RÉPUBLIQUE TUNISIENNE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

École Nationale d'Ingénieurs de Carthage

RAPPORT DE PROJET ACADÉMIQUE

Vision-LLM pour la Reconnaissance Faciale des Émotions et des Expressions Composées (FER-CE)

Réalisé par l'équipe Meow-AI :

M. Mohamed Dhia Eddine THABET

M. Ala Eddine MADANI

M. Mohamed BEN MADHI

M. Aymen SATOURI

M. Mouhanned DHAHRI

Année Universitaire : 2025 - 2026

Table des matières

Résumé Exécutif **3**

1 Contexte et Problématique **4**

1.1 Introduction 4

1.2 Le Défi des Émotions Composées 4

1.3 Vision Open Source 4

2 Méthodologie et Traitement Avancé **5**

2.1 Le Jeu de Données RAF-CE 5

2.2 Pipeline de Prétraitement Collaboratif 5

2.2.1 Étape 1 : Alignement et Normalisation 6

2.2.2 Étape 2 : Le Défi de la Résolution 6

2.2.3 Étape 3 : Super-Résolution par GAN 6

2.3 Stratégies de Régularisation 6

3 Expérimentations et Analyse **8**

3.1 Limites et Difficultés (Bottlenecks) 8

3.2 Résultats Comparatifs 8

3.2.1 Analyse des Résultats 9

3.3 Interprétabilité et Évaluation 10

3.3.1 Analyse Visuelle : Grad-CAM 10

3.3.2 Évaluation Textuelle : BLEU et ROUGE 10

4 Déploiement et Conclusion **11**

4.1 Architecture Cible (DockerHub) 11

5 Conclusion **12**

Résumé Exécutif

Synthèse du Projet

Ce rapport synthétise les travaux de l'équipe **Meow-AI** sur le développement d'un système de reconnaissance des émotions faciales composées. Face aux limites des modèles classiques (CNN/ViT) sur des datasets complexes et restreints, nous avons développé une approche innovante combinant **Super-Resolution par GAN** et **Vision-LLM**.

Notre projet vise non seulement à bencher les technologies actuelles, mais aussi à proposer une solution Open Source robuste. Après avoir surmonté des défis liés à la qualité des images et aux ressources matérielles limitées, nous avons convergé vers le modèle **Qwen-VL 7B**, atteignant une **Training Loss de 0.007**.

Mots-clés : Vision-LLM, GAN Super-Resolution, Qwen-VL, RAF-CE, Multimodalité, Open Source.

Chapitre 1

Contexte et Problématique

1.1 Introduction

La reconnaissance d'expressions faciales (FER) évolue vers l'analyse des **émotions composées** (*Compound Expressions*), qui reflètent la complexité réelle des interactions humaines.

1.2 Le Défi des Émotions Composées

Les émotions composées posent deux problèmes majeurs :

1. **Ambiguïté visuelle :** Les micro-variations musculaires sont subtiles et partagées entre plusieurs classes.
2. **Manque de contexte :** Un modèle classique manque de capacité de raisonnement pour interpréter ces nuances.

1.3 Vision Open Source

Notre solution ne se veut pas parfaite, mais constitue un benchmark technologique avancé. Nous croyons fermement que *"Open Source is Power"*. Notre objectif à terme est de publier une version optimisée et conteneurisée sur **DockerHub**, rendant ces modèles accessibles à la communauté.

Chapitre 2

Méthodologie et Traitement Avancé

2.1 Le Jeu de Données RAF-CE

Nous travaillons sur le dataset **RAF-CE** (4 500 images), comportant 15 classes d'émotions composées.

Mapping des 15 Classes

- 0: Happily surprised
- 1: Happily disgusted
- 2: Sadly fearful
- 3: Sadly angry
- 4: Sadly surprised
- 5: Sadly disgusted
- 6: Fearfully angry
- 7: Fearfully surprised
- 8: Fearfully disgusted
- 9: Angrily surprised
- 10: Angrily disgusted
- 11: Disgustedly surprised
- 12: Happily fearful
- 13: Happily angry
- 14: Happily sad

2.2 Pipeline de Prétraitement Collaboratif

La qualité des données étant critique, notre équipe a travaillé collectivement sur plusieurs approches de traitement d'image pour optimiser les performances des modèles.

2.2.1 Étape 1 : Alignement et Normalisation

Les images brutes variaient énormément en taille et en cadrage. **Nous avons standardisé** le dataset en détectant et alignant les visages (centrage des yeux), produisant ainsi une base cohérente d'images de dimension 100×100 pixels.

2.2.2 Étape 2 : Le Défi de la Résolution

Les modèles standards comme ResNet sont entraînés sur des entrées de 224×224 pixels. **Nous avons initialement tenté** une mise à l'échelle classique (interpolation) des images de 100×100 vers 224×224.

- **Constat :** Cette approche floutait les bords et les micro-traits essentiels (rides, contour des yeux), dégradant la performance du modèle.

2.2.3 Étape 3 : Super-Résolution par GAN

Pour résoudre ce problème de flou, **l'équipe a implémenté** une approche innovante basée sur l'IA générative :

1. Upscaling des données pré-traitées (100×100) vers 400×400 pixels via un modèle **GAN (Generative Adversarial Network)** renommé pour la restauration d'images.
2. Redimensionnement final (Downscaling) vers les résolutions cibles : 224×224 pour ResNet et $\sim 336 \times 336$ pour le LLM.

Impact

Cette technique a permis de restaurer une netteté cruciale et a été l'un des facteurs déterminants pour l'amélioration significative de nos résultats.

