INP ENSEEIHT

Institut national polytechnique en France

L'École nationale supérieure d'électrotechnique, d'électronique, d'informatique, d'hydraulique et des télécommunications

Deuxième année

Reconnaissance d'emotions faciales

(Apprentissage Profond - Projet)

Rapport du projet

Projet réalisé par *(Ayoub Bouchama)*
et *(Oussama Elguerraoui)*

Dirigé par *(Axel Carlier)*

Lien utile vers les page des realisateurs du projet :

Ayoub Bouchama, *2A Génie Logiciel, Groupe L12 (Profil LinkedIn)*
Oussama ElGuerraoui, *2A Génie Logiciel, Groupe L12 (Profil LinkedIn)*

Table des matières

Table des figures

1 Introduction

FIGURE 1 – Images de plusieurs emotions sur un visage humain

La reconnaissance des émotions faciales est une discipline fascinante et cruciale dans le domaine de l'intelligence artificielle (IA) et de la vision par ordinateur. Elle vise à détecter, analyser et interpréter les expressions émotionnelles humaines à partir d'images ou de vidéos, ouvrant ainsi la voie à une compréhension plus profonde et plus précise des interactions humaines.

L'importance de la reconnaissance des émotions faciales réside dans son potentiel à améliorer de nombreux aspects de notre vie quotidienne, de la technologie à la santé mentale en passant par les interactions sociales. En effet, la capacité des machines à comprendre et à réagir aux émotions humaines peut conduire à des avancées significatives dans des domaines tels que la reconnaissance automatique d'émotions dans les interactions homme-machine, l'amélioration des interfaces utilisateur, la détection précoce des troubles mentaux, la personnalisation des services et bien d'autres.

Cependant, malgré ses nombreuses applications potentielles, la reconnaissance des émotions faciales est confrontée à des défis majeurs. Parmi ceux-ci, on trouve la variabilité interindividuelle des expressions faciales, les différences culturelles dans l'expression des émotions, le bruit et la qualité des données, ainsi que la nécessité de développer des modèles robustes capables de généraliser efficacement à partir de données limitées.

Dans ce contexte, l'intelligence artificielle joue un rôle essentiel en fournissant des outils et des techniques avancés pour relever ces défis. Les modèles d'apprentissage automatique, en particulier les réseaux de neurones profonds, ont démontré leur capacité à apprendre des caractéristiques discriminantes à partir de données complexes et à réaliser des tâches de reconnaissance des émotions faciales avec une précision remarquable.

Dans ce rapport, nous explorerons en détail les différentes approches que nous avons suivies pour résoudre le problème de la reconnaissance des émotions faciales, de la collecte et du prétraitement des données à la modélisation et à l'évaluation des performances. Nous analyserons également les résultats obtenus, en mettant en lumière les défis rencontrés et les leçons apprises tout au long du processus. Enfin, nous discuterons des implications de notre travail et des pistes de recherche futures dans ce domaine passionnant de l'IA.

2 Méthodologie

Dans notre projet, nous avons appliqué les connaissances acquises lors des travaux pratiques, en particulier lors du TP3, pour effectuer le traitement des données et concevoir des réseaux de neurones convolutifs capables de reconnaître les émotions faciales sur des images humaines. Nous avons également exploré différentes méthodes pour améliorer l'entraînement des modèles. Notre objectif principal était d'optimiser la précision de la courbe de validation après l'entraînement afin d'assurer une détection efficace des émotions faciales. Initialement, nous avons travaillé avec six classes d'émotions, mais avons constaté que la précision ne dépassait pas 30%. Par conséquent, nous avons envisagé de réduire le nombre de classes, d'améliorer la précision à l'aide de différentes techniques, puis d'ajouter d'autres classes par la suite.

Ainsi, nous avons implémenté les différents modèles CNN vus lors du TP3, notamment un **modèle simple** ainsi que le **VGG19**, tant dans sa version initiale que dans sa version fine-tunée, afin de comparer leurs performances.

Nous avons également mis en œuvre un nouveau modèle, le **mécanisme d'attention**, cependant, les résultats obtenus n'étaient pas suffisamment convaincants.

3 Collecte et prétraitement des données

3.1 Collecte des données

Dans un premier temps, nous avons automatisé la récupération d'images sur Google en utilisant un script Python de web scraping. Ce script, disponible sur la repository GitHub de notre projet sous le nom de "scrapping.py", nous a permis de gagner du temps en spécifiant des mots-clés de recherche pertinents.

Par la suite, afin d'enrichir notre jeu de données, nous avons étendu notre collecte d'images en explorant d'autres sources en ligne. L'objectif était d'obtenir un minimum de 200 images pour chaque émotion.

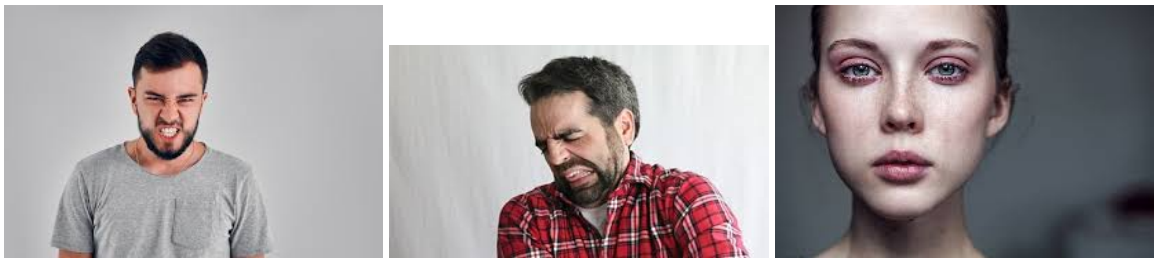

FIGURE 2 – Quelques images de notre base de données

Pour garantir la qualité de notre ensemble de données, nous avons mis en œuvre des scripts pour détecter et supprimer les doublons, minimisant ainsi les redondances.

Nous avons accordé une attention particulière à la diversité des images en capturant des visages de différentes tranches d'âge, incluant des bébés, des enfants, des adultes et des personnes âgées.

