

Technologies d'assistance pour les personnes malvoyantes basées sur la vision : avancées, limites et perspectives

Aela Le Sommer, Panagiotis Papadakis, Christophe Lohr

Introduction

Défis 🚩 : Les personnes malvoyantes peinent à se déplacer et à comprendre des environnements inconnus.



Introduction

Défis  : Les personnes malvoyantes peinent à se déplacer et à comprendre des environnements inconnus.



Intérêt : Solutions basées sur des modèles IA

Introduction

Défis 🚩 : Les personnes malvoyantes peinent à se déplacer et à comprendre des environnements inconnus.



Intérêt : Solutions basées sur des modèles IA

→ Vision

→ Langage

Introduction

Défis  : Les personnes malvoyantes peinent à se déplacer et à comprendre des environnements inconnus.



Intérêt : Solutions basées sur des modèles IA

→ Vision

→ Langage

Image captioning, video summarization et Visual-Question Answering (VQA) ➔ solutions prometteuses pour aider les personnes malvoyantes dans leur vie quotidienne.



1. État de l'art

Présentation des tâches

Image Captioning :



*

Image
Captioning

Un homme en chemise
noire joue de la guitare.

Présentation des tâches

Image Captioning :



*

Image
Captioning

Un homme en chemise
noire joue de la guitare.

VQA :



**

Quels animaux sont sur la photo ?

Visual Question
Answering

Il y a des chats sur la photo.

Présentation des tâches

Image Captioning :



*

Image
Captioning

Un homme en chemise
noire joue de la guitare.

VQA :



**

Quels animaux sont sur la photo ?

Visual Question
Answering

Il y a des chats sur la photo.

Video Summarization :

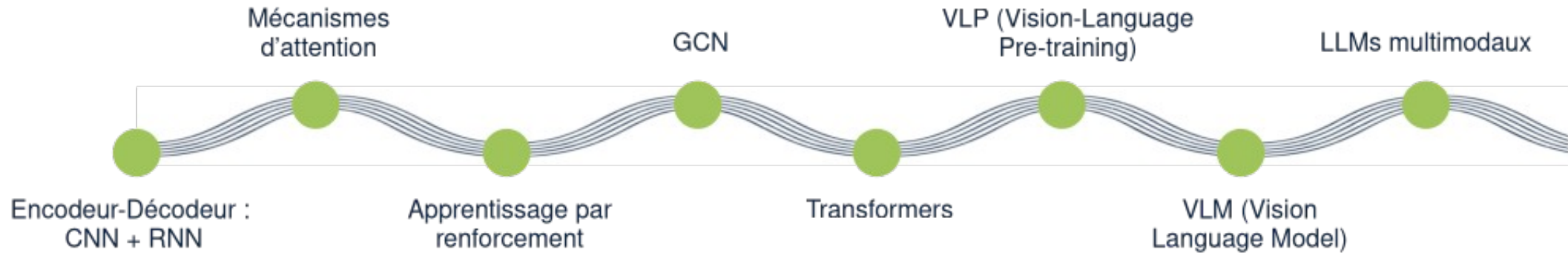


Video
Summarization

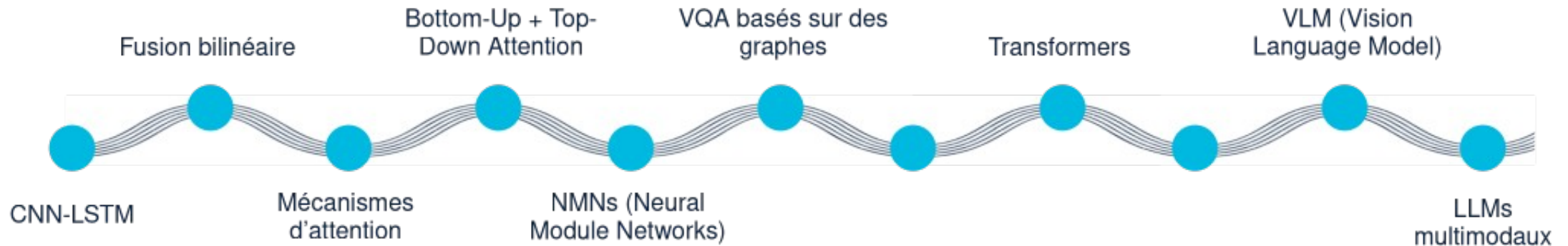
Une petite fille avec un manteau
rose joue dans une aire de jeux.

Évolution des architectures des modèles

Image Captioning :



VQA :



Solutions basées sur la vision disponible sur le marché pour les personnes malvoyantes

- Lunettes intelligentes pour les personnes malvoyantes : Envision Glasses, OrCam MyEye, etc.
- Exemples d'applications mobiles d'assistance basées sur la vision pour les personnes malvoyantes :

App. Mobile	Reconnaissance d'objets	Lecture de textes	Description de scène	VQA	Reconnaissance faciale	Nb Téléchargements
Be My Eyes (Be My AI)			✓	✓		1M+*
Lookout	✓	✓	✓	✓		500k+*
TapTapSee	✓	✓				500k+*
Envision	✓	✓	✓	✓	✓	100k+*
Seeing AI		✓	✓	✓	✓	100k+*
Supersense	✓	✓				50k+*
Aipoly	✓	✓				-
Oorion	✓	✓				-
VizWiz				✓		-
VoiceVision			✓			-

* sur Google Play en février 2025

Approches principales

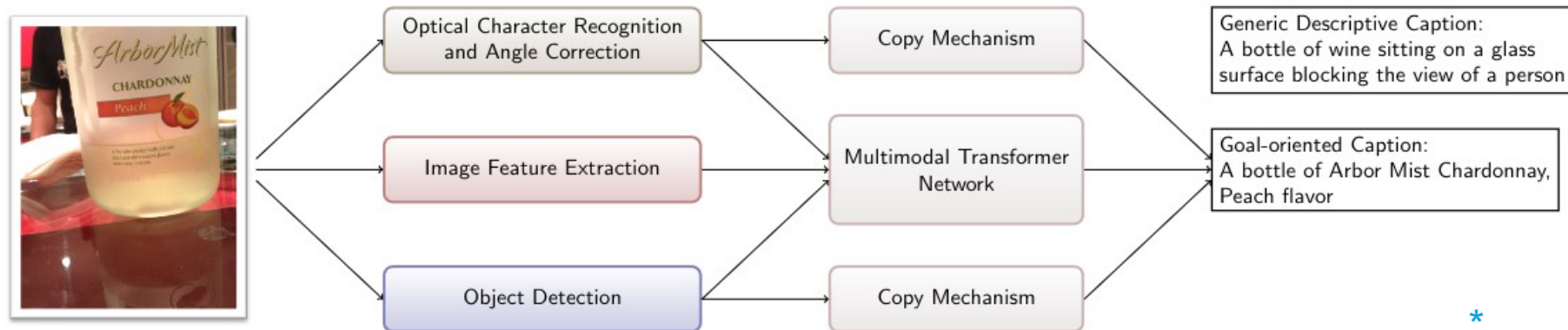
(Image Captioning, Video Summarization, VQA) + Text-to-Speech

➤ ⚠ Limites liées à leur structure en plusieurs étapes :

- Pertes d'informations,
- Délai de traitement plus important.

➤ Cas d'application pour les malvoyants :

- Ajouts de modules complémentaires (OCR, détection d'objets, correction d'angle, etc.)
- **BUT** : Améliorer la fiabilité des descriptions/réponses.



Approches principales

(Image Captioning, Video Summarization, VQA) + Text-to-Speech

➡ ⚠ Limites liées à leur structure en plusieurs étapes :

- Pertes d'informations,
- Délai de traitement plus important.

➡ Cas d'application pour les malvoyants :

- Ajouts de modules complémentaires (OCR, détection d'objets, correction d'angle, etc.)
 - **BUT** : Améliorer la fiabilité des descriptions/réponses.

Image-to-Speech end-to-end ➡ convertit directement l'image en parole

➡ ⚠ Limites :

- Complexité d'évaluation,
- Entraînement nécessitant de grandes quantités de paires image/audio ➡ difficile et coûteux à obtenir.