2.3 Stratégies de Régularisation

Pour contrer la petite taille du dataset et le déséquilibre des classes, nous avons appliqué :

- **Dropout :** Désactivation aléatoire de neurones pour forcer le réseau à apprendre des caractéristiques robustes et éviter l'overfitting.
- **Data Augmentation :** Rotations, ajustements de luminosité et flips pour augmenter artificiellement la variété des données.
- **Weighted Loss :** Attribution de poids plus importants aux classes minoritaires (ex: Fearfully Disgusted) dans la fonction de perte pour empêcher le modèle de les ignorer.

Chapitre 3

Expérimentations et Analyse

3.1 Limites et Difficultés (Bottlenecks)

Notre démarche a été confrontée à plusieurs obstacles majeurs :

Volume et Déséquilibre des Données

Le dataset est très petit (~ 4500 images). Malgré l'augmentation de données, le déséquilibre persiste, rendant la prédiction des classes rares difficile. Ce n'est pas une solution parfaite, mais un pas vers le benchmark de nouvelles technologies.

Infrastructure et Temps

Nous manquons de GPU puissants. L'utilisation d'environnements Freemium (Kaggle/Colab) nous a limités en temps de calcul et en mémoire, nous obligeant à utiliser des versions quantizées des modèles.

3.2 Résultats Comparatifs

Nous présentons ici l'intégralité des benchmarks réalisés. Notre équipe a testé six architectures distinctes avant de converger.

Table 3.1: Benchmark Complet des Modèles Testés

Modèle	Type / Approche	Performance	Statut
ResNet-18	Binaire (One-vs-All)	Acc: 43.0%	Échec
ViT Base	Vision Transformer	Acc: 47.9%	Échec (Overfit)
ResNet-50	CNN (Baseline)	Acc: 51.0%	Moyen
VGG-Face	CNN Spécialisé	Acc: 54.0%	Bon
ResNet-18	CNN Multi-classe	**Acc: 55.3%**	Meilleur Vision-Only
BLIP-2	Multimodal	-	Trop Abstrait / Échec
Qwen-VL 7B	**Vision-LLM**	**Loss: 0.007**	**Solution Finale**

3.2.1 Analyse des Résultats

1. **Échec de l'approche Binaire :** La séparation en 15 classifieurs binaires (43%) a échoué, prouvant que les émotions composées partagent trop de traits communs pour être isolées brutalement.

2. **Surprise ResNet-18 :** Contre toute attente, le modèle ResNet-18 Multi-classe (55.3%) a surpassé le ResNet-50 et le VGG-Face. Sa structure plus légère a probablement évité l'overfitting sur notre petit dataset, surtout après l'amélioration des images par GAN.

3. **Limites de BLIP-2 :** Ce modèle a produit des descriptions trop vagues, incapables de capturer les micro-expressions nécessaires.

4. **Suprématie Qwen-VL :** Avec une perte quasi-nulle (0.007), ce modèle montre qu'il a "compris" la tâche, là où les modèles de classification pure plafonnent vers 55%.

3.3 Interprétabilité et Évaluation

3.3.1 Analyse Visuelle : Grad-CAM

Pour valider que nos modèles ne sont pas des ”boîtes noires”, nous avons appliqué l'algorithme **Grad-CAM** sur le ResNet-18. L'analyse des heatmaps générées révèle une corrélation physiologique forte : le modèle se focalise systématiquement sur les **Yeux**, les **Sourcils** et la **Bouche**, confirmant la pertinence de ses décisions.

3.3.2 Évaluation Textuelle : BLEU et ROUGE

Au-delà de la classification visuelle, la plus-value du Vision-LLM (Qwen-VL) réside dans sa capacité à générer des explications. Pour évaluer la qualité de ces descriptions générées, nous nous référons à deux métriques standards du Traitement Automatique du Langage (NLP) :

- **BLEU (Bilingual Evaluation Understudy) :** Cette métrique évalue la précision. Elle vérifie si les mots employés par le modèle (ex: ”sourcils froncés”, ”bouche ouverte”) correspondent au vocabulaire des annotations de référence.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation) :** Cette métrique évalue le rappel. Elle vérifie si le modèle a bien capturé *toute* l'information importante de la référence, sans omettre de détails cruciaux sur l'expression faciale.

Ces scores nous permettent de quantifier objectivement la qualité des ”explications” fournies par notre système, assurant qu'elles sont non seulement plausibles mais aussi fidèles à la réalité émotionnelle de l'image.

Chapitre 4

Déploiement et Conclusion

4.1 Architecture Cible (DockerHub)

Dans l'esprit "Open Source", nous avons conteneurisé notre solution.

Stack Technique

Service API Conteneur Python exposant le modèle via une API REST.

Interface Prototype Streamlit.

Conteneurisation Docker pour garantir la reproductibilité.

Chapitre 5

Conclusion

Le projet **Meow-AI** démontre que l'innovation (Vision-LLM, GAN Super-Resolution) permet de compenser partiellement le manque de données et de ressources.

1. Le pipeline de prétraitement collaboratif (Alignement $\rightarrow$ GAN Upscale) a été un facteur clé de succès.
2. **ResNet-18 (55.3%)** reste une option légère solide, surpassant les modèles plus lourds.
3. **Qwen-VL 7B** offre une compréhension supérieure (**Loss 0.007**) et explicable.

Nos résultats, bien que contraints par l'infrastructure, sont prometteurs et ouvrent la voie à des outils Open Source plus performants.