Malheureusement, les images disponibles sur Google Images étaient relativement limitées en nombre, ce qui a restreint notre ensemble de données initial. Après avoir testé ce jeu de données restreint et utilisé des techniques telles que l'ImageGenerator de TensorFlow pour augmenter la taille de l'ensemble de données, nous avons constaté que l'exactitude ne dépassait pas les 30%. Par conséquent, pour garantir la cohérence des données et améliorer la précision, nous avons été contraints d'ajouter quelques images provenant d'autres bases de données existantes, bien que cela ne concernait qu'une partie des images.

FIGURE 3 – Quelques images de la base de donnée déjà élaborée

3.2 Répartition des données

Pour la répartition des données, nous avons ajusté notre approche initiale qui prévoyait 50% (200 images) pour l'entraînement et 25% (100 images) pour chaque ensemble de validation et de test. Nous avons opté pour une répartition différente, avec plus de 600 images pour l'entraînement pour un pourcentage de plus de 60% et plus de 200 pour la validation et le test, car nous considérions l'entraînement comme crucial pour le modèle.

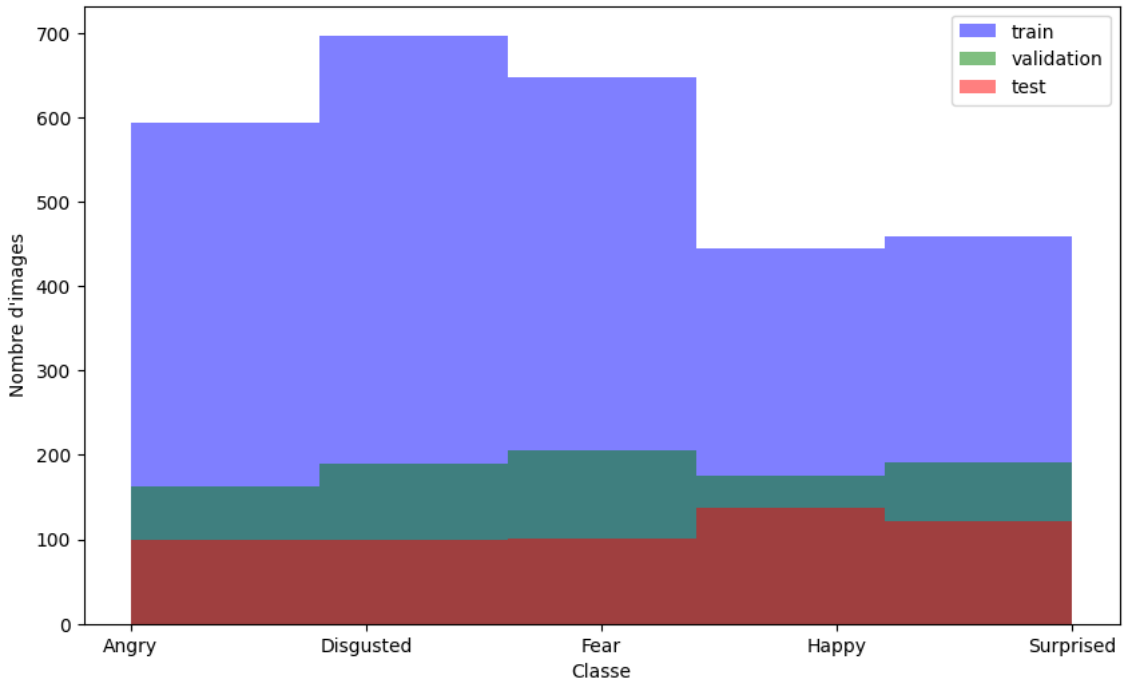

FIGURE 4 – Répartition de nos données

3.3 Prétraitement des données

3.3.1 Prétraitement intial (simple)

Nous avons entamé notre première étape de prétraitement des images en utilisant le script fourni sur Moodle. Ce script nous a permis de charger chaque image, de la redimensionner à une taille fixée de 64 par 64 pixels, de la convertir en format RGB, puis de la stocker dans un tableau nommé x.
Cette méthode s'est avérée relativement simple pour le chargement des données, étant donné la complexité associée à la reconnaissance des émotions faciales, qui met l'accent sur les visages humains.

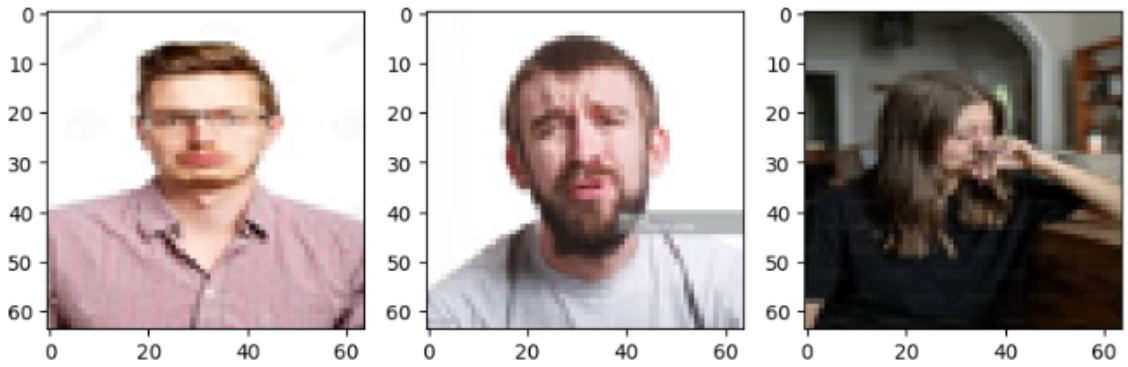

FIGURE 5 – Images après traitement simple

3.3.2 Détection du visage

Nous avons finalement opté pour un traitement des données qui impliquait la détection du visage dans chaque image en utilisant les fonctionnalités de la bibliothèque **OpenCV (cv2)**. Ainsi, les nouvelles données étaient constituées uniquement d'images de visages en niveaux de gris, redimensionnées à une taille de `(64, 64)`.

