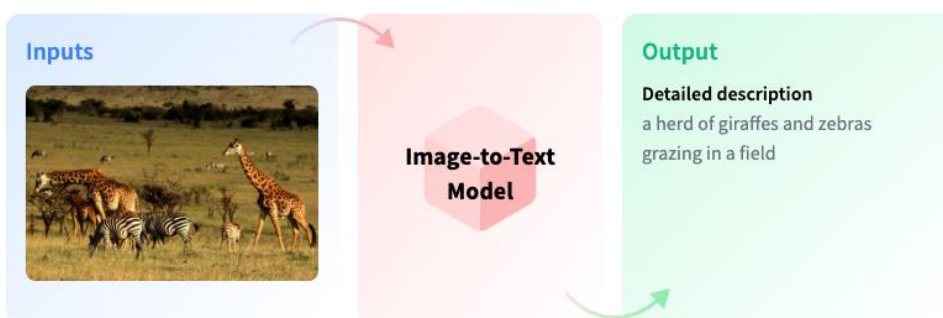
from flask import Flask
from flask import request
from flask import render_template
# from labeliserBlip import Labeliser
# from labeliserGit import Labeler
# from labeliserGPT2 import LabeliserURL
from translatorOpus import Translator
app = Flask(__name__)
@app.route("/")
def index():
    return render_template("form.html")
@app.route("/upload", methods=["GET", "POST"])
def upload_file():
    # traduction = None
    if request.method == "POST":
        uploaded_file = request.files["image"]
        label = Labeliser.generate(uploaded_file)
        traduction = Translator.translate(label)
        return render_template("final.html", traduction=traduction)
@app.route("/url", methods=["GET"])
def upload_url():
    url = request.args.get("url")
    label = LabeliserURL.generate(url)
    print(label)
    traduction = Translator.translate(label)
    return render_template("final.html", traduction=traduction)

```

A.6 SCRIPT TRADUCTION

```
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM
class Translator:
    def translate(label: str) -> str:
        tokenizer = AutoTokenizer.from_pretrained("t5-large")
        model2 = AutoModelForSeq2SeqLM.from_pretrained("t5-large")
        batch = tokenizer([label], return_tensors="pt")
        generated_ids = model2.generate(**batch)
        traductions = tokenizer.batch_decode(generated_ids, skip_special_tokens=True)
        print("Traductions : ", traductions)
        traduction = traductions[0]
        return traduction
```

A.7 SCHÉMA DE LA DESCRIPTION DE L'IMAGE



Source : <https://huggingface.co/tasks/image-to-text>

A.8 EXEMPLE D'IMAGE ET DE DESCRIPTIONS



Blip : Il y a un vase de fleurs sur une table sur un patio

GIT : un patio avec vue sur le jardin

VitGPT2 : un vase rempli de fleurs assis sur une banquette en bois

ChatGPT-4V : Une pergola moderne surplombe une terrasse en bois agrémentée d'un bouquet de fleurs, avec en arrière-plan un jardin verdoyant et un ciel nuageux.

A.9 EXTRAIT GRILLE D'ÉVALUATION

Images	Description des images Modèle A	Qualité du label					
		Description correcte [1]	Grammaire correcte [2]	Information correcte [4]	Description pertinente [5]	Description créative [6]	Rédigée par l'humain [7]
		La description décrit correctement l'image ? [3]	La description est grammaticalement correcte ?	La description contient des informations correctes ?	La description est pertinente pour cette image ?	La description est construite de manière créative ?	La description a été rédigée par un humain ?
1 01-33793_pergola-a-lames-orie-ntables-pure-line.jpeg	Une pergola moderne surplombe une terrasse en bois agrémentée d'un bouquet de fleurs, avec en arrière-plan un jardin verdoyant et un ciel nuageux.	ni d'accord ...	tout à fait d'...	ni d'accord ...	d'accord	pas d'accord	ni d'accord ...
