UNIVERSITÉ MOHAMMED V DE RABAT

Faculté des Sciences

Département Physique

Master Informatique et télécommunications

ML/DL : PROJET 3

Réalisé par :

IDRISSI Youssef

HAMIDI Souhail

Encadré par :

Pr. MAHMOUDI Abdelhak

Année universitaire : 2024 -2025

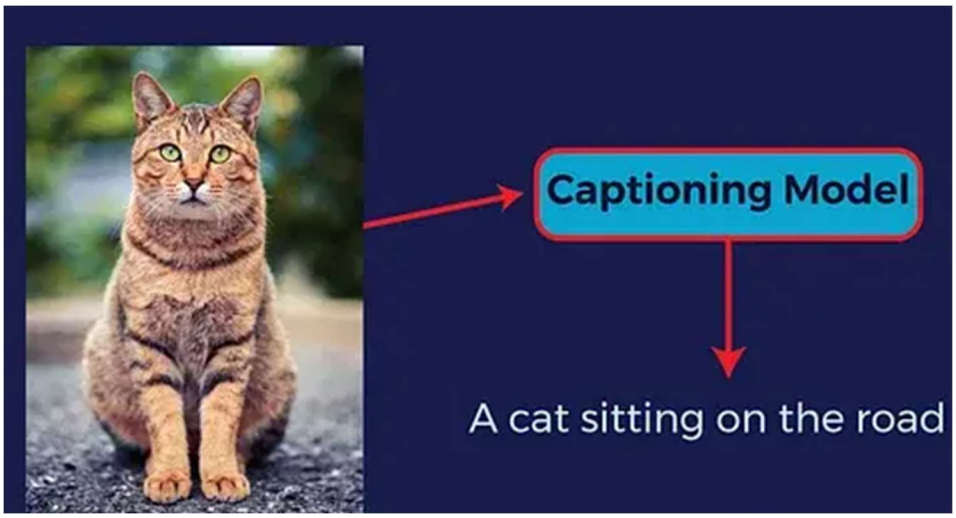
SOMMAIRE

Résumé

Dans le cadre du module ML/DL on travaille sur Ce projet a pour objectif de créer un modèle de deep learning capable de produire automatiquement une description textuelle (ou légende) pour une image donnée par un utilisateur. En utilisant un réseau de neurones convolutionnel (CNN) pré-entraîné pour extraire les caractéristiques visuelles de l'image donnée, associé à un réseau récurrent (LSTM) pour formuler des phrases, on a utilisé dataset Flickr8k (contient presque 8000 images avec descriptions). Les résultats révèlent une capacité de génération assez cohérente, bien qu'il existe certaines limites en ce qui concerne la précision lexicale.