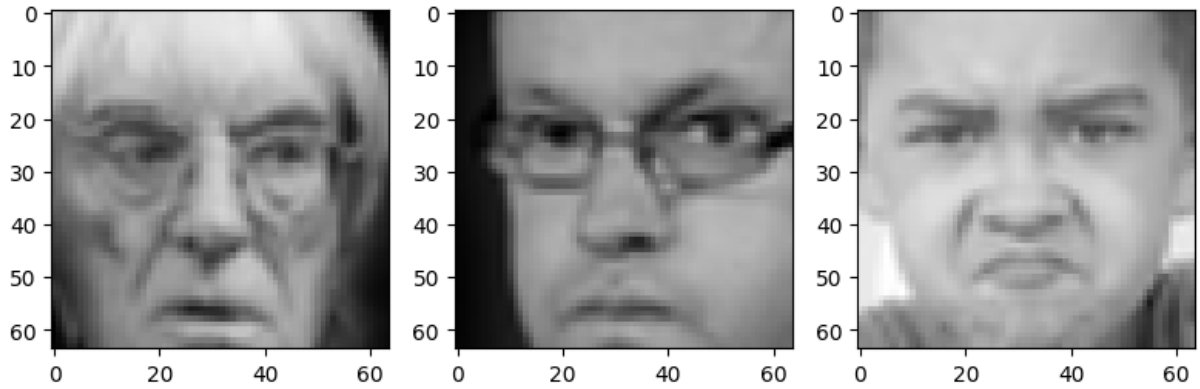

FIGURE 6 – Images après traitement avec la détection de visage

Ce prétraitement, qui consistait à recadrer les images autour du visage, s'est révélé considérablement bénéfique pour les différents modèles.

3.3.3 Détection de la bouche

Nous avons tenté d'entraîner les modèles sur des données qui avaient été prétraitées pour ne contenir que des images de la bouche, étant donné que la bouche joue un rôle crucial dans l'expression des émotions chez les humains. Cependant, cette approche n'a pas été plus avantageuse que l'entraînement sur des visages entiers.

4 Modélisation et entrainement

4.1 Hyperparamètres

```
IMAGE_SIZE = 64
BATCH_SIZE = 32
EPOCHS = 50
FT_EPOCHS = 20
labels = ['angry', 'happiness','fear','disgusted','surprised']
NUM_CLASSES = len(labels)
```

FIGURE 7 – Hyperparamètres

4.2 Modèle Simple

Le premier modèle que nous avons développé est un réseau de neurones convolutif relativement simple, conçu pour capturer les caractéristiques importantes des images de visages et leur expression émotionnelle. La structure du modèle est la suivante :

- Couche d'entrée : Une couche convolutive avec 32 filtres de taille `(3, 3)` et une fonction d'activation ReLU. Cette couche prend en entrée des images de taille `(IMAGE_SIZE, IMAGE_SIZE, 3)`.
- MaxPooling : Une couche de max pooling avec une fenêtre de `(2, 2)` pour réduire la dimensionnalité des caractéristiques extraites.
- Répétition de couches convolutives et de max pooling : Nous avons répété ce motif quatre fois, en augmentant le nombre de filtres convolutifs à chaque étape `(64, 96 et 128)`, tout en réduisant progressivement la taille spatiale des caractéristiques grâce aux opérations de max pooling.
- Couche Flatten : Une couche pour aplatir les caractéristiques extraites en un vecteur unidimensionnel afin de les passer à travers des couches entièrement connectées.
- Couches Dense : Deux couches denses, la première avec 512 neurones et une fonction d'activation `RELU`, suivie d'une couche de sortie avec un nombre de neurones égal au nombre de classes d'émotions (`NUM_CLASSES`). Nous avons utilisé une fonction d'activation `softmax` dans la couche de sortie pour obtenir des probabilités pour chaque classe d'émotion.

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 62, 62, 32)	896
max_pooling2d (MaxPooling2D)	(None, 31, 31, 32)	0
conv2d_1 (Conv2D)	(None, 29, 29, 64)	18,496
max_pooling2d_1 (MaxPooling2D)	(None, 14, 14, 64)	0
conv2d_2 (Conv2D)	(None, 12, 12, 96)	55,392
max_pooling2d_2 (MaxPooling2D)	(None, 6, 6, 96)	0
conv2d_3 (Conv2D)	(None, 4, 4, 128)	110,720
max_pooling2d_3 (MaxPooling2D)	(None, 2, 2, 128)	0
flatten (Flatten)	(None, 512)	0
dense (Dense)	(None, 512)	262,656
dense_1 (Dense)	(None, 5)	2,565

FIGURE 8 – Sommaire de notre modèle simple

4.3 Modèle VGG16

Le modèle VGG16 est un réseau neuronal convolutif (CNN) développé par Simonyan et Zisserman en 2014. Il fait partie de la série des modèles VGG (Visual Geometry Group) et est nommé ainsi car il comporte 16 couches (13 couches de convolution et 3 couches entièrement connectées). Sa profondeur et sa capacité à extraire des caractéristiques complexes en font un choix populaire pour la classification d'images.

Nous avons choisi d'utiliser VGG16 dans notre projet de reconnaissance des émotions faciales pour tirer parti de sa capacité à apprendre des caractéristiques visuelles complexes. En utilisant le modèle pré-entraîné sur le jeu de données ImageNet, nous pouvons bénéficier des connaissances acquises par le modèle sur une large gamme d'objets et de motifs visuels.

4.3.1 Configuration du modèle

Dans notre implémentation, nous avons chargé les poids pré-entraînés du modèle **VGG16**, initialement entraîné sur le jeu de données ImageNet. Nous avons exclu la couche supérieure du modèle (`include_top=False`) pour pouvoir ajouter nos propres couches entièrement connectées adaptées à notre tâche de reconnaissance des émotions faciales. La forme de l'entrée a été fixée à (`IMAGE_SIZE`, `IMAGE_SIZE`, 3), correspondant à la taille des images de notre ensemble de données.

4.3.2 Architecture du modèle

Après avoir chargé le modèle VGG16, nous avons ajouté deux couches entièrement connectées à la fin du modèle. La première couche Dense comporte 256 neurones avec une fonction d'activation `RELU`, suivie d'une couche de sortie avec un nombre de neurones égal au nombre de classes d'émotions (`NUM_CLASSES`), utilisant une fonction d'activation softmax pour obtenir des probabilités pour chaque classe d'émotion. Cette architecture permet au modèle de classifier les émotions faciales en fonction des caractéristiques extraites par **VGG16**.