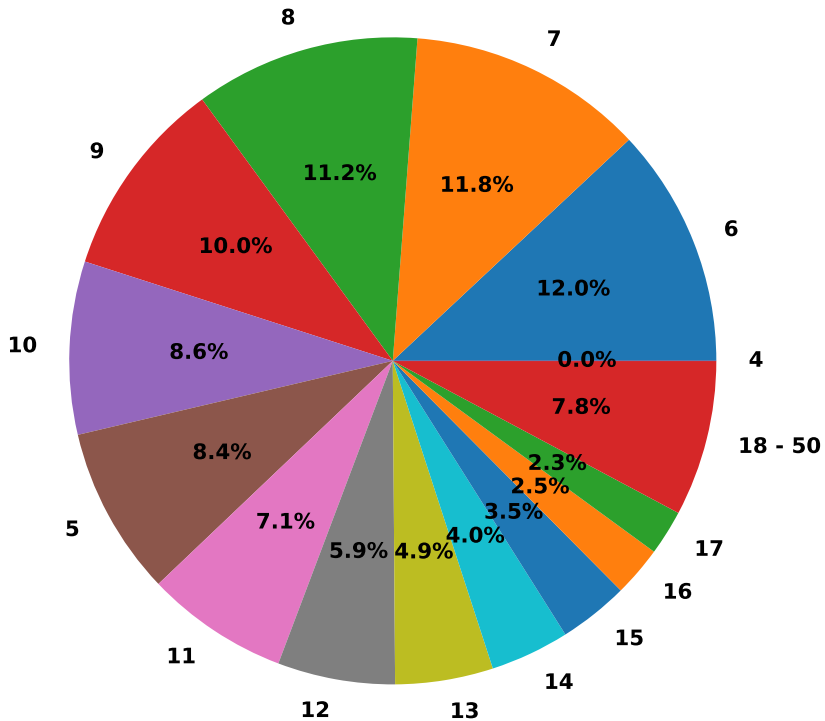
2. Limites

Principaux jeux de données d'image Captioning et de Visual-Question Answering

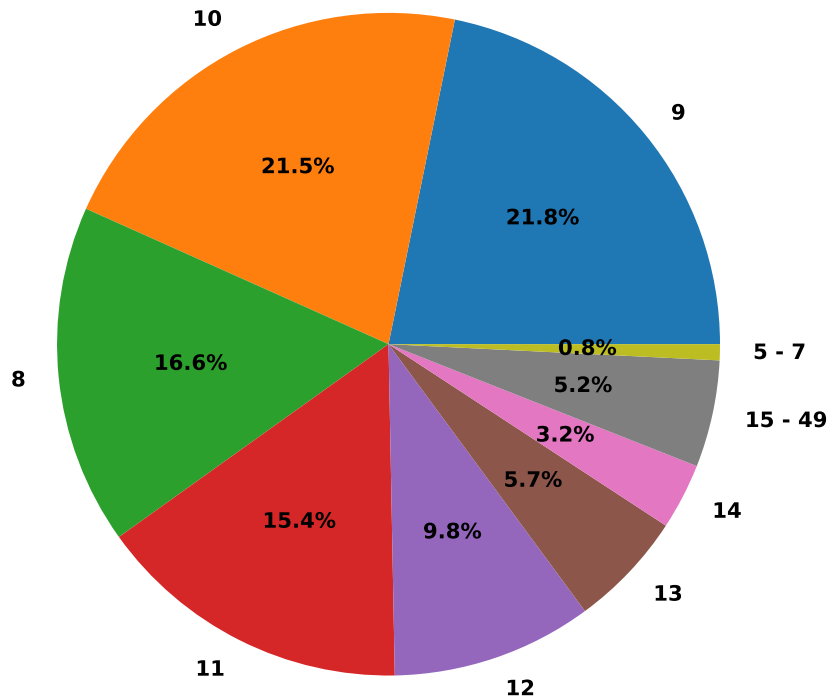
Jeux de données	Tâches	Source des images	Nb images	Nb descript°/quest°	Apports potentiels pour des modèles destinés aux malvoyants
VizWiz [11]	IC, VQA	Photo prises par des malvoyants	+ 39k / +32k	+ 195k / +32k	Questions et images réelles issues d'utilisateurs malvoyants
MSCOCO [11]	IC	Flickr	+123k	+ 616k	Grandes diversités d'environnement (intérieur, extérieur, objets courant), Diversité des annotations
Flickr30k [11]	IC	Flickr	+ 31k	+ 158k	Images d'activités, d'événements et de scènes de la vie quotidienne
Conceptual Captions [11]	IC	Pages Web	+ 3.3M	+ 3.3M	Très grand jeux de données, Grande variété de types d'images (ex : images naturelles, de produits, dessins animés, dessins, etc.)
NoCaps [11]	IC	OpenImages (validation + test)	+ 15k	+ 166k	Conçu pour tester la généralisation sur de nouveaux objets, ce qui peut être utile pour décrire des scènes inconnues
Visual Genome [11, 1]	IC, VQA	YFCC100M $\cap$ MS-COCO	+ 108k	5.4M descriptions de régions / 1.7M	Très dense au niveau des annotations : descriptions de régions, objets, attributs, relations, graphes de régions, graphes de scènes et VQA
GQA [1]	VQA	Visual Genome	+ 113k	+ 22M	Exploite les représentations sémantiques des questions et les graphes de scènes pour répondre à la question, permettant un raisonnement structuré
VQAv2 [1]	VQA	COCO	+ 204k	+ 1.1M	Biais réduit : Chaque question est associée à une paire d'images similaires qui donnent lieu à deux réponses différentes
OK-VQA [1]	VQA	COCO	+ 14k	+ 14k	Images du monde réel nécessitant des connaissances externes pour répondre aux questions
TDIUC [17]	VQA	MS-COCO + Visual Genome	+ 167k	+ 1.6M	Grande diversité de type de questions (12), comprenant des questions absurdes pour forcer le système à raisonner sur le contenu de l'image.

Nombre de mots dans les descriptions

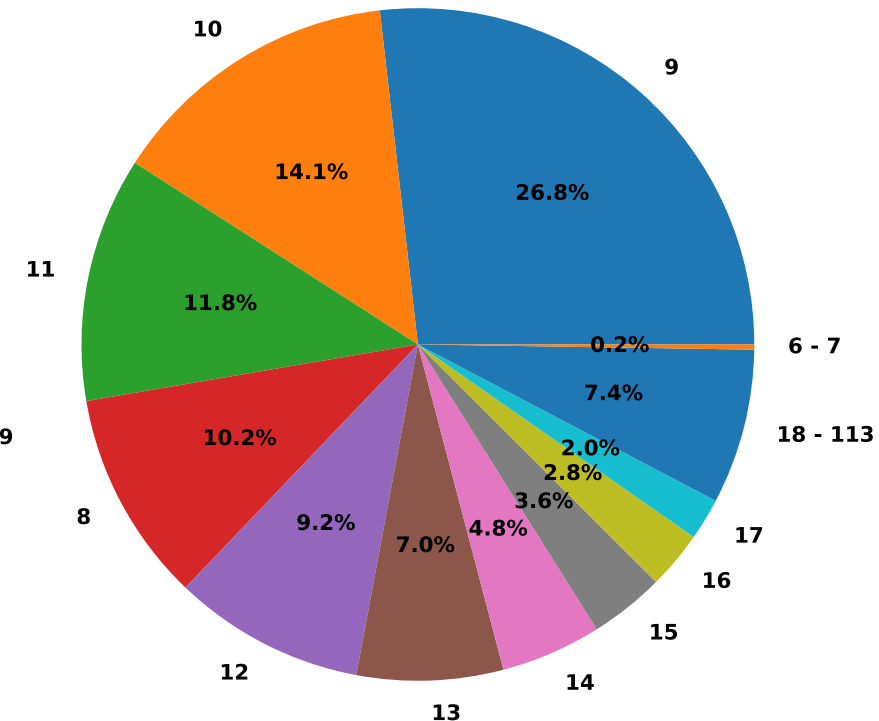
Conceptual Caption



MSCOCO



VizWiz

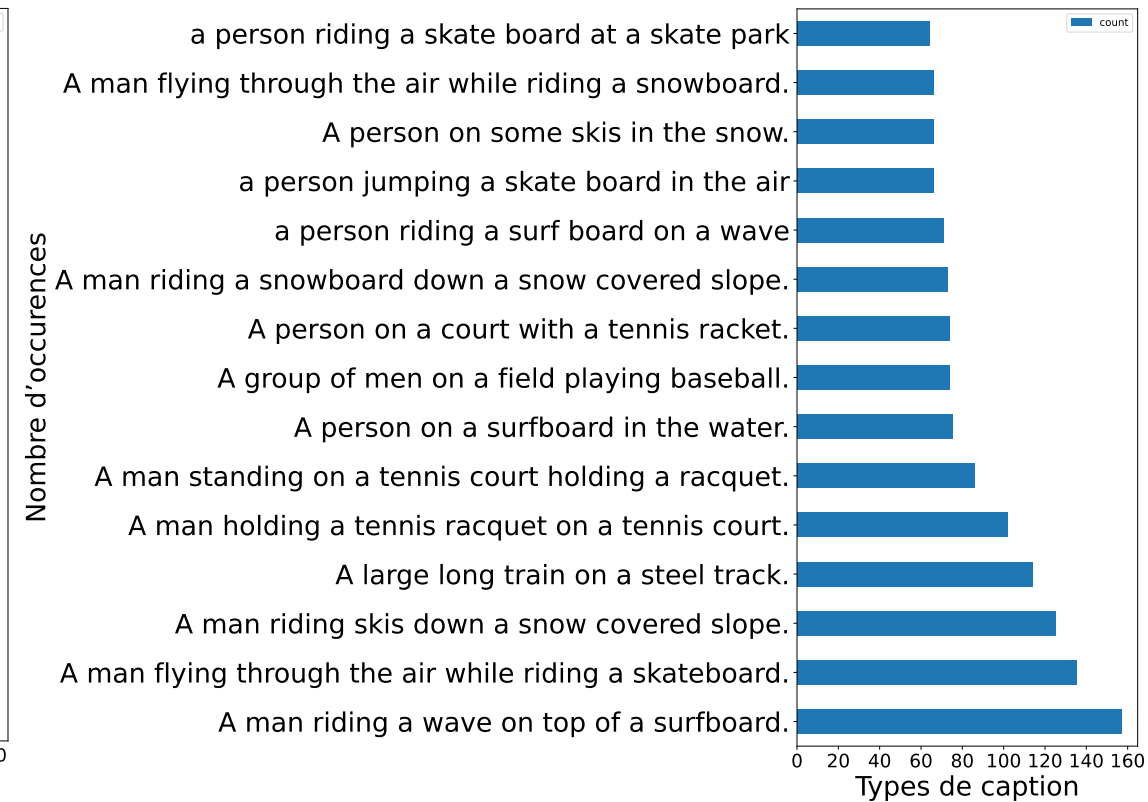
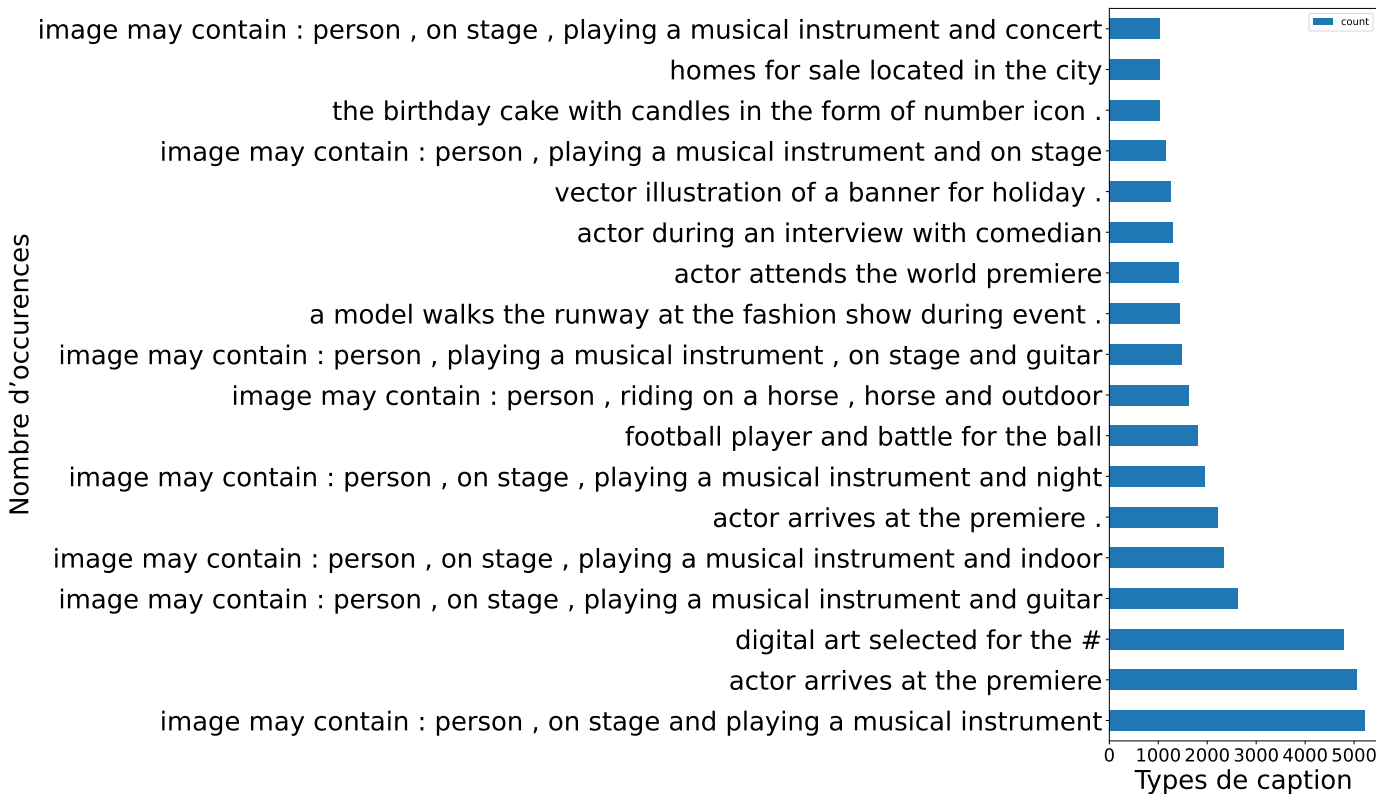


+90 % des descriptions < 20 mots ➔ descriptions générales, peu détaillées

Descriptions les plus courantes dans les jeux de données (1/2)

Conceptual Caption

MSCOCO



Descriptions les plus courantes dans les jeux de données (2/2)

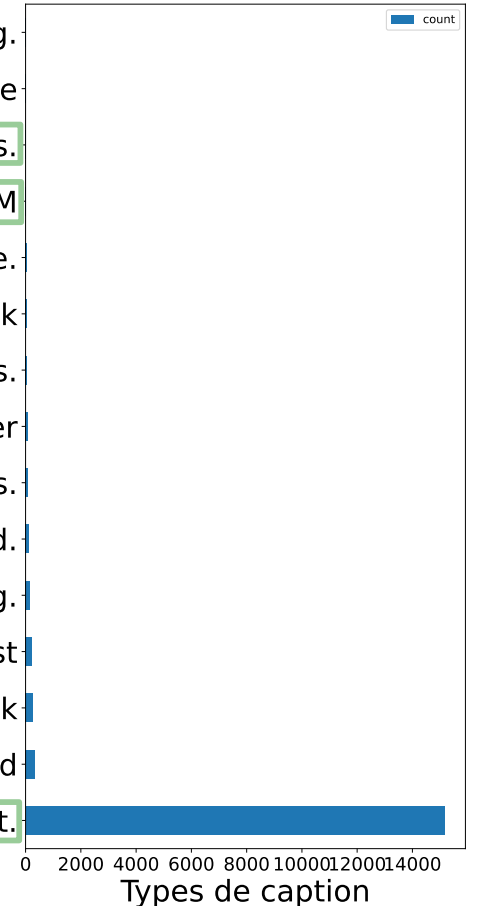
VizWiz

Prend en compte les problèmes de qualité des photos ...