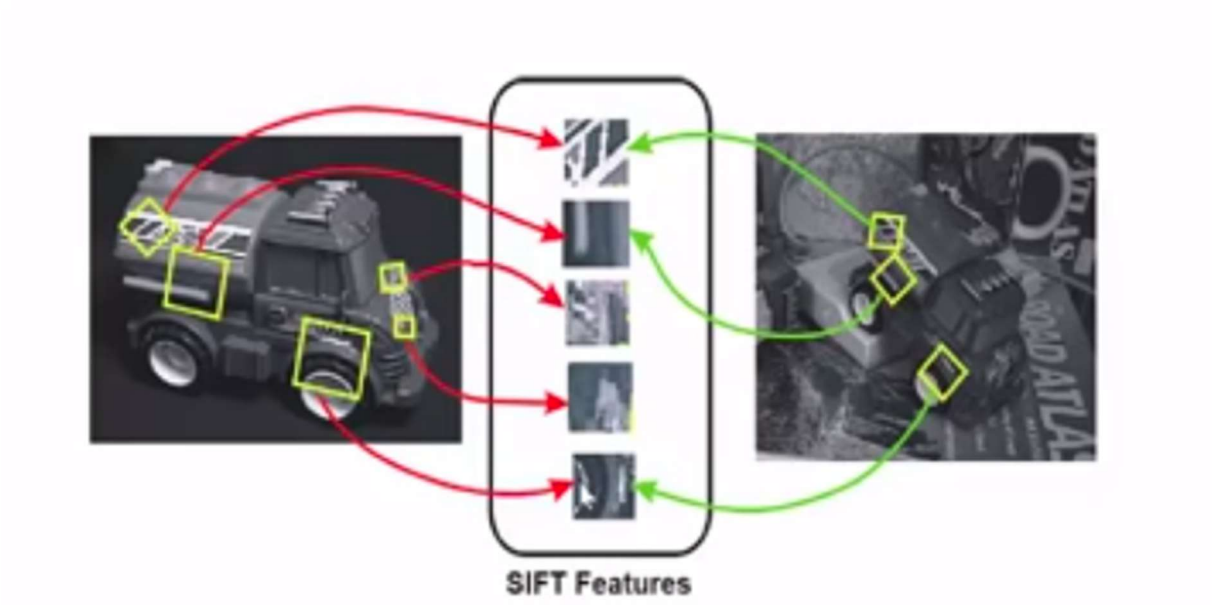
Introduction

La génération automatique de la description pour des images, également appelée "image captioning", représente un domaine à la croisée de la vision par ordinateur et du traitement automatique du langage naturel (TALN). Cette tâche vise à produire des descriptions textuelles pertinentes pour des images données par l'utilisateur, facilitant ainsi l'accessibilité aux contenus numériques pour les personnes malvoyantes, améliorant les systèmes d'indexation d'images dans les moteurs de recherche, ou encore enrichissant les interactions homme-machine. Ce rapport présente une démarche complète de développement d'un modèle d'image captioning entraîné spécifiquement sur le dataset Flickr8k, et illustré via une application web interactive réalisée avec Streamlit.



État de l'art

Historiquement, des approches traditionnelles comme SIFT (Scale-Invariant Feature Transform) ou les représentations Bag-of-Words ont été utilisées pour extraire des caractéristiques visuelles. Cependant, ces méthodes ont rapidement montré leurs limites en termes d'efficacité et de précision sémantique. Le deep learning a véritablement transformé ce domaine avec des modèles comme "Show and Tell" développé par Google, qui associe un réseau convolutionnel (CNN) pour l'extraction des caractéristiques visuelles et un réseau récurrent de type LSTM pour la génération de texte. Aujourd'hui, les architectures basées sur les transformers, telles que ViT (Vision Transformer) et GPT, offrent des performances remarquables grâce à leur capacité à gérer des contextes longs et complexes, bien qu'elles nécessitent des ressources computationnelles importantes.



Méthodologie

- ***Dataset :***
  - Le modèle a été entraîné sur le dataset Flickr8k, composé de 8091 images accompagnées chacune de cinq descriptions dans un fichier texte nommé captions.txt donnée par des humains, offrant ainsi une richesse sémantique significative.
- ***Prétraitement :***
  - Redimensionnement systématique des images à une taille standardisée de 224x224 pixels pour garantir une uniformité lors de l'extraction des caractéristiques.
  - Traitement linguistique des légendes comprenant la suppression des ponctuations, la conversion en minuscules et une tokenisation afin de préparer les données textuelles pour l'apprentissage
- ***Construction du vocabulaire***
  - Application d'un seuil de fréquence minimal fixé à 5 occurrences afin de ne retenir que les mots pertinents et fréquemment utilisés.
  - Utilisation de tokens spéciaux pour marquer les débuts (), fins (), les mots inconnus () et le padding () nécessaire à l'alignement des séquences.
- ***Architecture :***
  - CNN : utilisation du réseau ResNet18 pré-entraîné (weights=ResNet18_Weights.DEFAULT) pour extraire efficacement un vecteur caractéristique compact représentant chaque image.
  - Réseau récurrent : un LSTM avec 256 neurones dans sa couche cachée est utilisé pour générer séquentiellement les mots constituant la légende.
  - Embedding des mots : représentation dense des mots avec une taille d'embedding fixée à 256.
  - Couche de sortie : une classification via softmax appliquée sur l'ensemble du vocabulaire défini.

- ***Entraînement :***

    - Le modèle a été entraîné pendant 15 epochs avec l'optimiseur Adam et un taux d'apprentissage de 3e-4.
    - La fonction de perte utilisée est la CrossEntropyLoss, avec une gestion spécifique du padding pour éviter son impact négatif durant l'apprentissage

Implémentation

Lors de l'implémentation, nous avons utilisé plusieurs outils tels que :

- Langage : Python
- Framework : PyTorch, permettant une flexibilité et une efficacité accrue pour l'implémentation des réseaux neuronaux profonds.
- Structure du projet : notebook interactif (image_captioning.ipynb) pour l'entraînement et une interface Streamlit (streamlit_app.py) pour visualiser les résultats de manière interactive.
- Modules Python utilisés : model.py intégrant les classes essentielles telles que ImageCaptioningModel, Vocabulary, FlickrDataset et MyCollate pour faciliter la structuration et la réutilisation du code.
- Compatibilité : développement et entraînement possibles via Google Colab, tandis que l'application Streamlit peut être exécutée localement.
  Dans notre cas nous avons exécuté le code localement

