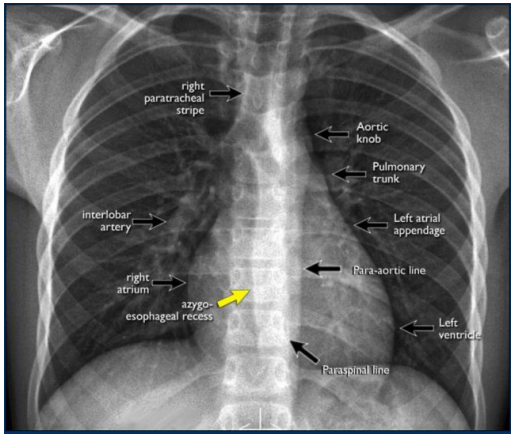
# Fine-tuning Image-Captioning Models for Chest X-ray Interpretation




**Amr MOHAMED - Thu DOAN**

**ING3 - IA - Group 2**

**Deep Learning**

**17/01/2024**

# Sommaire:

# Introduction

# Introduction

- Les récentes avancées dans la **légendage d'images** n'ont pas été exhaustivement appliquées à l'imagerie médicale.
- L'exploitation des techniques avancées de légendage d'images pour **interpréter** et décrire des **images médicales complexes** peut aider les praticiens de santé à mieux diagnostiquer et interpréter les images médicales, **accélérer le processus de diagnostic**, la **planification du traitement**, et **la gestion globale des patients**.
- Notre objectif est de développer un **modèle robuste** capable de générer des **légendes précises et informatives** pour les images radiologiques, visant à **améliorer les processus diagnostiques.**

# Méthodes

# Methods: Dataset

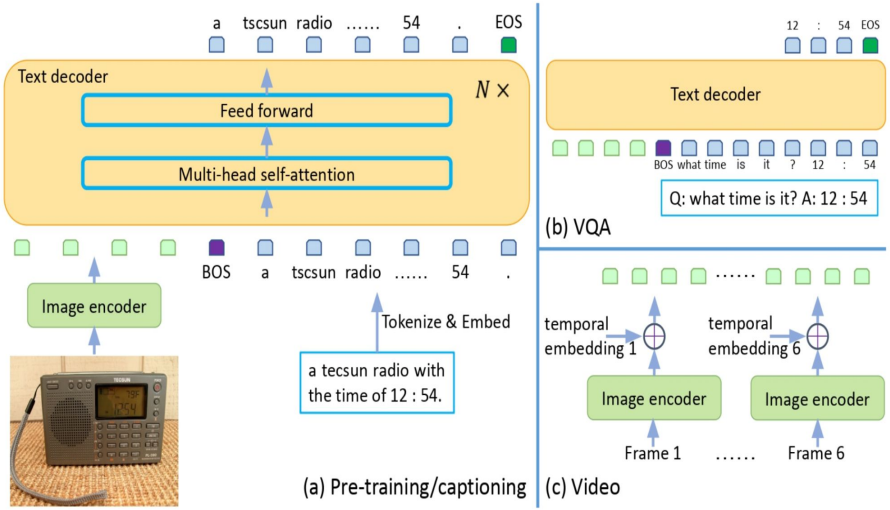
- **ROCO Dataset: Aperçu**
  a. **Objectif :** Conçu pour la image-captioning d'images en imagerie médicale.
  b. **Contenu** : Comprend une variété d'images radiologiques (radiographies, IRM, scanners CT) issues de la littérature médicale, réparties en environ **65 000 pour l'entraînement**, environ **8 200 pour les tests** et environ **8 200 pour la validation**.
  c. **Annotations** : Accompagnées de **textes descriptifs pour chaque image**, fournissant des informations détaillées sur les conditions médicales et les techniques d'imagerie.