... MAIS descriptions qui restent vagues.

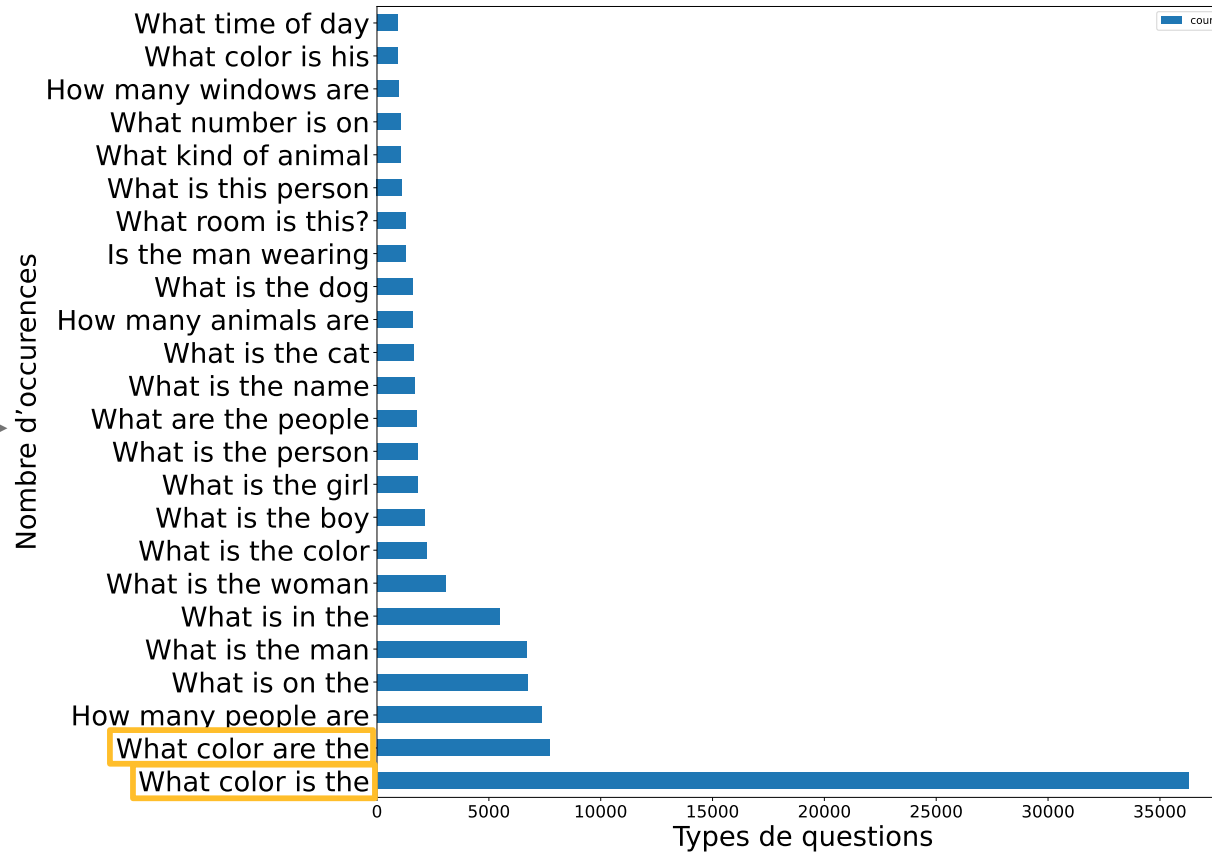
Nombre d'occurrences

The nutrition facts for the food are on the back of the packaging.
 A computer screen is displaying a window with a message
 Focus is not well and image has some drawbacks.
 IMAGE WAS UNCLEAR BUT IT IS NOT ITEM
 A package of food is on top of a table.
 The packaging for the product has the ingredients on the back
 A container / package that contains various goods / edible / liquid items.
 some type of liquid that is in a container
 a container/ box / bottle that contains liquid / goods.
 imagine how you would describe this image on the phone to a friend.
 A picture of the food is on the packaging.
 Beautiful view from behind the walls hidden under dark mist
 A wonderful view of the fog windows in the room is very thick
 A very wonderful view and worth seeing at all times, my friend
 Quality issues are too severe to recognize visual content.



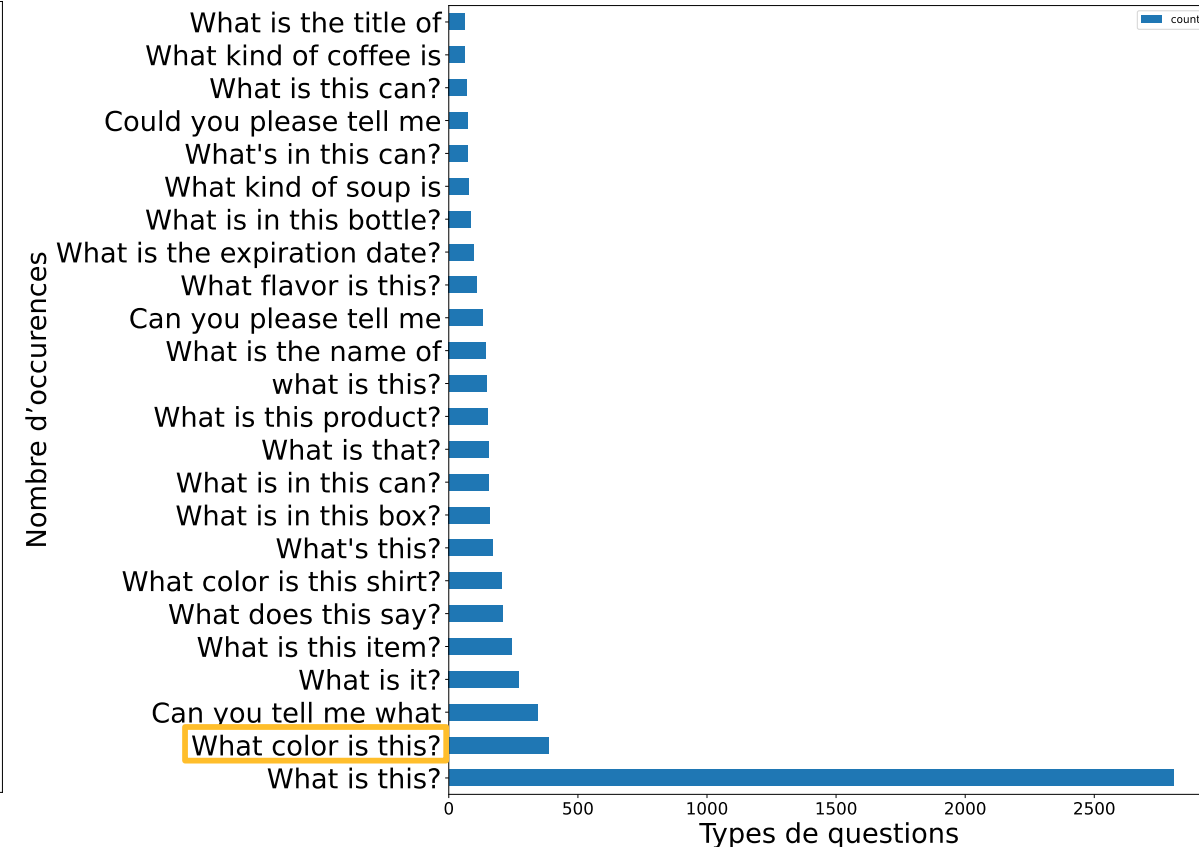
Types de questions les plus posées dans les jeux de données

VQA v2



Grand nombre de questions qui supposent une connaissance préalable de la scène

VizWiz



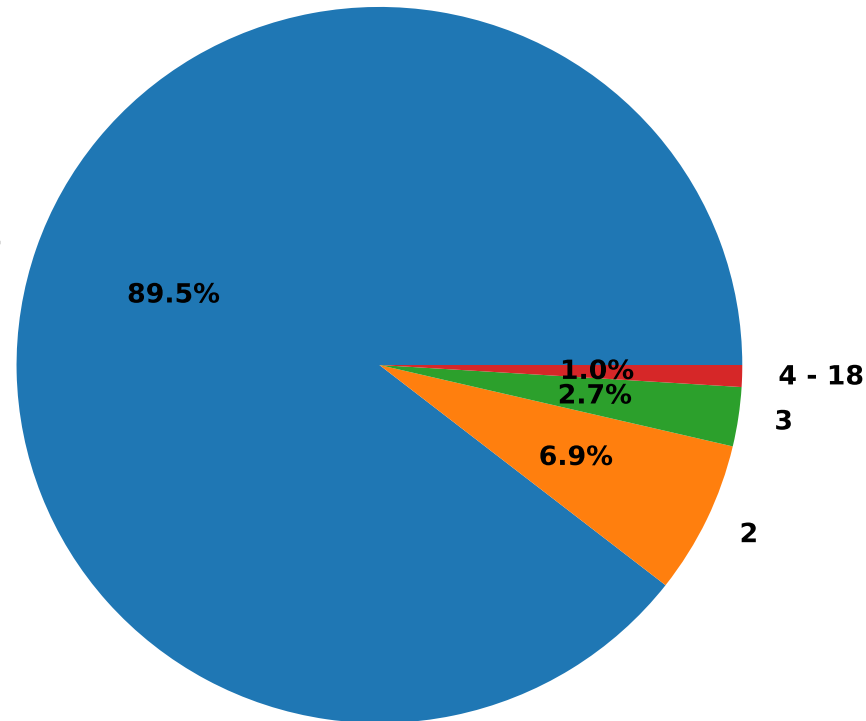
Formulation des questions :
VizWiz (malvoyants) $\neq$ autres jeux de données