Layer (type)	Output Shape	Param #
input_layer_1 (InputLayer)	(None, 64, 64, 3)	0
block1_conv1 (Conv2D)	(None, 64, 64, 64)	1,792
block1_conv2 (Conv2D)	(None, 64, 64, 64)	36,928
block1_pool (MaxPooling2D)	(None, 32, 32, 64)	0
block2_conv1 (Conv2D)	(None, 32, 32, 128)	73,856
block2_conv2 (Conv2D)	(None, 32, 32, 128)	147,584
block2_pool (MaxPooling2D)	(None, 16, 16, 128)	0
block3_conv1 (Conv2D)	(None, 16, 16, 256)	295,168
block3_conv2 (Conv2D)	(None, 16, 16, 256)	590,080
block3_conv3 (Conv2D)	(None, 16, 16, 256)	590,080
block3_pool (MaxPooling2D)	(None, 8, 8, 256)	0
block4_conv1 (Conv2D)	(None, 8, 8, 512)	1,180,160
block4_conv2 (Conv2D)	(None, 8, 8, 512)	2,359,808
block4_conv3 (Conv2D)	(None, 8, 8, 512)	2,359,808
block4_pool (MaxPooling2D)	(None, 4, 4, 512)	0
block5_conv1 (Conv2D)	(None, 4, 4, 512)	2,359,808
block5_conv2 (Conv2D)	(None, 4, 4, 512)	2,359,808
block5_conv3 (Conv2D)	(None, 4, 4, 512)	2,359,808
block5_pool (MaxPooling2D)	(None, 2, 2, 512)	0

FIGURE 9 – Sommaire de notre modèle VGG16

4.3.3 Modèle VGG16 avec fine tuning

Dans notre approche de fine-tuning du modèle VGG16, nous avons ajouté des couches supplémentaires au modèle de base pour adapter le réseau à notre tâche spécifique de reconnaissance

des émotions faciales.

Tout d'abord, nous avons chargé les poids pré-entraînés du modèle VGG16, initialement entraîné sur le jeu de données ImageNet, en utilisant la classe Sequential de Keras. Ensuite, nous avons ajouté une couche Flatten pour aplatir les caractéristiques extraites par les couches convolutionnelles de VGG16 afin de les passer à travers des couches entièrement connectées.

Après la couche Flatten, nous avons ajouté une couche Dense avec 256 neurones et une fonction d'activation ReLU pour apprendre des représentations de plus haut niveau à partir des caractéristiques extraites par VGG16. Enfin, nous avons ajouté une couche Dense de sortie avec 4 neurones, correspondant au nombre de classes d'émotions que nous cherchons à prédire, avec une fonction d'activation softmax pour obtenir des probabilités pour chaque classe d'émotion.

Il est important de noter que nous avons gelé les poids des couches du modèle VGG16 en les définissant sur non-entraînable (trainable=False). Cela signifie que lors de l'entraînement, seules les nouvelles couches que nous avons ajoutées seront mises à jour, tandis que les poids des couches pré-entraînées de VGG16 resteront inchangés.

Layer (type)	Output Shape	Param #
vgg16 (Functional)	(None, 2, 2, 512)	14,714,688
flatten_2 (Flatten)	(None, 2048)	0
dense_8 (Dense)	(None, 256)	524,544
dense_9 (Dense)	(None, 5)	1,285

FIGURE 10 – Sommaire de notre modèle VGG16 avec fine tuning

Cette approche de fine-tuning nous permet de tirer parti des connaissances préalables apprises par VGG16 sur ImageNet tout en adaptant le modèle aux caractéristiques spécifiques de notre ensemble de données de reconnaissance des émotions faciales.

5 Evaluation des performances

Dans l'ensemble de notre projet de reconnaissance des émotions faciales, l'évaluation des performances revêt une importance capitale. Cette phase constitue en effet le baromètre de l'efficacité et de la fiabilité de notre système dans la détection et l'interprétation des expressions émotionnelles humaines à partir d'images faciales.

Cette section vise à fournir une analyse détaillée des performances de notre système, en examinant sa capacité à identifier avec précision les émotions présentes dans les visages capturés. Nous aborderons ainsi les métriques d'évaluation utilisées, les résultats obtenus, ainsi que les implications de ces résultats pour notre projet.

Du coup, Nous avons tracé les courbes d'apprentissage pour l'accuracy et la perte, ainsi que calculé la matrice de confusion pour évaluer les performances de notre modèle. De plus, nous avons calculé la précision globale et par classe pour fournir une analyse détaillée de ses performances.

5.1 Performances du modèle simple

Commençons par évaluer les performances du modèle simple.

5.1.1 Courbe d'accuracy

Les courbes d'accuracy du modèle simple présentent des comportements intéressants. Initialement, l'accuracy sur les données d'entraînement et de validation augmente rapidement au fil des époques, ce qui suggère que le modèle apprend efficacement à partir des données. Cependant, après un certain nombre d'époques, on peut observer un écart entre l'accuracy sur les données d'entraînement et celle sur les données de validation.