# Méthodes: Data preprocessing

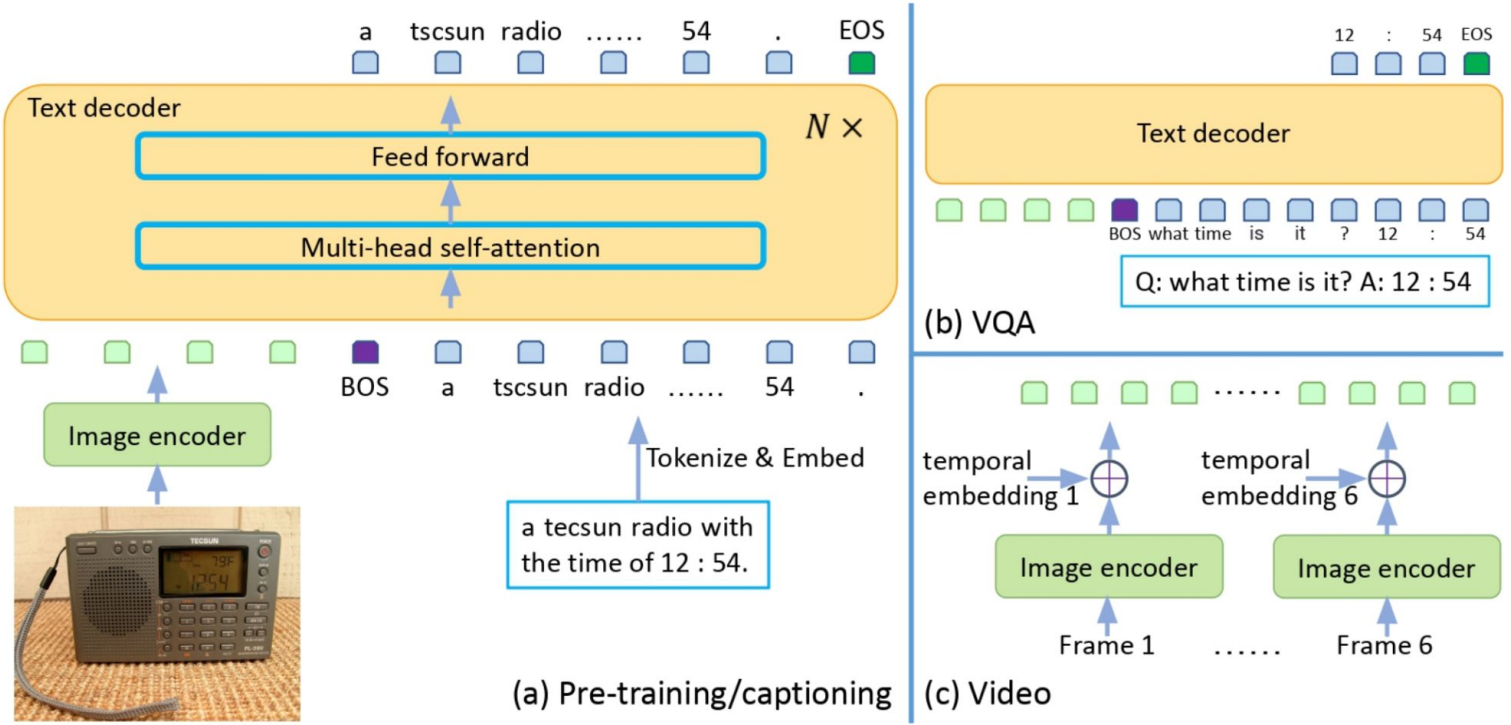
1. **Filtrage des données : La sélection exclusive d'images de radiographie thoracique** a entraîné une réduction de la taille des données à :
    a. Images **d'entraînement** : ~1,7k
    b. Images de **test** : ~200
    c. Images de **validation** : ~200
2. **Prétraitement des images :** (Adapté aux configurations de chaque modèle)
        i. **Redimensionnées** à **224 x 244 x 3** pour **GIT** et **384 x 384 x 3** pour **BLIP**
        ii. **Mises à l'échelle** par un facteur de **1/255**
        iii. **Normalisées** par leur **moyenne**