Section 2

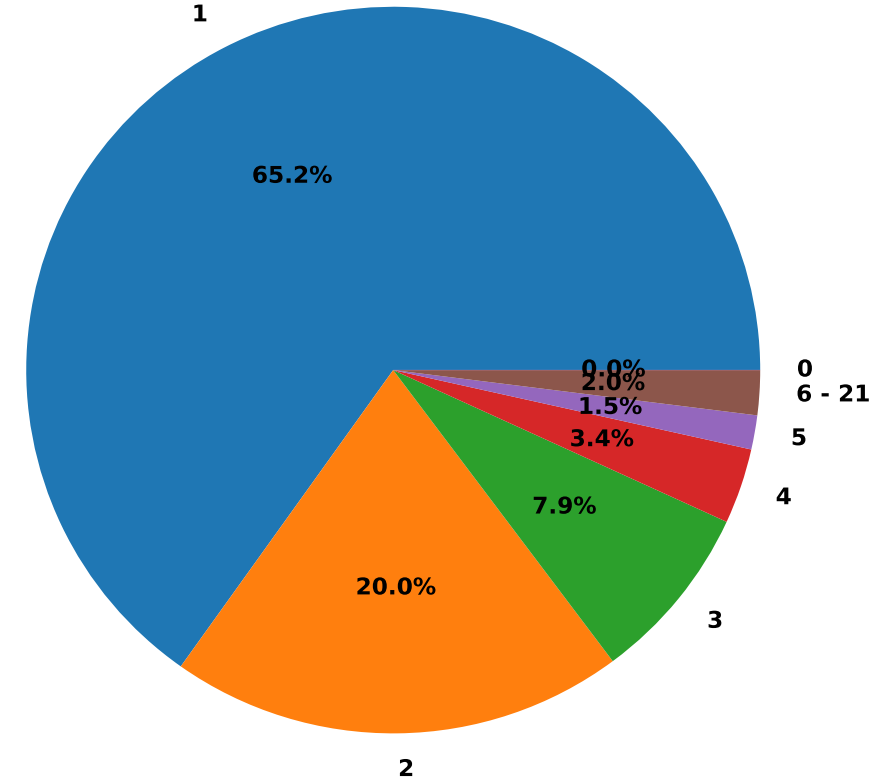
Limites

Nombre de mots dans les réponses

VQA v2



VizWiz



Réponses courtes

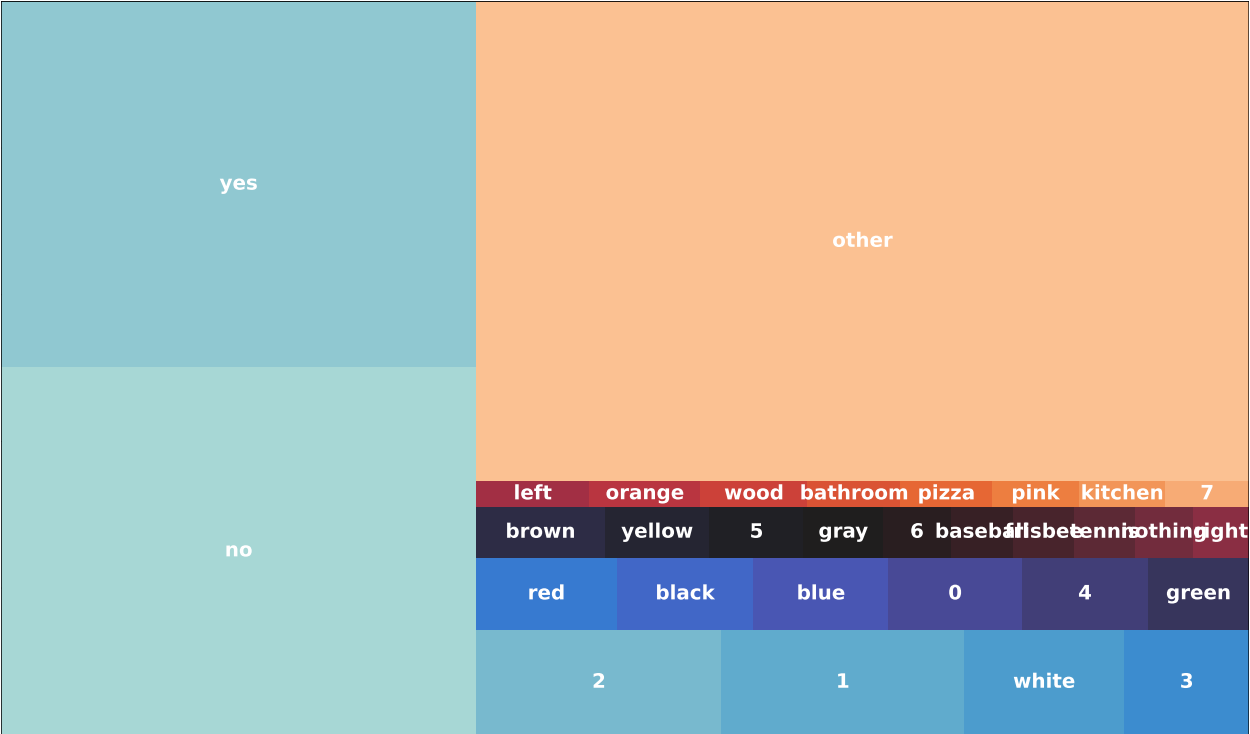
➔ Efficacité

➔ Facilité d'évaluation

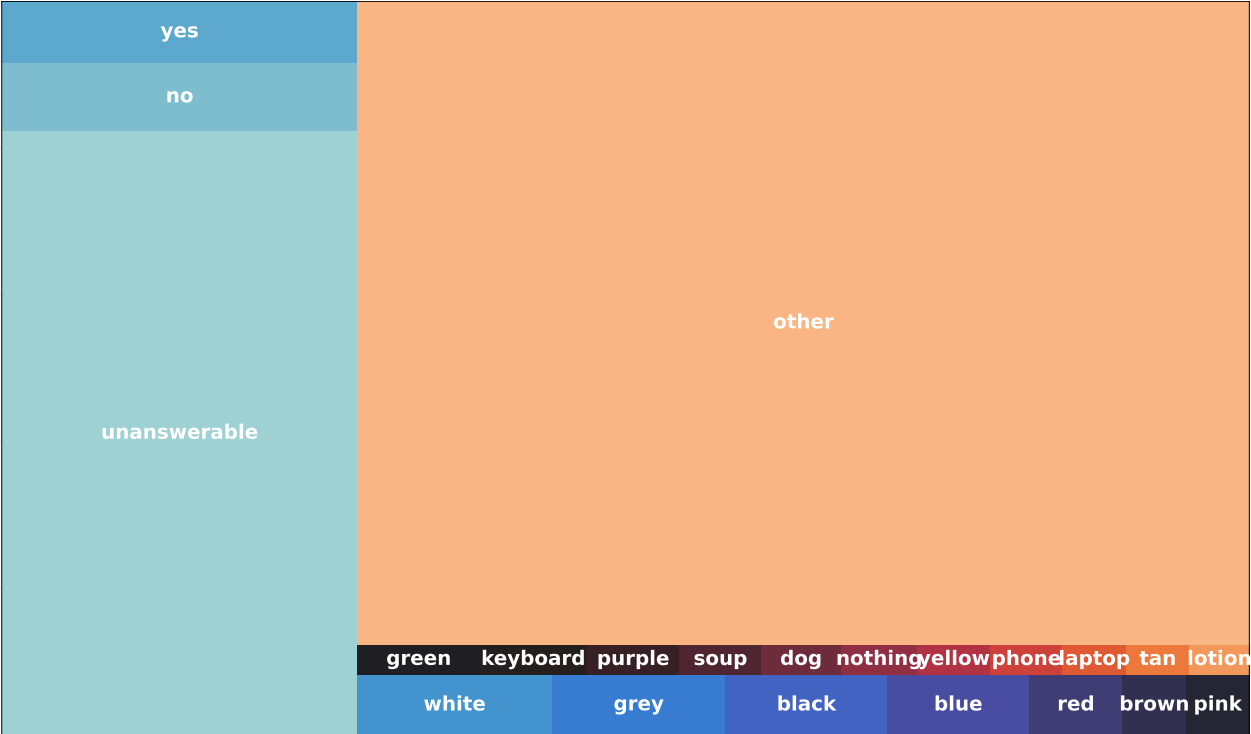
Modèles pas entraînés à générer
des réponses complexes

Répartition des réponses dans les jeux de données

VQA v2



VizWiz



Section 2

Limites

Principales problématiques

Section 2

Limites

Principales problématiques

■ Mauvaise qualité des photos prises par les utilisateurs malvoyants,



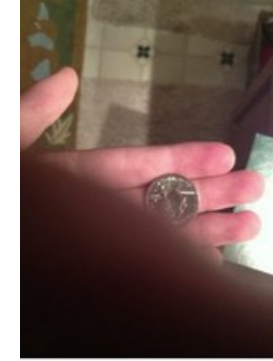
Photos provenant du jeu de données
VizWiz prises par des malvoyants

Section 2

Limites

Principales problématiques

- Mauvaise qualité des photos prises par les utilisateurs malvoyants,
- Complexité d'évaluation des modèles en raison de leur nature générative,



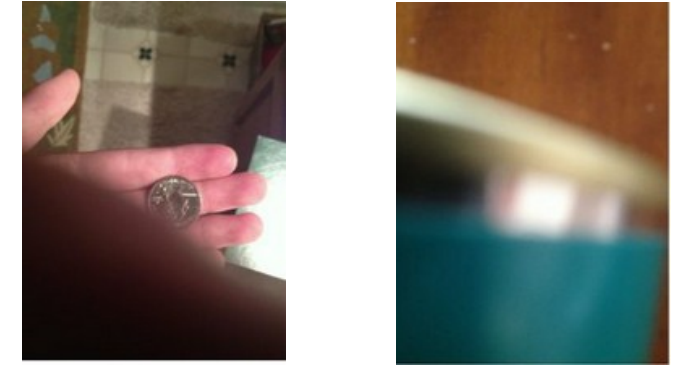
Photos provenant du jeu de données VizWiz prises par des malvoyants

Section 2

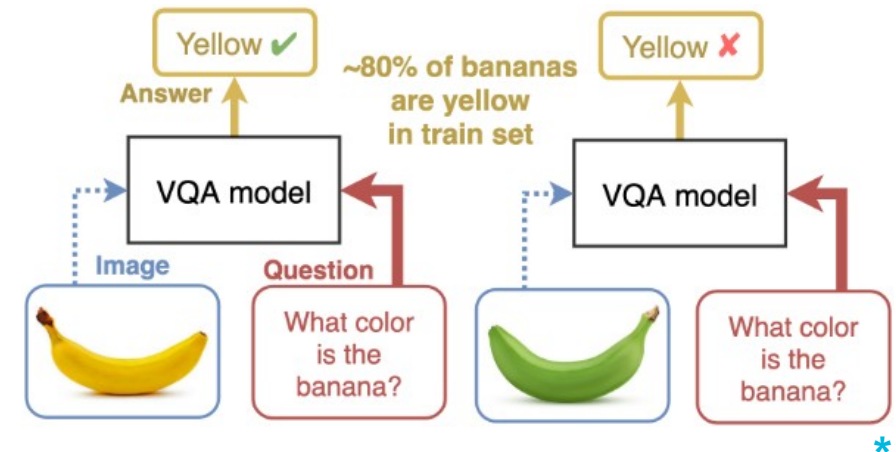
Limites

Principales problématiques

- Mauvaise qualité des photos prises par les utilisateurs malvoyants,
- Complexité d'évaluation des modèles en raison de leur nature générative,
- Biais dans les modèles,