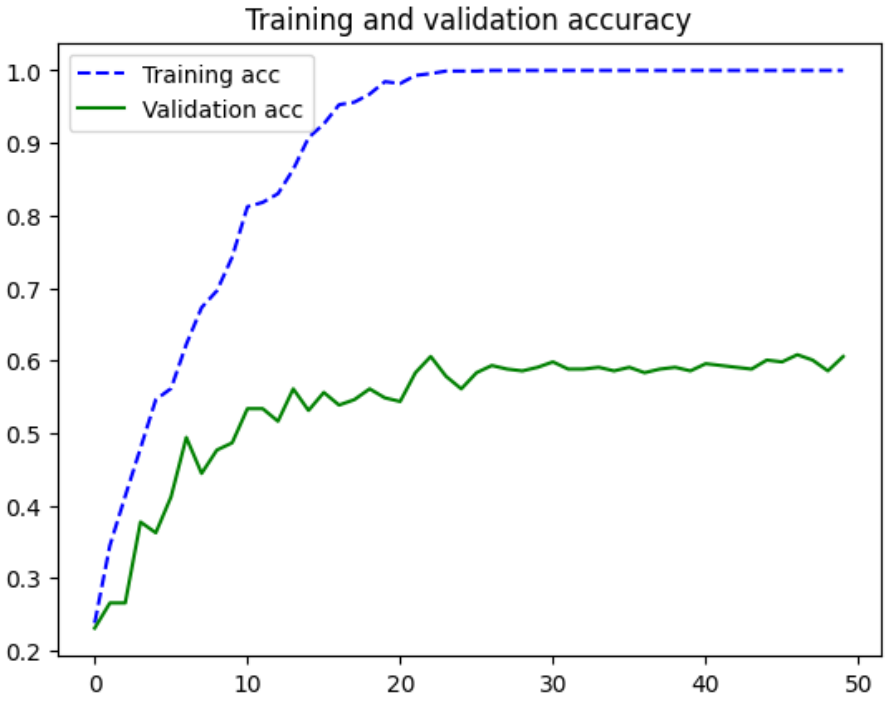

FIGURE 11 – Courbe d'accuracy du modèle simple

Les courbes d'accuracy du modèle simple atteignent un maximum d'accuracy sur les données d'entraînement d'environ 1.0, ce qui suggère que le modèle parvient à ajuster parfaitement les données d'entraînement. Cependant, sur les données de validation, le pic d'accuracy est d'environ 0.61, ce qui montre une certaine différence par rapport à l'accuracy maximale sur les données d'entraînement. Cette différence indique un certain degré de surapprentissage, où le modèle s'est trop adapté aux données d'entraînement spécifiques et ne généralise pas bien aux données non vues.

5.1.2 Courbe du loss

Le loss diminue progressivement au fil des époques, passant d'une valeur initiale d'environ 10.59 à environ 2.78 à la dernière époque. Cependant, il reste notablement plus élevé sur les données de validation que sur les données d'entraînement, ce qui suggère un certain niveau de surapprentissage. Le loss sur les données de validation atteint son minimum autour de la 21ème époque avec une valeur d'environ 1.28, tandis que le loss sur les données d'entraînement continue de diminuer jusqu'à la dernière époque.

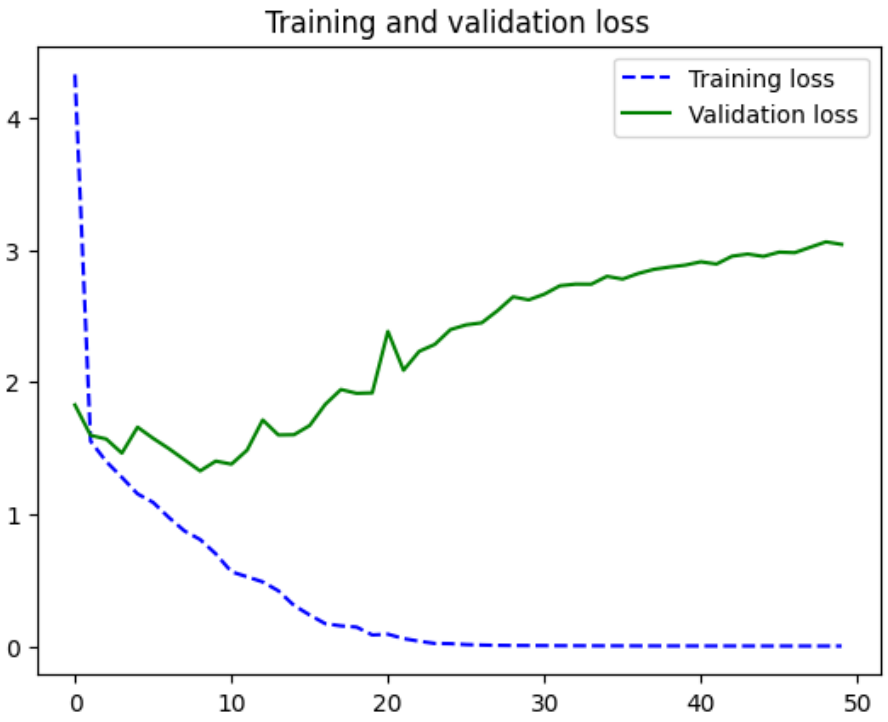

FIGURE 12 – Courbe du loss du modèle simple

5.1.3 Matrice de confusion

Dans la matrice de confusion, on observe que toutes les émotions dépassent le seuil de 50% après le test, indiquant une certaine capacité du modèle à reconnaître différentes émotions. Cependant, l'émotion de la colère se distingue par une confusion assez équilibrée entre plusieurs autres émotions, notamment la peur, le dégoût et la surprise. Cette confusion peut être interprétée comme étant due à la similarité dans les expressions faciales associées à ces émotions. Par exemple, une expression faciale de colère peut également présenter des caractéristiques de peur ou de dégoût, ce qui rend la distinction entre ces émotions plus difficile pour le modèle. En conséquence, malgré une performance globale satisfaisante, la reconnaissance précise de l'émotion de la colère peut être plus délicate en raison de ces similitudes dans les expressions faciales.