# Model: GIT (Generative Image2Text), base-sized

- GIT est un **décodeur Transformer conditionné** à la fois sur les jetons d'image CLIP et les jetons de texte. Le modèle est entraîné en utilisant la méthode de **'teacher forcing'** sur un grand nombre de paires (image, texte).
- L'objectif du modèle est simplement de prédire le **prochain jeton de texte, en tenant compte des jetons d'image et des jetons de texte précédents**

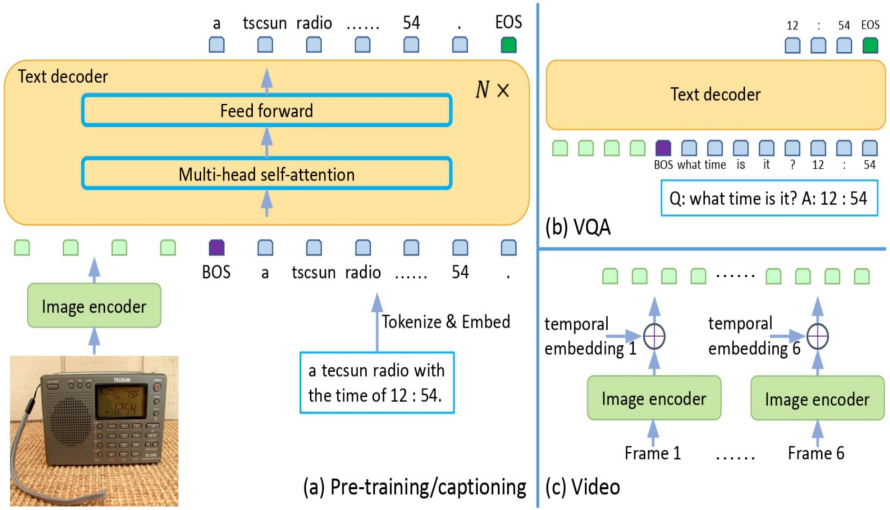


# Model: GIT (Generative Image2Text)



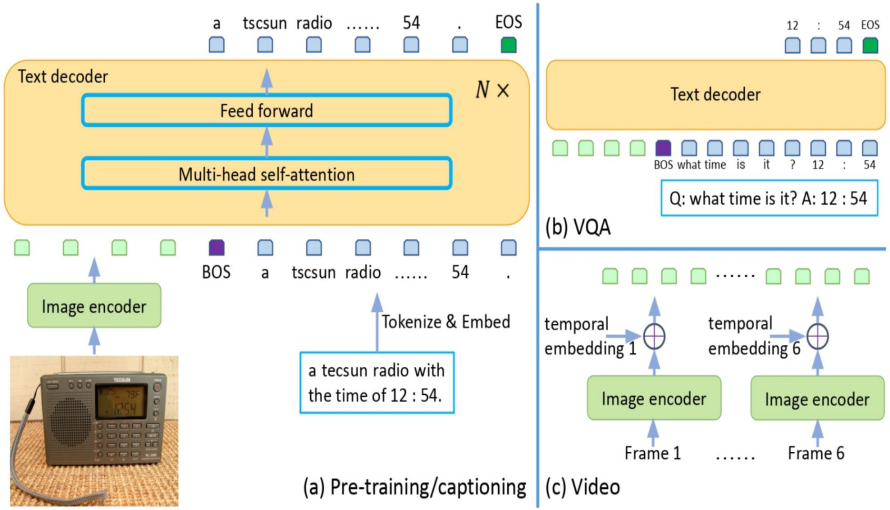
# Model: GIT (Generative Image2Text), base-sized

- **Encoder d'Images :**
  - **Modèle de Base :** L'encodeur d'images est initialement **pré-entraîné** avec des **tâches contrastives**.
  - **Processus :** Il prend une **image brute** et produit une **carte de caractéristiques 2D** compacte. Cette carte est ensuite aplatie en une **liste de caractéristiques.**
  - **Projection :** Ces caractéristiques sont **projetées dans 'D' dimensions** via une couche **linéaire** et une couche de **normalisation**.
  - **Objectif :** Les caractéristiques projetées servent d'entrée pour le **décodeur de texte.**