Photos provenant du jeu de données VizWiz prises par des malvoyants



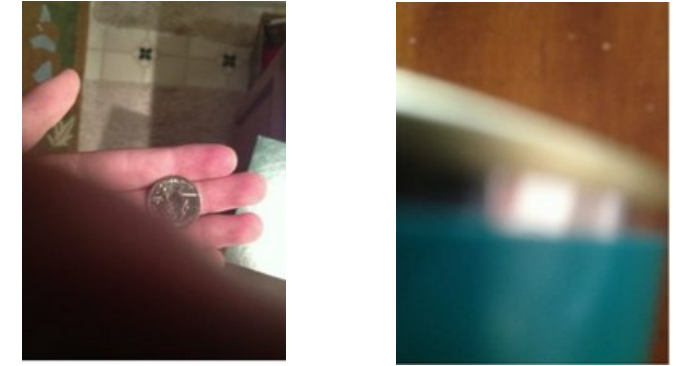
Exemple de biais hérités par les modèles

Section 2

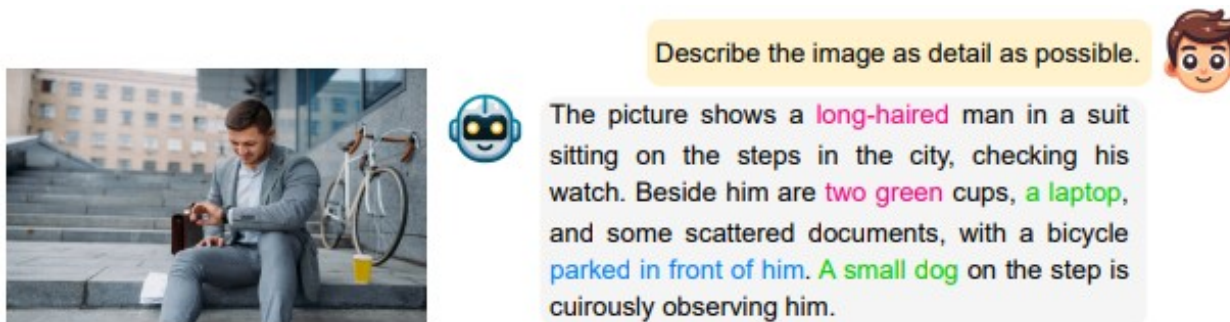
Limites

Principales problématiques

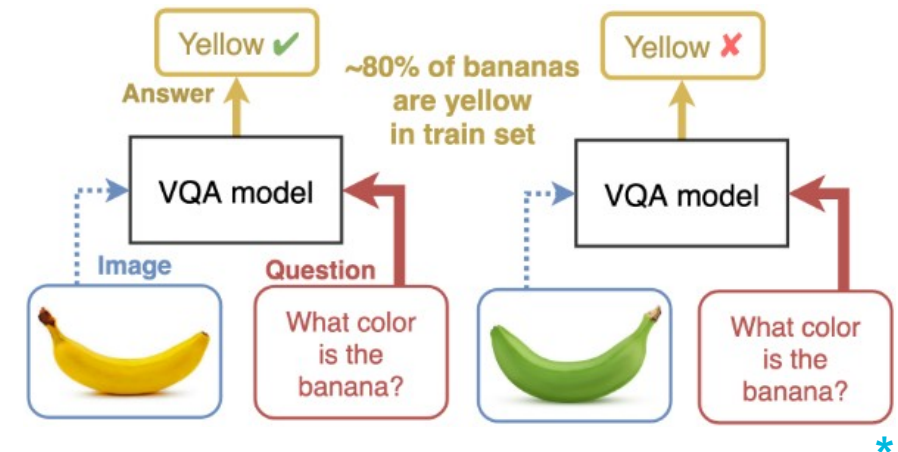
- Mauvaise qualité des photos prises par les utilisateurs malvoyants,
- Complexité d'évaluation des modèles en raison de leur nature générative,
- Biais dans les modèles,
- Hallucinations des modèles,



Photos provenant du jeu de données VizWiz prises par des malvoyants



Exemple d'hallucinations (objets, attributs, relations) **



Exemple de biais hérités par les modèles *



3. Pistes d'exploration

Section 3

Pistes d'exploration

► Poser le problème comme un problème d'intelligence incarnée (embodied intelligence)

► BUT :

- Pallier les limites des benchmarks existants,
- Contrôle total sur la scène,
- Facilité de génération de scénarios variés adaptés au cas d'application,
- Annotations automatiques, précises et exhaustives (segmentation, coordonnées, matériaux, etc.)
- Contourner les problèmes de confidentialité,

Section 3

Pistes d'exploration

► Poser le problème comme un problème d'intelligence incarnée (embodied intelligence)

► BUT :

- Pallier les limites des benchmarks existants,
- Contrôle total sur la scène,
- Facilité de génération de scénarios variés adaptés au cas d'application,
- Annotations automatiques, précises et exhaustives (segmentation, coordonnées, matériaux, etc.)
- Contourner les problèmes de confidentialité,

► Distinction des niveaux de gravité des hallucinations

Section 3

Pistes d'exploration

► Poser le problème comme un problème d'intelligence incarnée (embodied intelligence)

► BUT :

- Pallier les limites des benchmarks existants,
- Contrôle total sur la scène,
- Facilité de génération de scénarios variés adaptés au cas d'application,
- Annotations automatiques, précises et exhaustives (segmentation, coordonnées, matériaux, etc.)
- Contourner les problèmes de confidentialité,

► Distinction des niveaux de gravité des hallucinations

► Intégrer l'utilité des réponses dans l'évaluation

Section 3

Pistes d'exploration

Évaluation de la qualité et sélection des images :

■ Actuellement :

- Évaluation de la qualité de l'image avant Image Captioning/VQA.
 - ☒ Bonne qualité ➔ génération d'une réponse/description
 - ☒ Mauvaise qualité ➔ Notification à l'utilisateur ➔ nouvelle photo demandée
 - ⚠ Procédé long et contraignant au quotidien.

Section 3

Pistes d'exploration

Évaluation de la qualité et sélection des images :

■ Actuellement :

- Évaluation de la qualité de l'image avant Image Captioning/VQA.
 - ☒ Bonne qualité ➔ génération d'une réponse/description
 - ☒ Mauvaise qualité ➔ Notification à l'utilisateur ➔ nouvelle photo demandée
 - ⚠ Procédé long et contraignant au quotidien.

● Piste d'amélioration :

- Capture de l'environnement (série de photos) ➔ filtrage des images de bonne qualité ➔ Image Captioning/VQA sur la série
 - **BUT :**
 - Réduire le risque de réponses/descriptions peu fiables liées à la mauvaise qualité de l'image,
 - Limiter le nombre de reprises de photos.
 - ⚠ Limites : Étapes de calculs supplémentaires



IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom



MERCI

Aela Le Sommer



Annexes

Annexes

Lunettes intelligentes pour les personnes malvoyantes

Envision Glasses :

- ✓ Instant Text,
- ✓ Scan Text,
- ✓ Batch Scan,
- ✓ Call a Companion,
- ✓ Call Aira,
- ✓ Describe Scene,
- ✓ Detect Light,
- ✓ Recognise Cash,
- ✓ Detect Colors,
- ✓ Find People,
- ✓ Find Objects,
- ✓ Teach a Face and Explore.



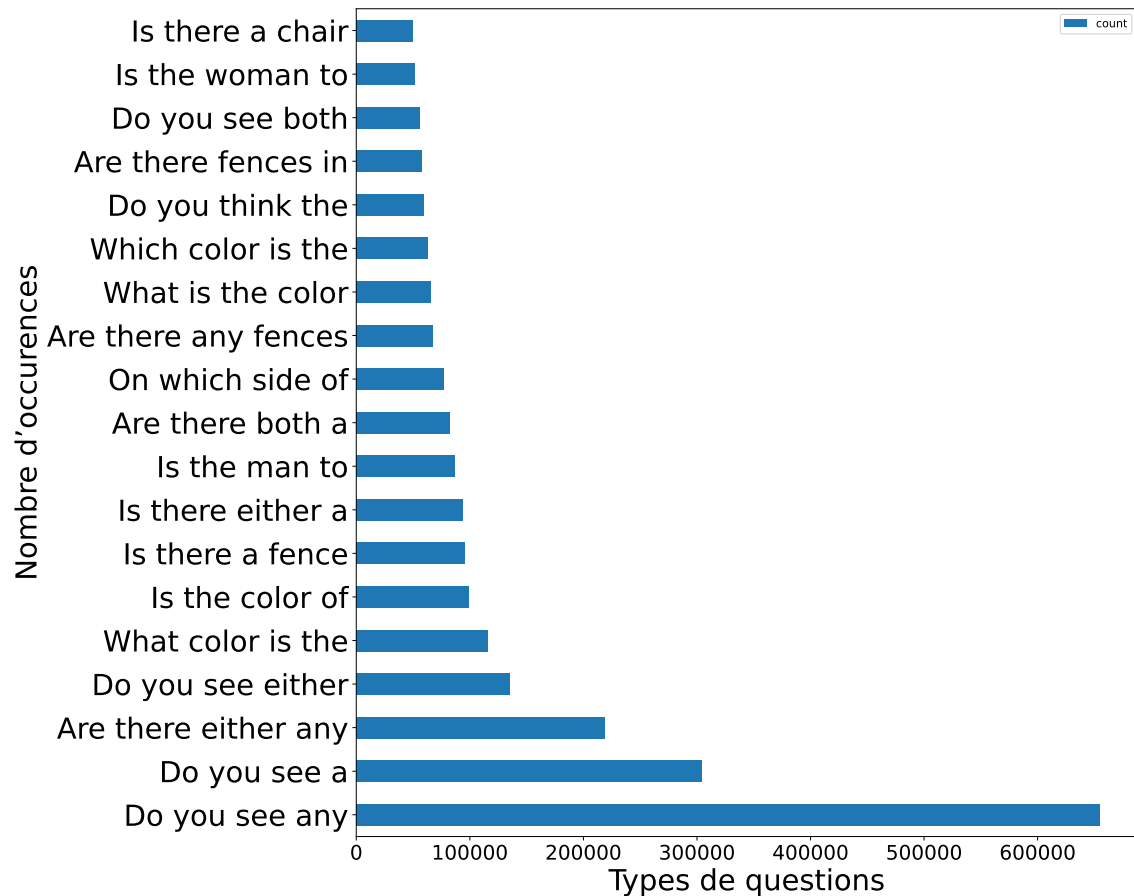
OrCam MyEye 3 Pro :

- ✓ AI Assistant-Just Ask
- ✓ Summarize text into main topics
- ✓ Utilizes intext and outsourced data
- ✓ Reads handwriting
- ✓ Zoom in&out
- ✓ Change contrast
- ✓ Extract and copy text
- ✓ Converts image text to digital text
- ✓ Connect to any screen
- ✓ Supports over 140 languages
- ✓ Voice commands
- ✓ Text reading
- ✓ Smart reading
- ✓ Recognize Faces
- ✓ Identifying Products
- ✓ Money Notes
- ✓ Barcodes
- ✓ Colors