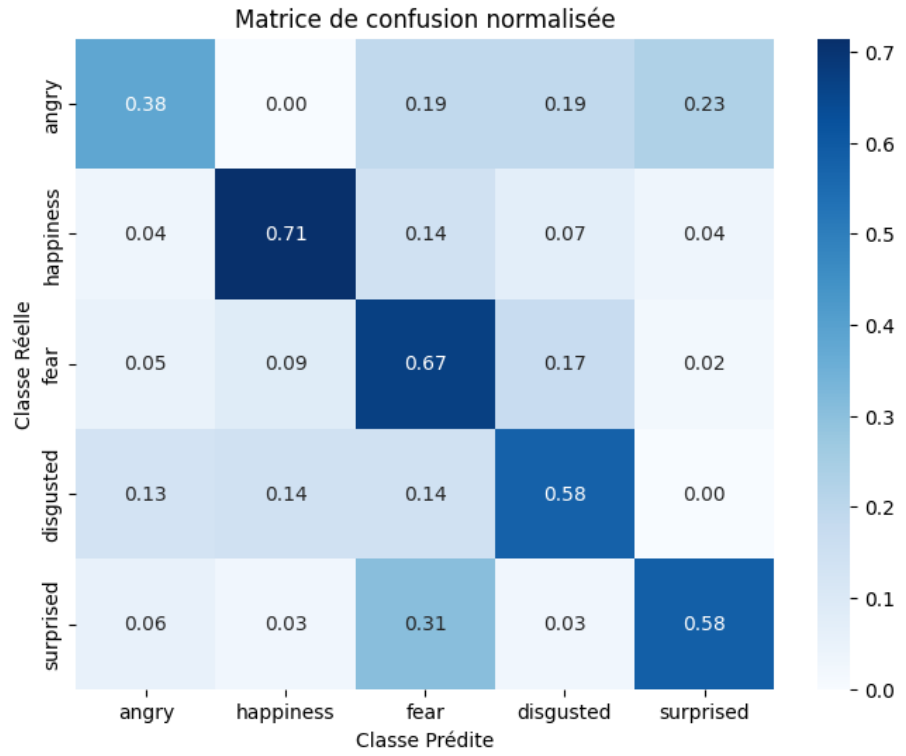

FIGURE 13 – Matrice de confusion du modèle simple

Par exemple, dans notre base de données, nous avons une image qui est principalement associée à l'émotion de colère. Cependant, cette même image pourrait facilement être interprétée comme représentant un enfant effrayé ou triste.

FIGURE 14 – Image d'enfant en colère

5.1.4 Précision globale et par classe

Les performances par classe révèlent des résultats variables. La précision la plus élevée est observée pour les émotions de joie (71%) et de peur (67%). Cependant, l'émotions de colère et de présente des performances plus faibles avec une précision de 38.46%. La précision globale du modèle est de 59.71%. Ces résultats suggèrent que le modèle peut mieux reconnaître certaines émotions par rapport à d'autres, ce qui pourrait nécessiter une analyse plus approfondie pour identifier les raisons sous-jacentes à ces différences de performance.

5.1.5 Test avec des images réelles

Nous avons évalué les performances de notre modèle simple en utilisant des images réelles capturées en selfie, représentant diverses émotions. Cependant, les résultats étaient décevants, avec une précision globale de seulement 35%. Cette faible performance peut être attribuée à la disponibilité limitée de données d'images réelles pour les tests, avec environ 12 images par classe seulement.

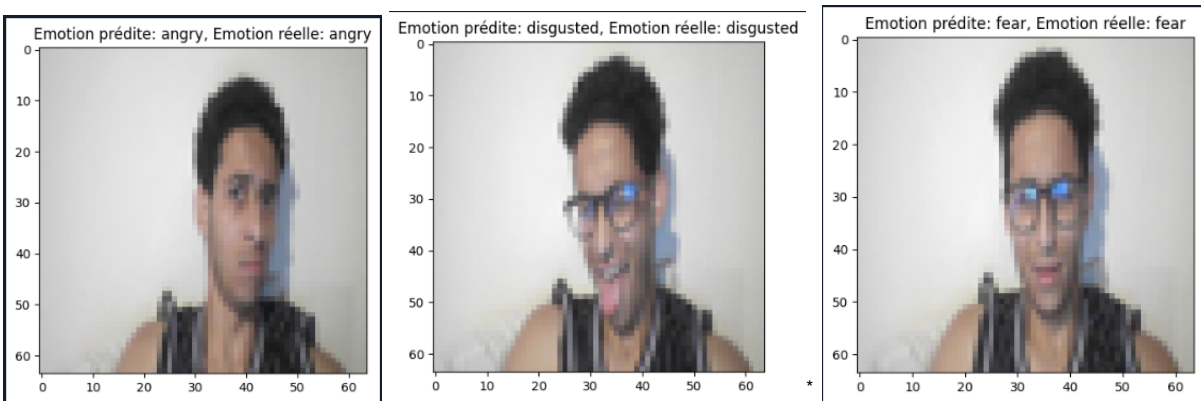

FIGURE 15 – Quelques images de tests réels après traitement

5.2 Performance du modèle VGG16

Pour améliorer les performances de notre système de reconnaissance d'émotions, nous avons opté pour l'utilisation du modèle VGG16, une architecture de réseau neuronal convolutif largement reconnue et utilisée dans le domaine de la vision par ordinateur. Le modèle VGG16 est réputé pour sa capacité à extraire des caractéristiques complexes à partir d'images grâce à sa profondeur et à sa structure de convolution en cascade. En exploitant les représentations apprises par VGG16 sur un vaste ensemble de données, nous cherchons à améliorer la capacité de notre système à reconnaître efficacement une gamme variée d'émotions.

5.2.1 Courbe d'accuracy

Les courbes d'accuracy du modèle VGG16 montrent une tendance similaire à celles du modèle simple. La précision sur les données d'entraînement augmente progressivement pour atteindre un pic d'environ 95.28%, indiquant une bonne adaptation aux données d'entraînement. Cependant, sur les données de validation, la précision atteint un plateau autour de 51.61%, montrant une légère différence par rapport à l'accuracy maximale sur les données d'entraînement. Cette disparité suggère également un certain degré de surapprentissage, où le modèle pourrait avoir du mal à généraliser efficacement aux données non vues.