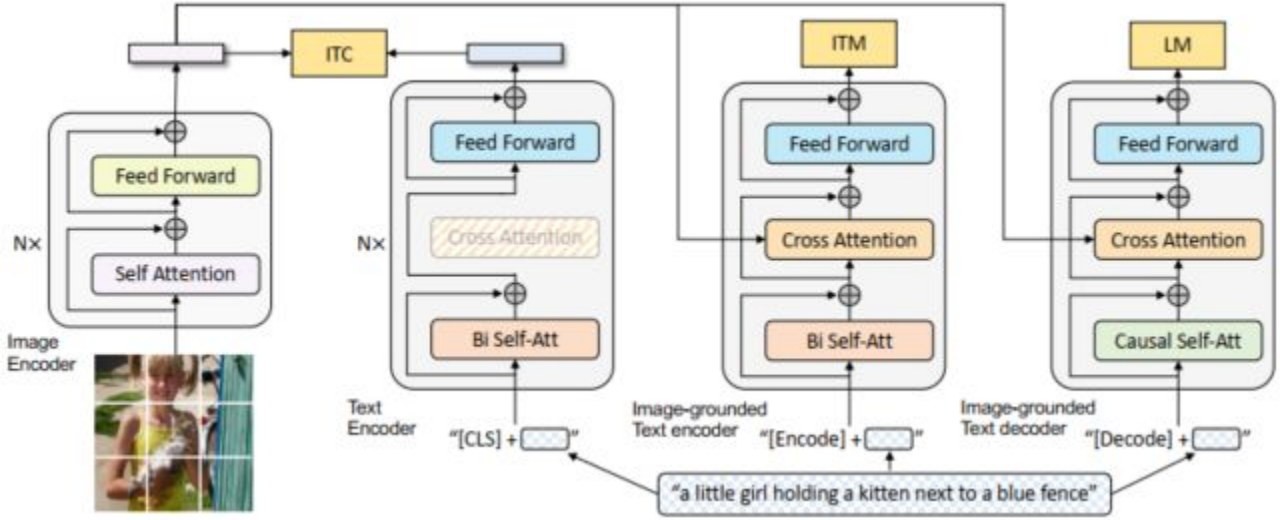
# Model: GIT (Generative Image2Text), base-sized

- **Text Decoder**:
  - **Structure :** composée d'un module de transformation avec plusieurs blocs, chacun contenant une couche **d'auto-attention** et une couche de **rétroaction.**
  - **Processus :** le texte est tokenisé, intégré dans les dimensions « D », ajouté avec un codage positionnel et une couche **layernorm**. Ces intégrations de texte sont concaténées avec des fonctionnalités d'image pour l'entrée du **module de transformation.**
  - **Décodage :** commence par un **jeton [BOS]** et est décodé de manière **auto-régressive** jusqu'à ce qu'un **jeton [EOS]** ou une étape maximale soit atteinte. Il utilise un masque d'attention **seq2seq**.



# Méthodes: Modèle

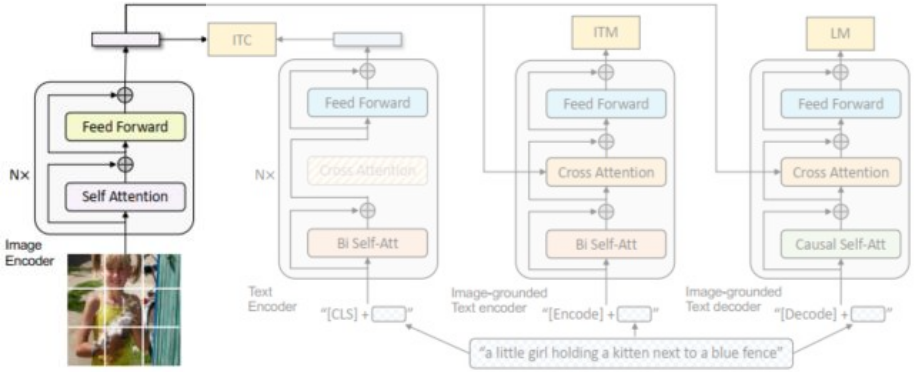
**Salesforce Blip** (Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation)