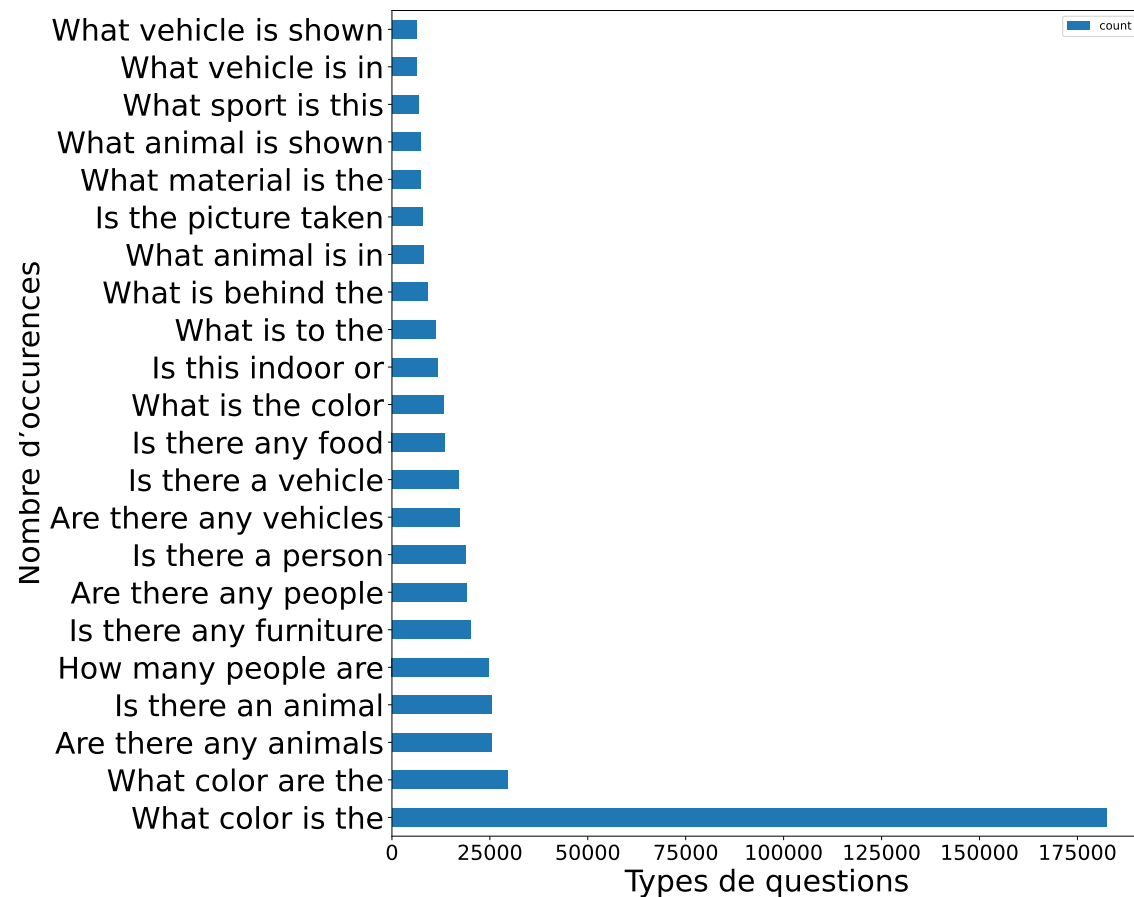


Types de questions les plus posées dans les jeux de données

GQA

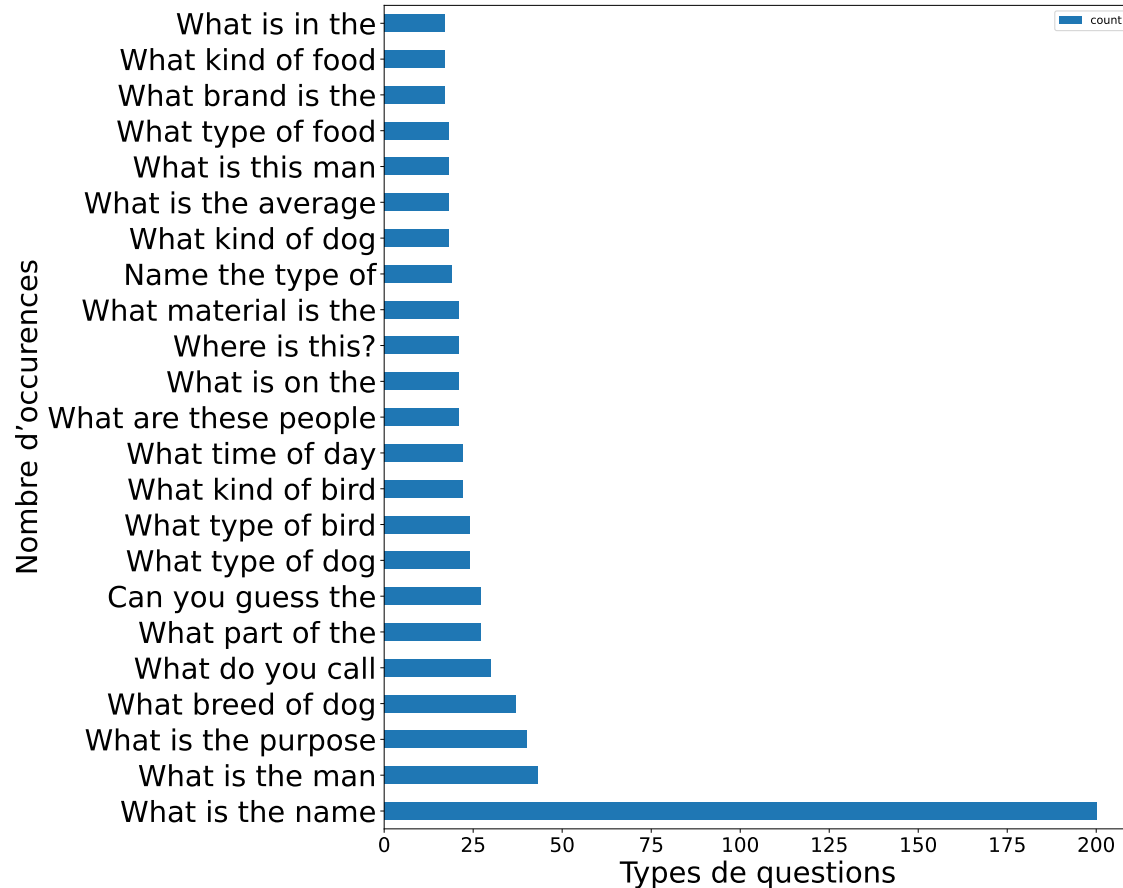


TDIUC

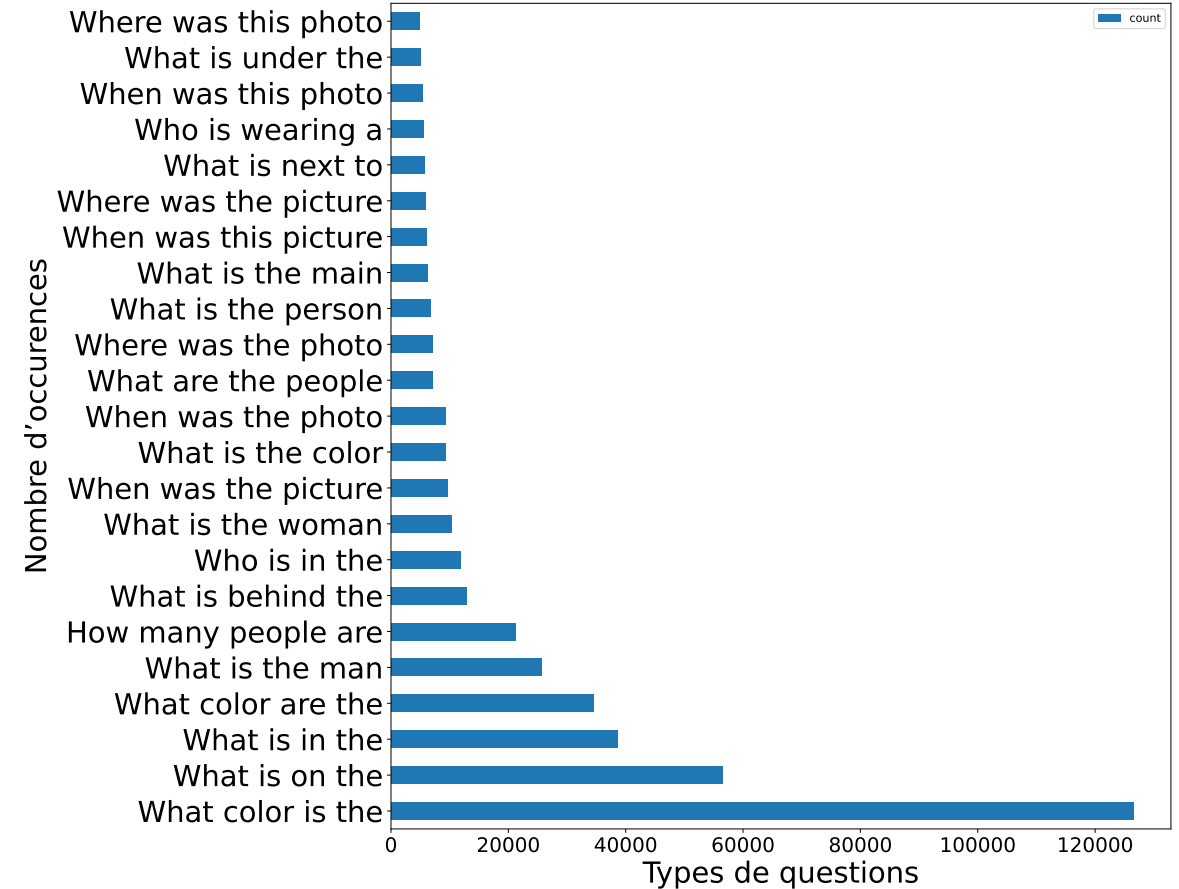


Types de questions les plus posées dans les jeux de données

OK-VQA

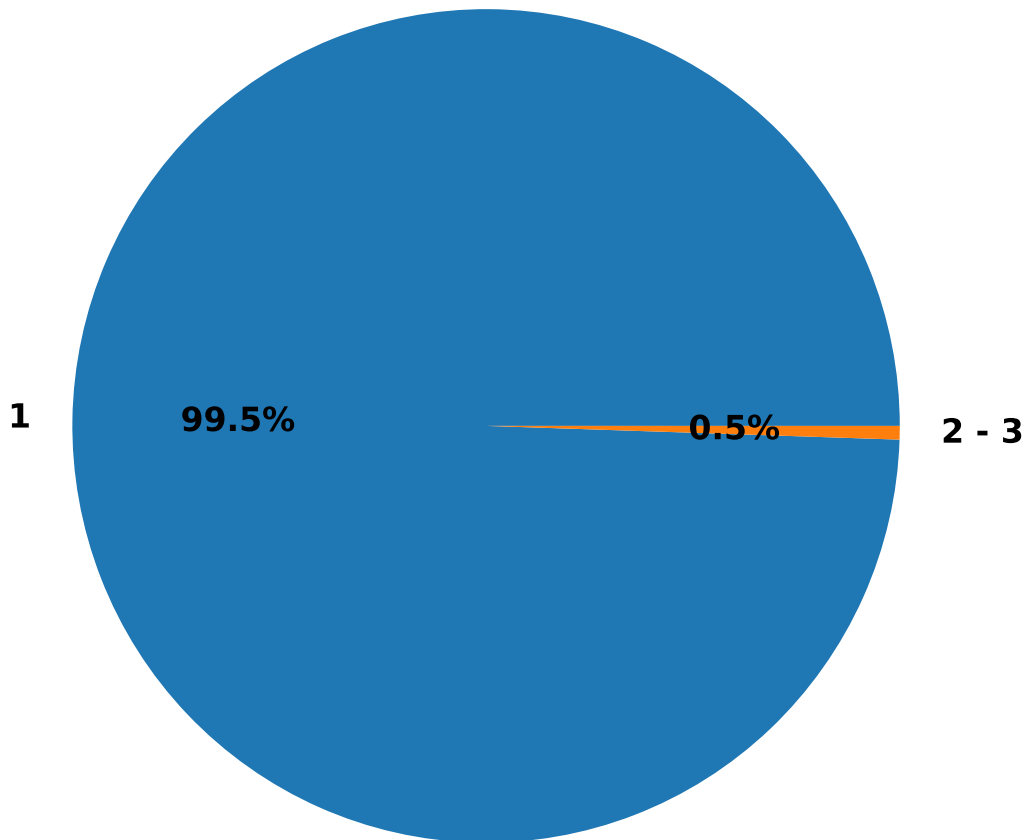


Visual Genome

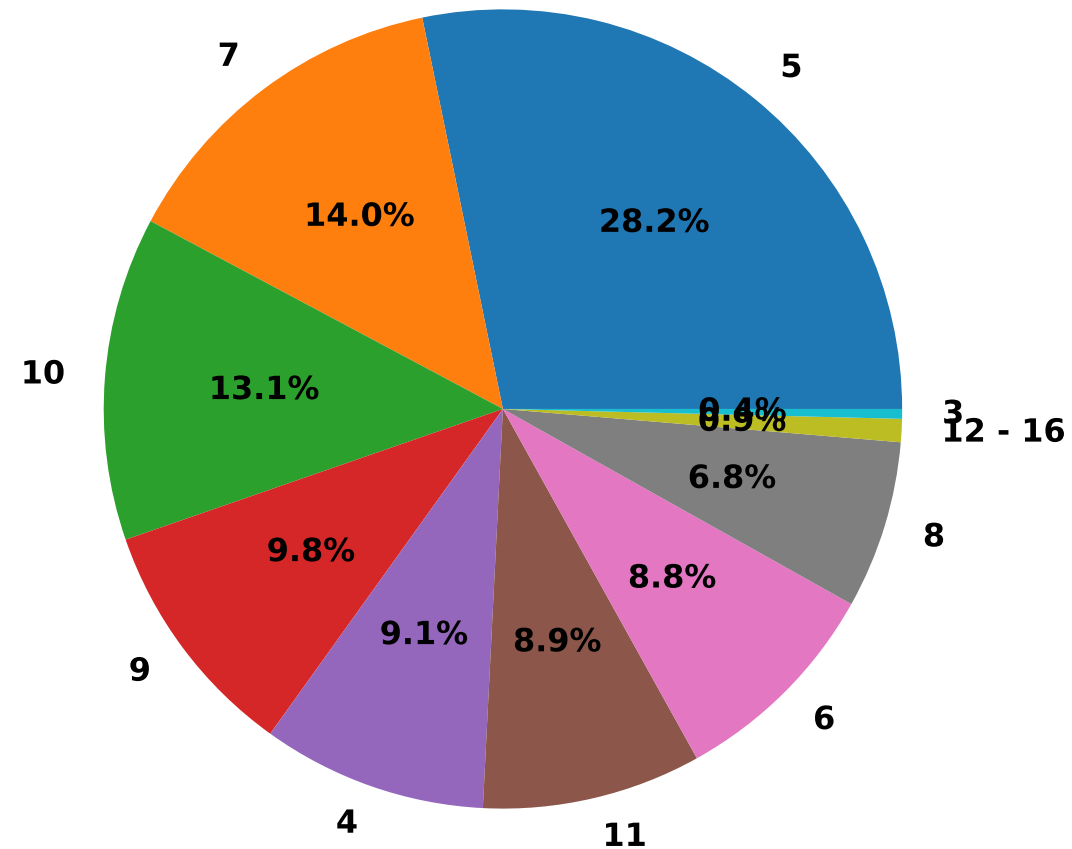


Nombre de mots dans les réponses

GQA : Réponses courtes

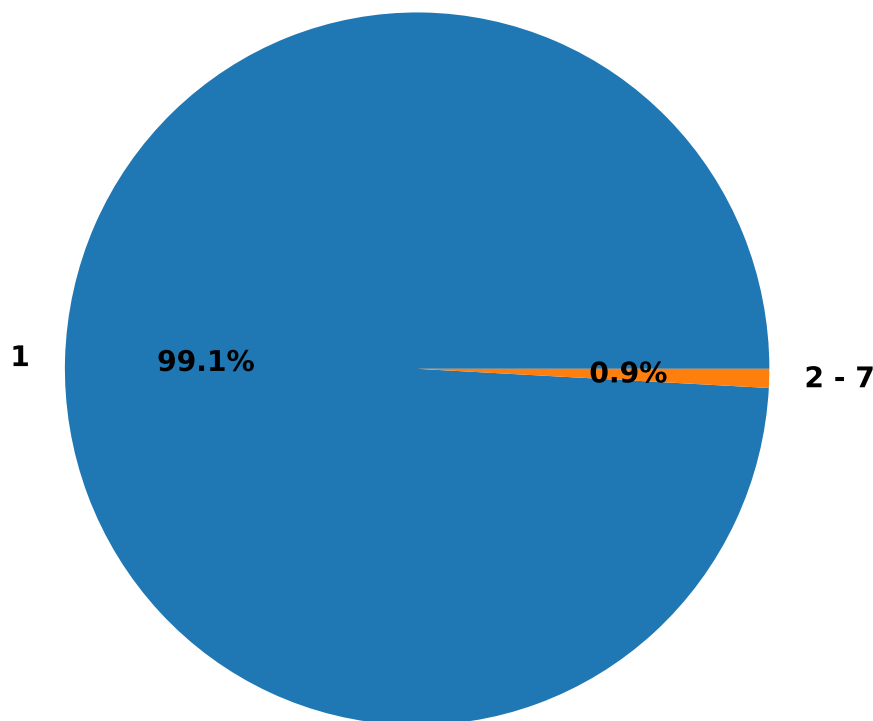


GQA : Réponses longues

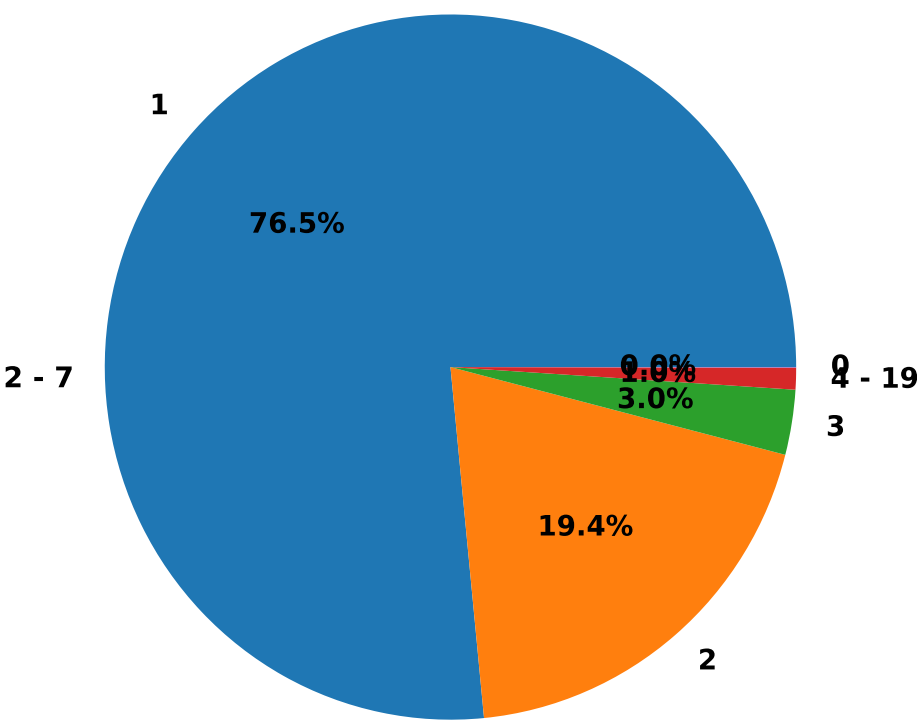