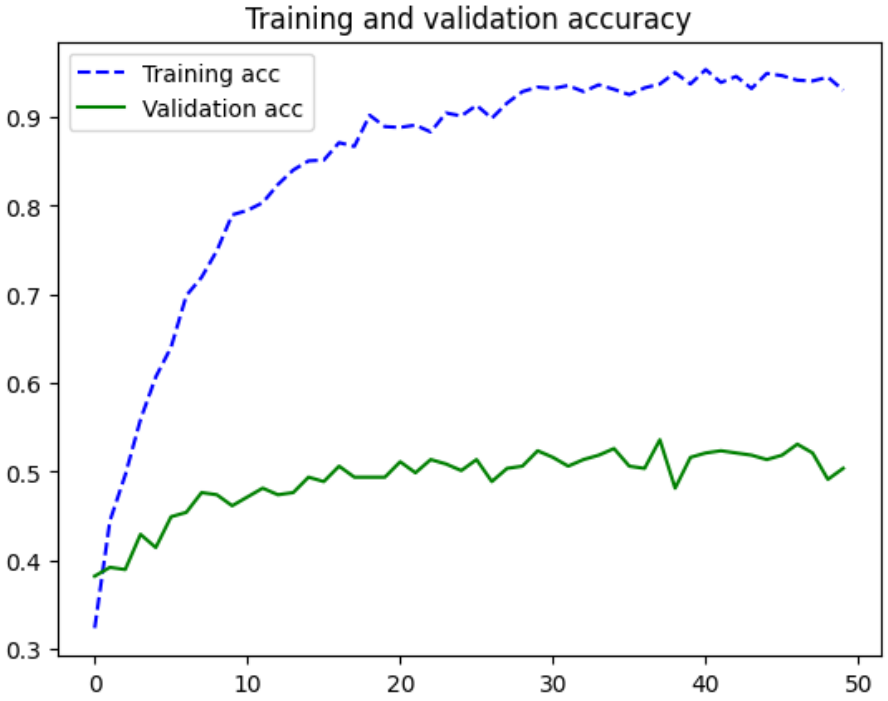

FIGURE 16 – Courbe d'accuracy du vgg16

5.2.2 Courbe du loss

Le loss du modèle VGG16 diminue progressivement au fil des époques, débutant à environ 19.42 et atteignant environ 0.14 à la dernière époque. Cependant, le loss reste notablement plus élevé sur les données de validation que sur les données d'entraînement, indiquant ainsi un certain niveau de surapprentissage. Sur les données de validation, le loss atteint son minimum autour de la 45ème époque, avec une valeur d'environ 3.11, tandis que le loss sur les données d'entraînement continue de diminuer jusqu'à la dernière époque. Cette divergence entre les losses des données d'entraînement et de validation soulève des préoccupations quant à la capacité du modèle à généraliser efficacement aux données non vues, mettant en évidence la nécessité d'explorer des stratégies de régularisation ou d'augmentation des données pour améliorer les performances sur de nouvelles instances.

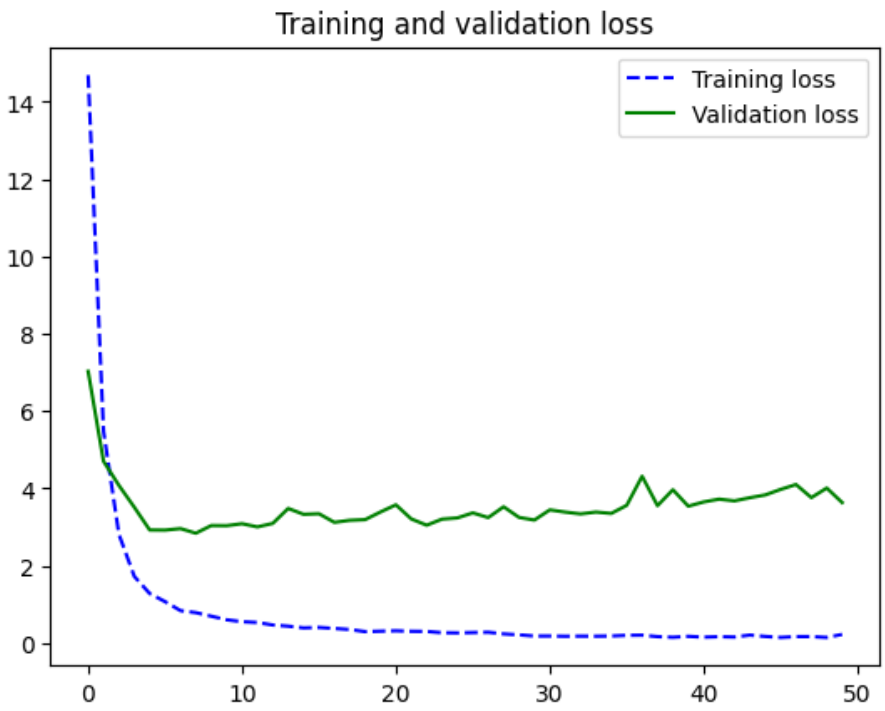

FIGURE 17 – Courbe du loss du vgg16

5.2.3 Matrice de confusion

La matrice de confusion révèle des performances encourageantes pour certaines émotions, avec des taux de reconnaissance élevés pour la classe du bonheur à environ 0.75, la peur à environ 0.69 et le dégoût à environ 0.71. Cependant, la classe de la surprise présente une précision relativement plus faible, avec un score d'environ 0.5. Une observation intéressante est que l'émotion