# Méthodes: Modèle

**Salesforce BLIP: Image Encoder**

BLIP utilise le Vision Transformer (ViT) pour diviser une image d'entrée en patchs et les **encode sous forme d'une séquence d'embeddings.**



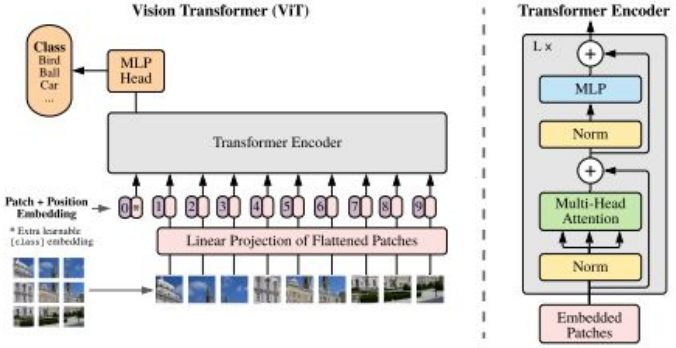
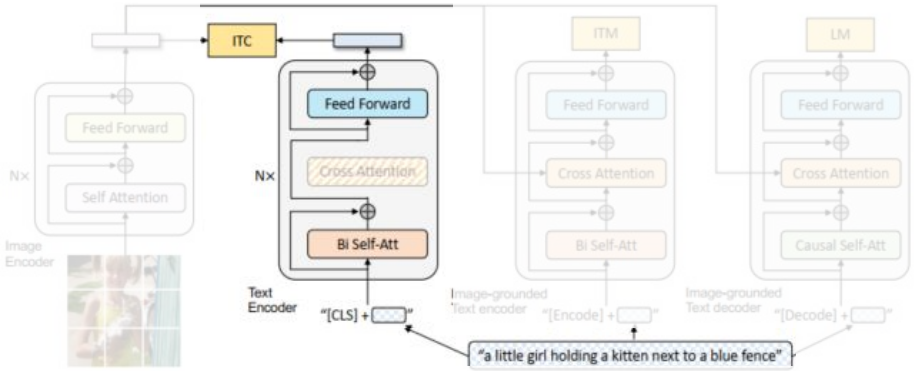




*Image source: An Image is Worth 16x16 Words: Transformers for Image Recognition at*

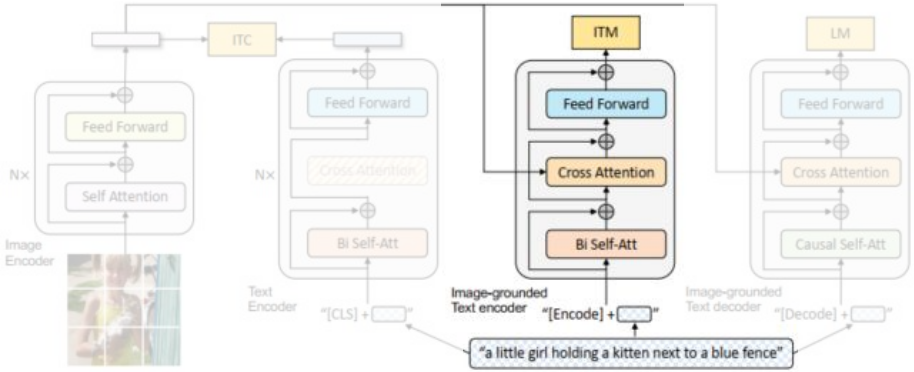
# Méthodes: Modèle

**Salesforce BLIP : Text Encoder**

- L'encodeur de texte **encode séparément l'image et le texte.**
- Il adopte **l'architecture** de **BERT**.
- **Image-Text Contrastive Loss** (ITC) est la fonction de perte pour cette partie du modèle. Son objectif est d'**aligner l'espace des caractéristiques du transformateur visuel et du transformateur textuel** en **encourageant** les paires **positives image-texte** à avoir des **représentations similaires**, contrairement aux paires négatives.