Nombre de mots dans les réponses

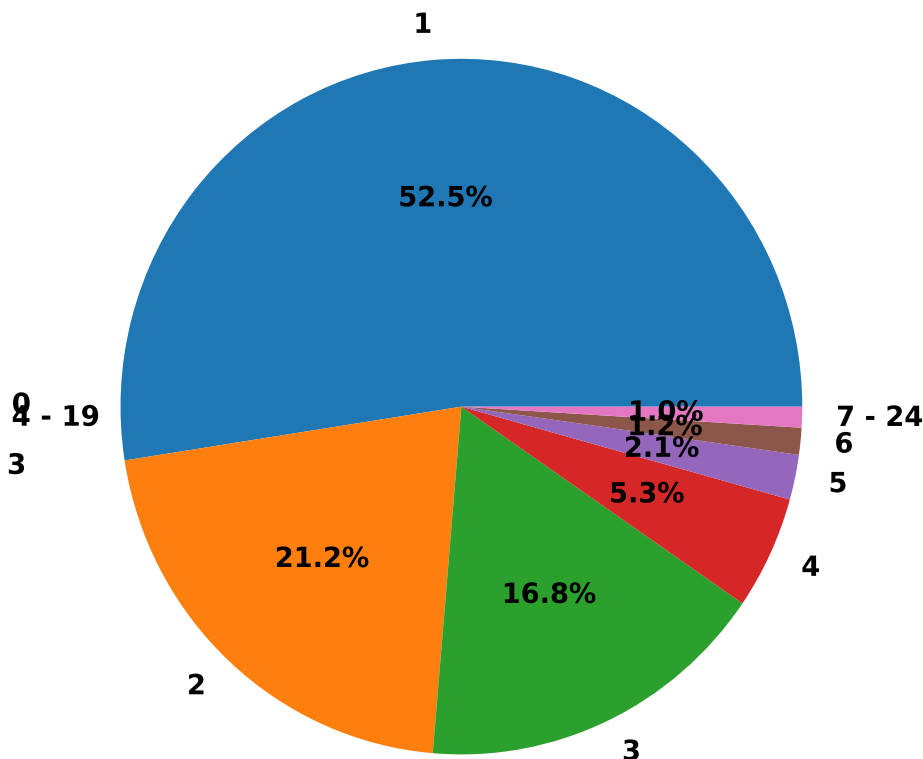
TDIUC



OK-VQA

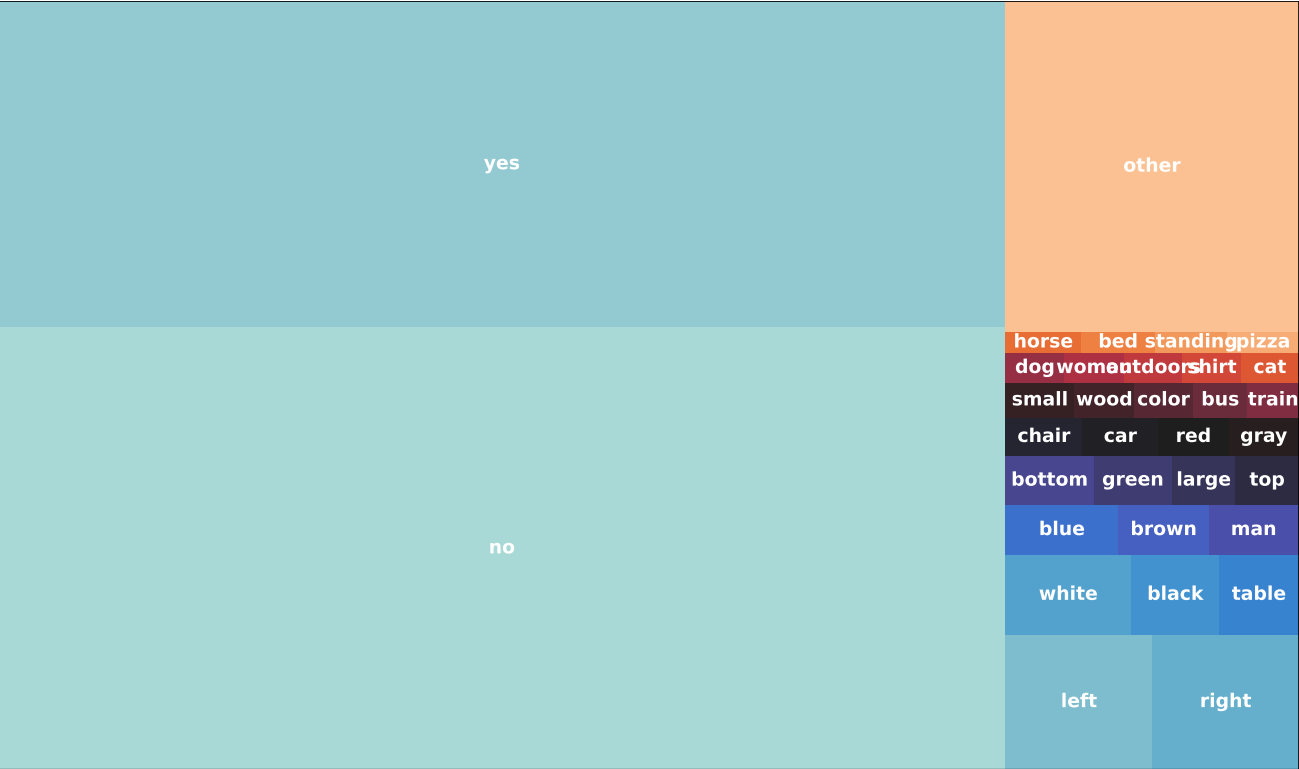


Visual Genome



Répartition des réponses dans les jeux de données

GQA : Réponses courtes



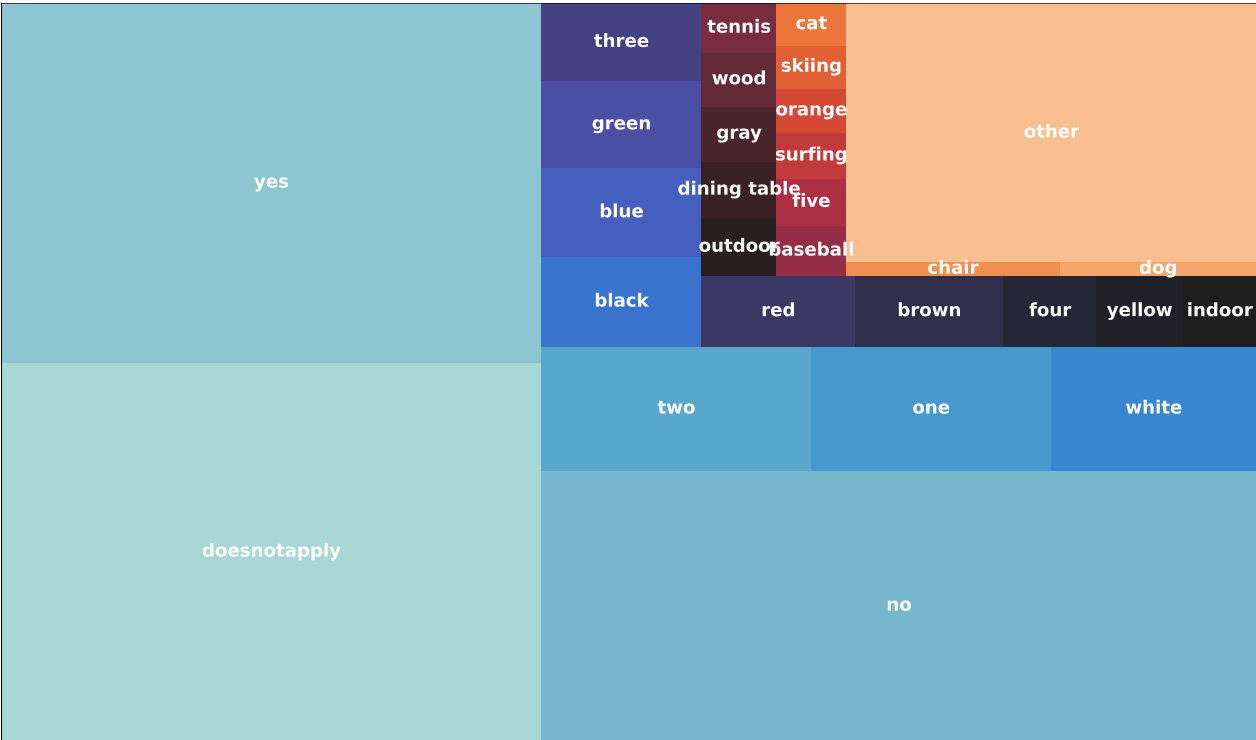
GQA : Réponses longues

Réponses	Nb d'apparitions	Pourcentage d'apparitions
No, there are no fences.	219099	1.53
No, there are no cars.	78880	0.55
Yes, there is a window.	60311	0.42
No, there are no glasses.	47528	0.33
No, there are no helmets.	43327	0.30