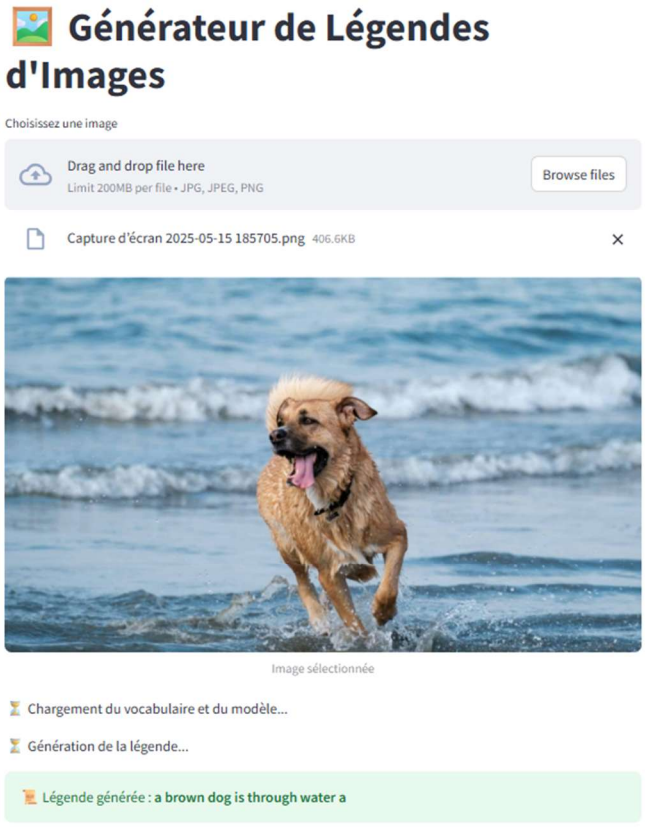
Résultats

Quelques résultats typiques :

- Image : un chien joyeux en train de courir sur une plage
- Légende prédite: « a **brown dog is through water** a"

Problèmes couramment rencontrés :

- Répétitions des termes, nuisant à la fluidité
- Erreurs grammaticales fréquentes
- Génération occasionnelle du token représentant des mots inconnus



Discussion

Le modèle proposé dans ce projet affiche une capacité encourageante à générer des légendes compréhensibles pour des images variées. Cependant, plusieurs limitations importantes ont été identifiées :

Limitations identifiées :

- **Difficulté des LSTM à gérer efficacement la grammaire complexe :**
  Les réseaux LSTM, bien que capables de capturer certaines séquences temporelles, rencontrent des difficultés significatives lorsqu'il s'agit de respecter des structures grammaticales complexes, ce qui conduit souvent à des erreurs syntaxiques ou des phrases incohérentes.
- **Surapprentissage fréquent dû au nombre restreint d'images :**
  Le dataset Flickr8k, bien qu'utile pour des démonstrations initiales, est relativement limité en taille. Ce nombre restreint d'exemples peut causer un surapprentissage, réduisant la capacité du modèle à généraliser efficacement à de nouvelles images.
- **Manque d'évaluation automatisée :**
  L'absence d'évaluation quantitative par des métriques standardisées telles que les scores BLEU, METEOR ou CIDEr, empêche une évaluation objective et systématique de la qualité des légendes générées.

Axes d'amélioration potentiels :

Afin d'améliorer significativement les performances du modèle, plusieurs axes de développement peuvent être envisagés :

- **Adopter des architectures à base de Transformers :**
  Le remplacement du réseau LSTM par des modèles basés sur des transformers, tels que BERT pour le traitement linguistique et ViT (Vision Transformer) pour l'analyse visuelle, permettrait une meilleure compréhension contextuelle et une génération plus précise de légendes.
- **Utilisation de datasets plus importants :**
  Entraîner le modèle sur des jeux de données plus volumineux et diversifiés tels que MS-COCO pourrait renforcer sa robustesse et réduire les risques de surapprentissage, améliorant ainsi sa généralisation.
- **Augmentation des données :**
  L'application de techniques d'augmentation de données, comme la rotation, l'ajout de bruit visuel ou encore l'utilisation de synonymes pour enrichir la variété linguistique des légendes, permettrait d'accroître la quantité effective d'informations disponibles pour l'apprentissage et ainsi améliorer les capacités prédictives du modèle.

Conclusion

Ce projet a été une belle occasion d'explorer et d'intégrer de manière efficace des techniques de vision par ordinateur et de traitement automatique du langage naturel dans un même modèle. Bien qu'il soit capable de produire des légendes assez cohérentes, le modèle présente encore des imperfections notables, surtout en raison des limites inhérentes aux LSTM et de la taille restreinte du dataset utilisé. L'avenir semble prometteur avec l'intégration d'architectures plus avancées comme les transformers, qui devraient apporter des améliorations significatives en termes de précision et de fluidité linguistique.

Bibliographies

Vinyals et al. (2015), "Show and Tell: A Neural Image Caption Generator"
Documentation officielle de PyTorch
Dataset Flickr8k : https://www.kaggle.com/datasets/adityajn105/flickr8k
Ressources additionnelles sur le site Papers With Code, section Image Captioning