Université Sidi Mohamed Ben Abdellah — Année universitaire : 2025/2026
Faculté des sciences Dhar El Mahraz — **Master MLAIM**

TP5

Objectif :

Le but de ce TP est de concevoir et d'implémenter un système de **reconnaissance des émotions à partir de vidéos** en utilisant une approche basée sur le modèle **Time-Space Transformer (TimeSformer)**. Ce modèle devra analyser une séquence vidéo d'une personne (par exemple, des extraits de conversation, des scènes d'émotion ou des expressions faciales) et prédire l'émotion de cette personne parmi une série de catégories.

Le système doit être capable de classer une vidéo en fonction des émotions humaines telles que la **joie**, la **tristesse**, la **colère**, la **peur**, la **surprise**, le **dégoût** et un état **neutre**.

Contexte :

Les émotions humaines sont souvent exprimées à travers des indices visuels (expressions faciales, gestes corporels) et peuvent être analysées grâce à des modèles de vision par ordinateur. Les vidéos sont particulièrement riches en informations temporelles et spatiales, ce qui les rend particulièrement adaptées à l'utilisation de modèles comme le **Transformer**.

Le **Time-Space Transformer** (TimeSformer) est une variante du Transformer qui intègre à la fois l'attention spatiale (sur les caractéristiques d'une seule image) et l'attention temporelle (sur l'évolution des images au fil du temps dans une vidéo). Cela le rend particulièrement adapté à des tâches de classification vidéo où les informations visuelles et temporelles sont essentielles.

Étapes du projet :

1. **Collecte des données :**
    - Sélectionner un **dataset vidéo** annoté avec des émotions humaines. Vous pouvez utiliser des datasets populaires comme **EmoReact**, **AffectNet**, **RAVDESS**, ou d'autres datasets de vidéos avec des émotions étiquetées.
    - Les vidéos doivent être étiquetées avec une émotion précise. Les émotions peuvent être parmi les classes suivantes : **joie**, **tristesse**, **colère**, **peur**, **surprise**, **dégoût** et **neutre**.
    - **Traitement des vidéos** : Extraire les frames des vidéos et les prétraiter pour les rendre compatibles avec les exigences du modèle.
2. **Prétraitement des données :**
    - Convertir les vidéos en séquences d'images (frames) à une fréquence d'images donnée (par exemple, 16 images par seconde).
    - Redimensionner les images à une taille uniforme (par exemple, 224x224 pixels) et normaliser les valeurs des pixels.
    - Diviser les données en **ensemble d'entraînement**, **ensemble de validation**, et **ensemble de test**.

3. **Création du modèle Time-Sformer :**
    - Implémenter un modèle **Time-Sformer** en utilisant une architecture basée sur le Transformer, mais adaptée aux vidéos.
    - Le modèle doit comporter deux mécanismes d'attention :
        - **Spatiale** : pour extraire les informations pertinentes à partir des caractéristiques de chaque frame.
        - **Temporelle** : pour capturer les relations entre les frames et traiter les informations sur la durée de la vidéo.
4. **Entraînement du modèle :**
    - Définir les critères d'entraînement, tels que la **fonction de perte** (par exemple, **Cross-Entropy Loss** pour la classification multi-classes).
    - Choisir un **optimiseur** comme **Adam** et définir des hyperparamètres comme le taux d'apprentissage et le nombre d'époques.
    - Entraîner le modèle sur les données d'entraînement et surveiller les performances sur les données de validation.
5. **Évaluation du modèle :**
    - Une fois l'entraînement terminé, évaluer le modèle sur l'ensemble de test pour mesurer sa capacité à prédire les émotions des vidéos.
    - Utiliser des **métriques d'évaluation** telles que la précision, le **F1-score**, la **métrique d'accuracy** ou la **matrice de confusion** pour analyser les résultats du modèle.
6. **Amélioration du modèle :**
    - Effectuer des ajustements sur les hyperparamètres (par exemple, nombre de couches de transformer, taille du batch, etc.).
    - Implémenter des techniques d'**augmentation de données** (par exemple, rotation, zoom, modification de l'éclairage) pour améliorer la robustesse du modèle.