No, there are no fences or cars.	36227	0.25
The vehicle is a car.	35888	0.25
Yes, there is grass.	35585	0.24
No, there are no cars or fences.	34614	0.24
No, there are no chairs	33826	0.23
The piece of furniture is a chair.	30121	0.21

⚠ Réponses qui semblent très structurées

Répartition des réponses dans les jeux de données

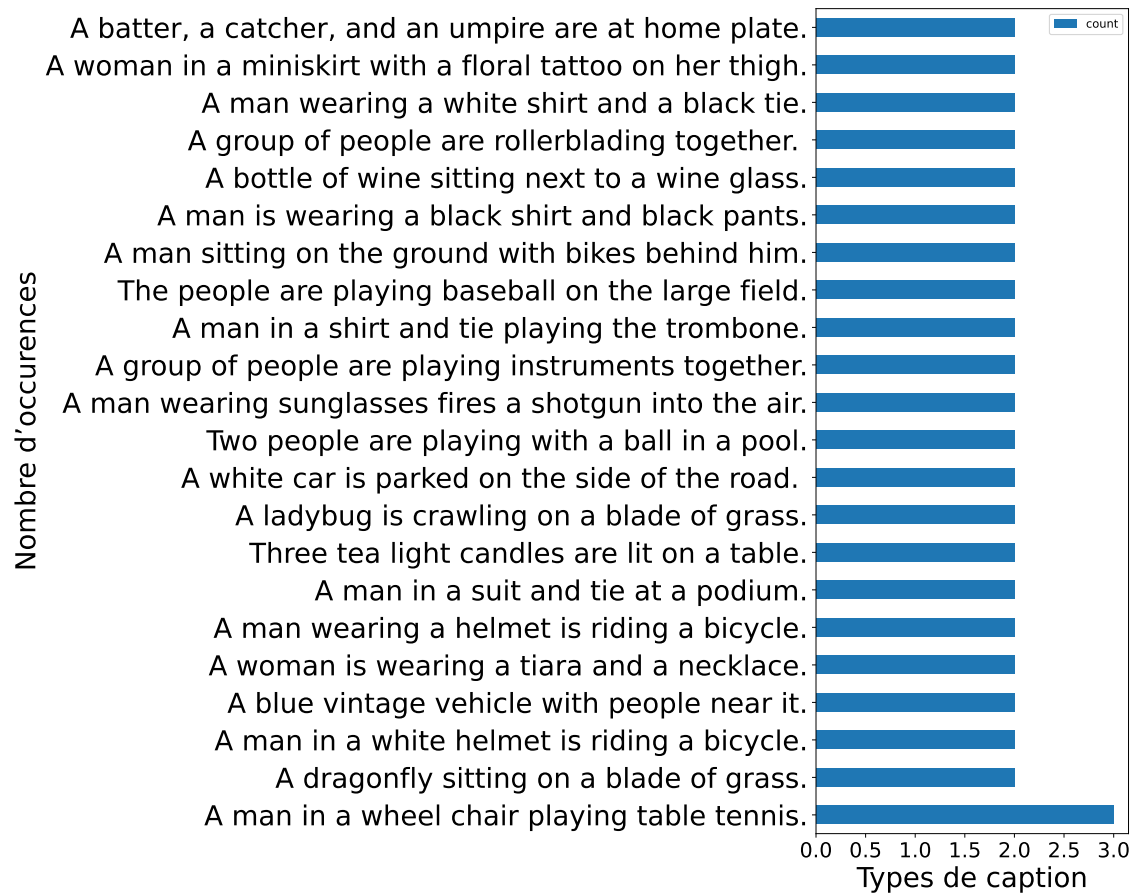
TDIUC



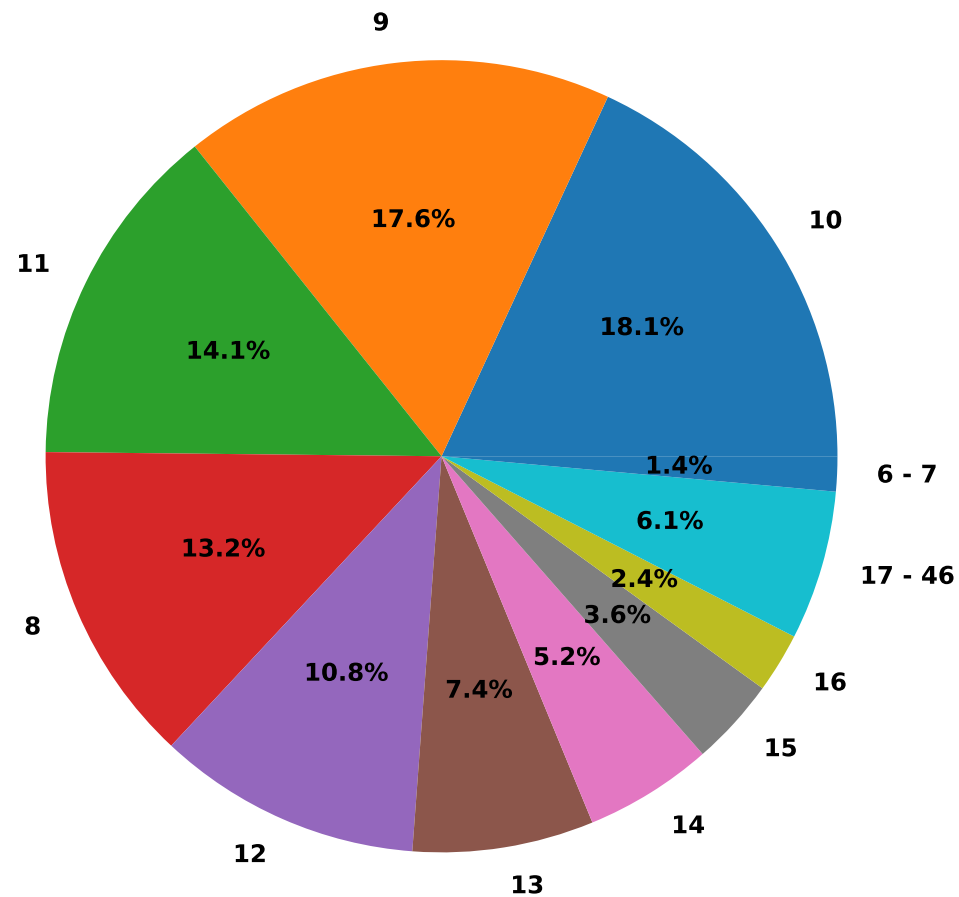
Visual Genome



Jeux de données NoCaps



Descriptions les plus courantes dans le jeu de données



Nombre de mots dans les descriptions