# Méthodes: Modèle



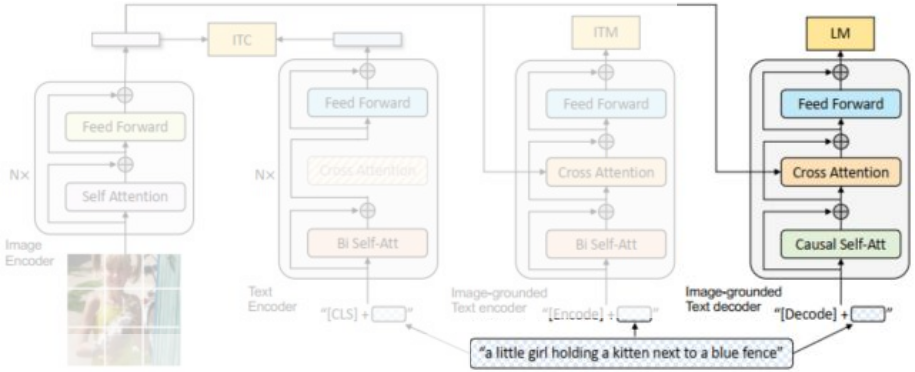

**Salesforce Blip: Image-grounded Text Encoder**

- Injecte des informations visuelles en **insérant une couche de cross-attention supplémentaire (CA) entre la couche de self-attention (SA) et le réseau d'avant-propos (FFN)** pour chaque bloc transformateur de l'encodeur de texte.
- Un **token [Encode] spécifique à la tâche est ajouté au texte**, et **l'embedding** de sortie de [Encode] est **utilisé comme représentation multimodale** de la paire image-texte.

- **Image-Text Matching Loss (ITM)** est minimisée, visant à **apprendre une représentation multimodale image-texte qui capture l'alignement détaillé entre la vision et le langage**. ITM est une tâche de **classification binaire**, où le modèle utilise une tête ITM (une couche linéaire) pour **prédire si une paire image-texte correspondante** est positive ou non.

# Méthodes:  Modèle

## Salesforce BLIP:  Image-grounded text decoder




- **Remplace les couches de self-attention bidirectionnelle** dans l'encodeur de texte basé sur l'image **par des couches de causal self attention** pour favoriser la **génération autorégressive** de légendes.

- **Language Modeling Loss (LM)** vise à **générer des descriptions textuelles à partir d'une image**. Elle **optimise une perte de Cross Entropy** qui entraîne le modèle à maximiser la vraisemblance du texte de manière autorégressive.

# Méthodes: Fine tuning Plan

Pour le Fine-tuning du modèle sur notre dataset, nous avons défini les paramètres suivants :

- **Taux d'apprentissage :** initialement réglé à **5e-5.**
- **Optimiseur :** AdamW
- **Poids de la décroissance :** 1e-08
- **Nombre d'époques d'entraînement : 10** époques
- **Fonction de perte :** Cross Entropy
- **Paramètres efficaces pour GPU :**
    - **Entraînement en précision mixte (fp16) :** L'entraînement a été réalisé en utilisant la précision mixte pour exploiter efficacement la mémoire des GPU de moindre capacité.
    - **Taille de lot d'entraînement par périphérique :** Chaque patch d'entraînement était constitué de 8 échantillons.
    - **Taille de lot d'évaluation par périphérique :** Chaque patch d'évaluation était constitué de 2 échantillons.
    - **Étapes d'accumulation de gradient :** Les gradients ont été accumulés sur 2 étapes avant d'effectuer une passe en arrière.