2 02-33232_baies-couissantes-et-porte-d-entree-en-aluminium.jpeg	Une véranda avec de grandes baies vitrées et un toit en tuiles rouges est en construction à côté d'une maison, avec un camion "TRYBA" garé à proximité et un jardin en arrière-plan.	tout à fait d'...	tout à fait d'...	tout à fait d'...	tout à fait d'...	pas d'accord	tout à fait d'...
3 03-32670_portail-aluminium.jpeg	Un portail gris moderne est encadré par deux piliers en pierre, avec à gauche un abri en béton et à droite une vue partielle d'une maison blanche, le tout entouré d'arbres et d'un tapis de feuilles au sol.	d'accord	tout à fait d'...	tout à fait d'...	d'accord	ni d'accord ...	pas d'accord
4 04-32251_porte-d-entree-aluminium.jpeg	Une maison beige présente une porte d'entrée noire élégante avec des motifs dorés et un garage assorti, le tout encadré par un arbre aux feuilles pourpres et un buisson verdoyant.	d'accord	tout à fait d'...	d'accord	tout à fait d'...	ni d'accord ...	d'accord
5 05-32203_pergola-alu-pure-line.jpeg	Une maison à la façade claire avec une charpente en bois est dotée d'un balcon métallique et d'une terrasse couverte au rez-de-chaussée, le tout donnant sur un jardin avec des plantes et un ciel nuageux en arrière-plan.	d'accord	tout à fait d'...	d'accord	d'accord	ni d'accord ...	d'accord
6 06-33961_fenêtres-ta84oc-avec-volets-integras-et-porte-d-entree-modele-tech-aluminium.jpeg	Une maison en briques rouges avec un toit en ardoise présente plusieurs fenêtres et une lucarne, entourée d'un jardin soigné avec des arbustes fleuris, un jeune arbre et un ciel légèrement nuageux en arrière-plan.	ni d'accord ...	tout à fait d'...	ni d'accord ...	ni d'accord ...	ni d'accord ...	ni d'accord ...
7 07-32109_fenetre-bois.jpeg	Une maison en pierre traditionnelle arbore une fenêtre encadrée de volets bleus, ornée de jardinières fleuries et de rideaux blancs à motifs.	ni d'accord ...	pas d'accord	d'accord	ni d'accord ...	ni d'accord ...	pas d'accord
8 08-31941_fenêtres-et-portes-fenêtres-ta84oc-avec-volets-battants-du-gîte-a-plelo.jpeg	Une grande maison blanche aux volets verts est nichée au pied d'une forêt dense, avec un jardin bien entretenu devant, sous un ciel clair.	d'accord	pas d'accord	ni d'accord ...	d'accord	d'accord	pas d'accord
9 09-28945_deux-portes-fenêtres-pvc-70-leguevin.jpeg	Une maison moderne avec des volets blancs donne sur une piscine bleue éclatante, avec un robot nettoyeur de piscine sur le bord et des lampes posées sur une terrasse en bois.	ni d'accord ...	tout à fait d'...	ni d'accord ...	d'accord	pas d'accord	pas d'accord
10 10-27631_volets-stores-a-villeneuve-les-maguelone.jpeg	Il s'agit d'une maison contemporaine aux lignes épurées, avec des murs gris et blancs, des fenêtres équipées de persiennes horizontales et une pelouse verdoyante à l'avant.	d'accord	tout à fait d'...	tout à fait d'...	d'accord	pas d'accord	pas d'accord
11 11-20951_fenêtres-bois-a-l-ancienne.jpeg	Il s'agit d'une élégante façade de maison en briques rouges et blanches, dotée de grandes fenêtres, d'une porte en fer forgé et de décorations dorées à côté de l'entrée principale.	d'accord	ni d'accord ...	tout à fait d'...	ni d'accord ...	pas d'accord	pas d'accord
12 12-32507_menuiseries-aluminium.jpeg	L'image présente une cuisine lumineuse avec de grandes portes-fenêtres noires donnant sur un jardin, des étagères blanches contenant divers objets et vaisselles, un fauteuil noir, et un plan de travail moderne avec un évier et des ustensiles de cuisine.	tout à fait d'...	tout à fait d'...	tout à fait d'...	ni d'accord ...	pas d'accord	pas d'accord