Université Chouaib Doukkali
Ecole Nationale des Sciences Appliquées
d'El Jadida

جامعة شعيب الدكالي
universite chouaib doukkali

Compte Rendu TP Attention - GenAI

Nom & Prénom: MAJDALANE Soufiane

Niveau universitaire: 2ème année Master SDIA

Année universitaire: 2025/2026

Objectif du TP

L'objectif de ce TP est de concevoir un système capable de générer une description textuelle (caption) à partir d'une image.
Nous avons combiné trois concepts majeurs :

- **Transfer Learning :** Utilisation d'un **ResNet50** pré-entraîné pour l'extraction de caractéristiques visuelles.
- **Traitement Séquentiel :** Un réseau de neurones récurrent (**LSTM**) pour générer la phrase mot par mot.
- **Mécanisme d'Attention :** Un module permettant au décodeur de se focaliser sur des zones spécifiques de l'image à chaque étape de la génération.

Prétraitement et Chargement du Dataset

- **Dataset :** Utilisation de **Flickr30k**. Le téléchargement a été effectué via `kagglehub` pour pallier l'absence des fichiers locaux.
- **Nettoyage :** Implémentation d'un "parser" robuste pour gérer le fichiers de légendes `captions.txt` qui présentaient des formats hétérogènes (séparateurs `|` ou `,`), évitant ainsi les erreurs de chargement.

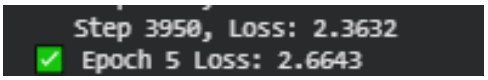
- **Transformations :** Redimensionnement des images en (224, 224) et normalisation (moyenne/écart-type ImageNet) pour correspondre à l'entrée attendue par ResNet6.
- **Embeddings :** Utilisation de vecteurs **Word2Vec** (GoogleNews negative300) pour initialiser la couche d'embedding, facilitant la convergence par rapport à une initialisation aléatoire.

Entraînement et Hyperparamètres

- **Fonction de perte :** `CrossEntropyLoss` (avec `ignore_index` pour le padding).
- **Optimiseur :** Adam avec un Learning Rate de 0.001.
- **Scheduler :** `StepLR` pour réduire le LR après un certain nombre d'epochs13.
- **Environnement :** Exécution sur Google Colab avec GPU (T4).

Résultats et Analyse

- **Performance :**
  Au terme de l'entraînement, j'ai atteint une **validation loss de 2.66**. Ce résultat indique une bonne convergence du modèle, démontrant que le décodeur LSTM parvient à généraliser les relations entre les caractéristiques visuelles extraites par le ResNet50 et les séquences de mots issues des embeddings Word2Vec.



- **Génération :**
  J'ai évalué la qualité du modèle par son aptitude à produire des descriptions sémantiquement cohérentes sur des données jamais vues. Pour cette phase d'inférence, j'utilise l'ensemble de test (20% du dataset initial). Le modèle reçoit une image en entrée et génère une légende mot par mot jusqu'à l'apparition du token de fin `<end>`

True: A red-uniformed hockey player is attempting to control the puck while two white-suited hockey players try to disrupt him .
Generated: hockey player in red uniform is playing game

True: A puppy plays with an adult dog in the snow .
Generated: dog is playing in the snow

Difficultés Rencontrées et Solutions

- **Disponibilité des données :** Le lien Word2Vec original était mort (erreur 404). Solution : Téléchargement via une source Kaggle alternative.
- **Parsing des légendes :** Le fichier de légendes contenait des erreurs de formatage causant des `RecursionError`. Solution : Écriture d'une logique de parsing manuelle pour isoler proprement l'ID de l'image.

Conclusion

Ce TP a permis de comprendre l'interaction complexe entre la vision par ordinateur (CNN) et le traitement du langage naturel (RNN). L'ajout de l'attention améliore l'interprétabilité du modèle.