# Méthodes: **Evaluation metrics**

- **BLEU (Bilingual Evaluation Understudy):** Évalue la qualité du texte traduit par machine par rapport à une référence en mesurant le chevauchement des n-grammes (séquences de mots) entre le texte généré par la machine et les textes de référence.

Precision of the sentence $$P_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Conut_{clip}(n-gram)}{\sum_{\mathscr{C}' \in \{Candidates\}} \sum_{n-gram' \in \mathscr{C}'} Conut(n-gram')}$$

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \log P_n\right)$$

$$BP = \begin{cases} 1 & if\ c < r \\ e^{1-r/c} & if\ c > r \end{cases}.$$

# Méthodes: **Evaluation metrics**

- **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation)**:** Initialement développé comme un package pour l'évaluation de résumés de texte. Le rappel est utilisé pour encourager une description détaillée.
  - ROUGE-1
    - Concentration : Chevauchement des unigrammes (mots individuels) entre le texte généré et la référence.
    - Mesure : Similarité lexicale sur une base mot à mot.
  - ROUGE-2
    - Concentration : Chevauchement des bigrammes (deux mots consécutifs) entre le texte généré et la référence.
    - Mesure : Similarité lexicale au niveau des phrases et cohérence structurelle de base.
  - ROUGE-L
    - Concentration : Plus Longue Sous-séquence Commune (LCS) entre le texte généré et la référence.
    - Mesure : Similarité de la structure au niveau des phrases et ordre des mots.
  - ROUGE-Lsum
    - Variation de ROUGE-L.
    - Concentration : LCS, mais appliquée à chaque phrase séparément avant l'agrégation.
    - Mesure : Plus sensible à la structure et à la cohérence au niveau des phrases dans les résumés multi-phrases.

# Méthodes: **Evaluation metrics**

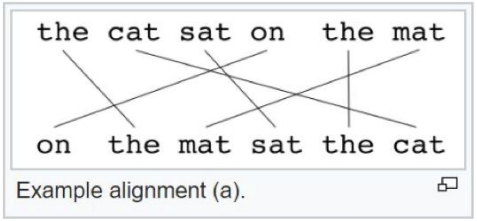
- **Meteor (Metric for Evaluation of Translation with Explicit ORdering):** Il est basé sur une correspondance explicite mot à mot entre la sortie et une la référence. Il peut également correspondre à des synonymes. Calculez le mappage entre la légende candidate et la légende de référence. En cas de conflit, le mappage avec le moins de croisements est sélectionné.

$$P = \frac{m}{w_t} \text{ and } R = \frac{m}{w_r}$$

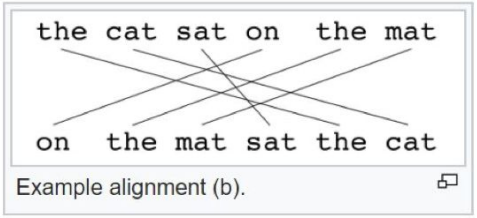
$$F_{mean} = \frac{PR}{\alpha P + (1-\alpha)R}$$ (Harmonic mean)

$$METERO = (1 - pen) \times F_{mean}$$

$$pen = \gamma \left(\frac{ch}{m}\right)^{\theta}$$ (penalty factor)



Example alignment (a).



Example alignment (b).

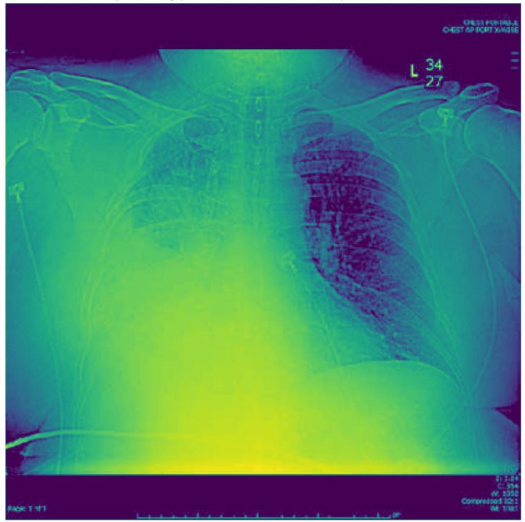
# Résultats

# Résultats: Evaluation Metrics Based

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-Lsum | BLEU | Meteor |
|---|---|---|---|---|---|---|
| Git-base | 0.1 | 0.01 | 0.8 | 0.8 | 0 | 0.05 |
| Git-base-ft | 0.35 | 0.22 | 0.33 | 0.34 | 0 | 0.14 |
| Blip-base | 0.2 | 0.06 | 0.19 | 0.19 | 0 | 0.08 |
| Blip-base-ft | 0.34 | 0.22 | 0.33 | 0.35 | 0.08 | 0.12 |

*Comparison of the different evaluation metrics used for each of the models*

# Résultats: Visualisation des radiographies et des légendes générées.



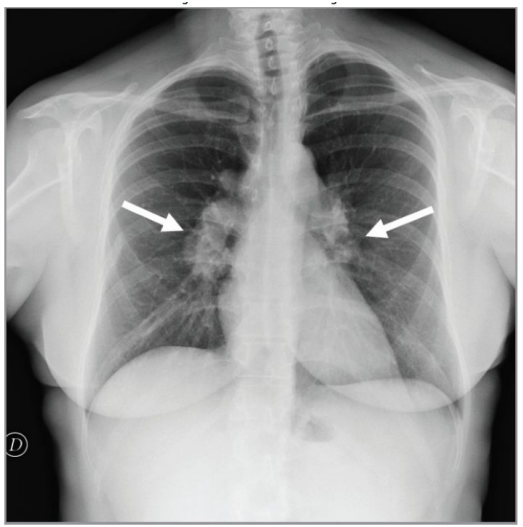
**Original caption**: coronary chest x-ray computed tomography (mediastinal window) showing massive pericardial effusion with an increased pericardial thickness (arrowheads)

**Git base**: what is the name of a pregnant woman?

**Git fine-tuned**: chest x - ray showing a large left - sided pneumothorax.',

**Blip base**: ct scan of the brain'

**Blip fine-tuned**: Chest - x ray showing pneumomediastinum and pneumothorax. arrow



**Original caption**: chest x-ray: multiple **bilateral opacities** and reticular pattern in **both thoracic fields**

**Git base**: a black and white image of human skeleton with a broken chest

**Git fine-tuned**: chest x - ray showing a large mass in the right hemithorax

**Blip base**: a chest with a chest with a chest

**Blip fine-tuned**: Chest - x ray showing a large right - sided mass with a mass - like **opacity** at the right **mid and lower lung zones**

# Discussion

# Discussion

- Blip a **surpassé** les autres en termes de métriques basées sur la sémantique.
- Le modèle doit être utilisé dans le contexte de la **fourniture de suggestions**, plutôt que de prendre des décisions finales.
- Pour des avancées ultérieures, l'engagement d'un **spécialiste du domaine sera crucial**. Cet expert sera responsable du choix minutieux des exemples, de l'évaluation des prédictions du modèle pour leur précision, et de la confirmation de la diversité des données utilisées.

# Discussion

## Limitations et perspectives

- Les limitations étaient centrées autour du manque de **connaissances spécialisées** pour **valider les prédictions du modèle.**
- **Le manque d'accès à des GPU haute performance** pour le Fine tuning de tels modèles d'architecture complexe.
- Les **données** disponibles étaient **limitées**, un manque dans l'expertise nécessaire pour **garantir la diversité de ces données**.
- Les efforts futurs se concentreront sur le développement de **stratégies de Fine Tuning plus complexes et avancées** pour permettre au modèle d'exceller dans le diagnostic, afin que nous puissions **avoir confiance dans les capacités prédictives** du modèle.

# References


- GIT: A Generative Image-to-text Transformer for Vision and Language
- An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
- BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation
- Learning to Evaluate Image Captioning

# Merci!