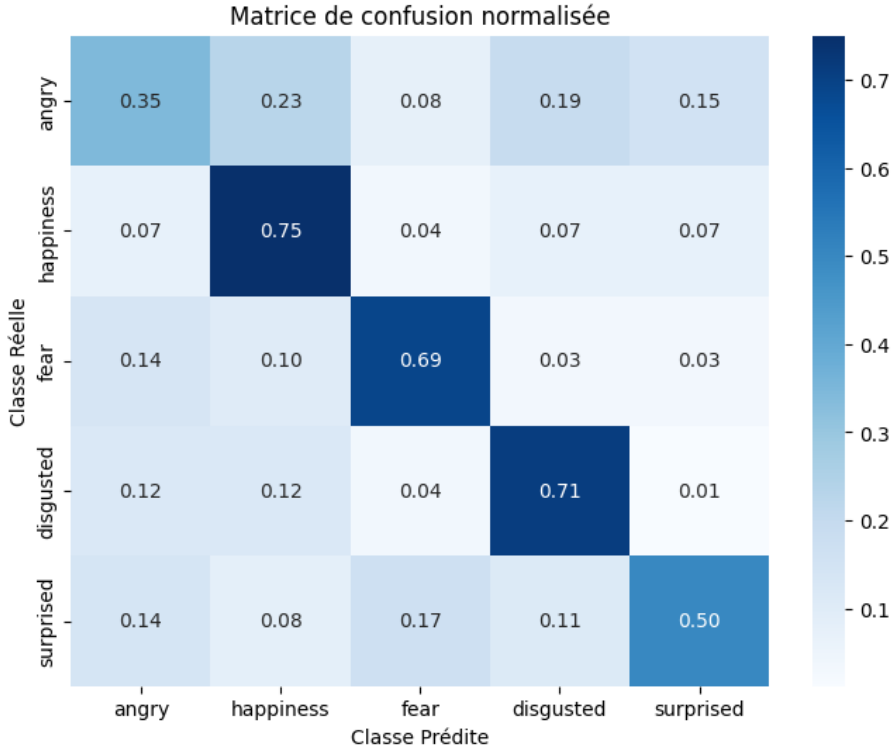

FIGURE 18 – Matrice de confusion du vgg16

de colère semble maintenant être confondue principalement avec la joie, avec un taux de confusion d'environ 0.23, et avec le dégoût, avec un taux d'erreur d'environ 0.19. De plus, la colère est également confondue avec la surprise, avec un taux de confusion d'environ 0.14, et avec la peur, avec un taux d'erreur d'environ 0.17. Comem mentionné auparavant, cette confusion accrue entre la colère et d'autres émotions peut indiquer une certaine similarité dans les expressions faciales ou les caractéristiques associées à ces émotions, soulignant ainsi la complexité de la tâche de reconnaissance des émotions et la nécessité d'améliorer la capacité du modèle à distinguer ces nuances subtiles.

5.2.4 Précision globale et par classe

L'analyse de l'accuracy par classe révèle des performances variables du modèle. Pour la classe "angry", l'accuracy est de 0.38, ce qui indique que le modèle a du mal à bien reconnaître cette émotion. En revanche, pour les classes "happiness", "fear" et "disgusted", l'accuracy est plus élevée, atteignant respectivement 0.71, 0.72 et 0.66. Cela suggère que le modèle est relativement meilleur pour identifier ces émotions. Pour la classe "surprised", l'accuracy est de 0.5, indiquant une performance moyenne.
L'accuracy globale du modèle est de 0.61, ce qui signifie que dans l'ensemble, le modèle prédit correctement les émotions pour environ 61% des données de test. Bien que cette performance soit modérée, elle montre que le modèle est capable de généraliser correctement sur un ensemble de données distinct de celui sur lequel il a été entraîné. Cependant, des améliorations sont nécessaires pour renforcer la capacité du modèle à reconnaître certaines émotions, notamment la colère, où il montre une performance relativement faible. Malgré nos multiples tentatives pour améliorer les performances, c'est le résultat que nous avons finalement obtenu.

5.3 Comparaison entre le modèle simple et le vgg16

Il est ainsi observé que le modèle VGG16 affiche une précision globale légèrement supérieure (61%) à celle du modèle simple (59%), ce qui est cohérent compte tenu de sa complexité accrue. Cependant, il est important de noter que l'utilisation du modèle VGG16 nécessite des ressources de calcul considérables, notamment l'accès à un GPU pour des temps d'entraînement et de calcul significativement prolongés. En effet, la mise en œuvre de ces calculs sur notre propre ordinateur personnel, en utilisant les capacités GPU disponibles, était nécessaire, étant donné les limitations de ressources rencontrées sur Google Colab.

5.4 Fine Tuning

Le fine-tuning du modèle VGG16 est essentiel pour améliorer la reconnaissance des émotions faciales. Il permet d'adapter les couches supérieures pré-entraînées sur ImageNet à notre dataset spécifique, capturant des caractéristiques pertinentes pour les émotions. Cette approche améliore la précision tout en réduisant le surapprentissage. Cependant, il augmente le temps de calcul, qui reste conséquent, surtout dans un environnement avec des ressources limitées comme Google Colab.

5.4.1 Courbe d'accuracy

L'entraînement du modèle VGG16 montre une amélioration notable de la précision sur le jeu d'entraînement, passant de 83.59% à 97.54% en 50 époques. La précision sur le jeu de validation fluctue autour de 63-67%, ce qui indique un potentiel surapprentissage.

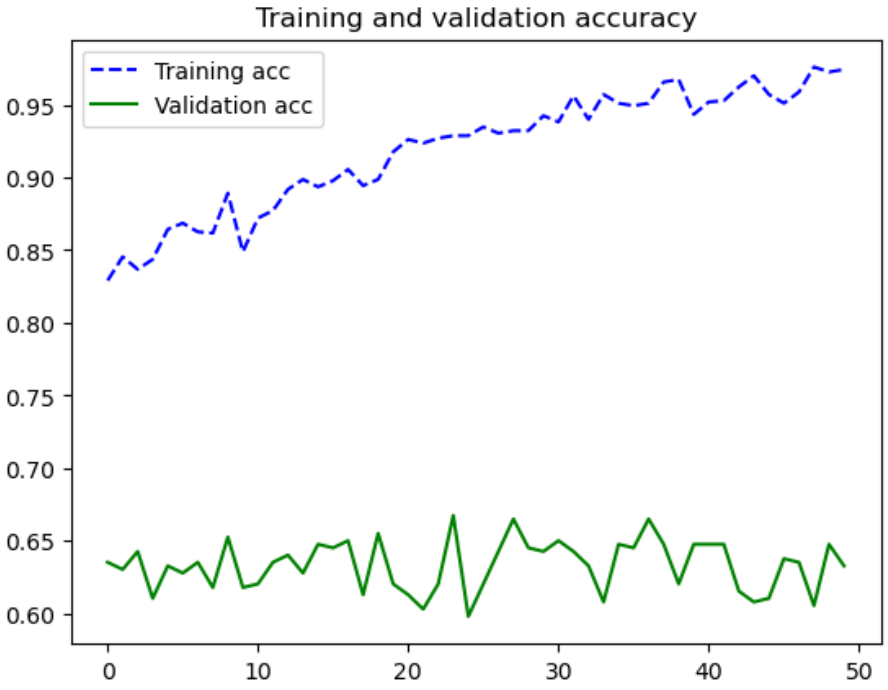

FIGURE 19 – Courbe d'accuracy du modèle vgg16 avec fine tuning

5.4.2 Courbe du loss

La courbe de perte (loss) montre une diminution régulière de la perte sur les données d'entraînement au fil des époques, passant d'environ 0.47 à 0.08. Cependant, la perte sur les données de validation montre une tendance à la hausse après environ 10 époques, suggérant un début de surapprentissage. Bien que la perte sur les données d'entraînement continue de diminuer, celle sur les données de validation augmente après un certain point, ce qui indique une perte de généralisation.

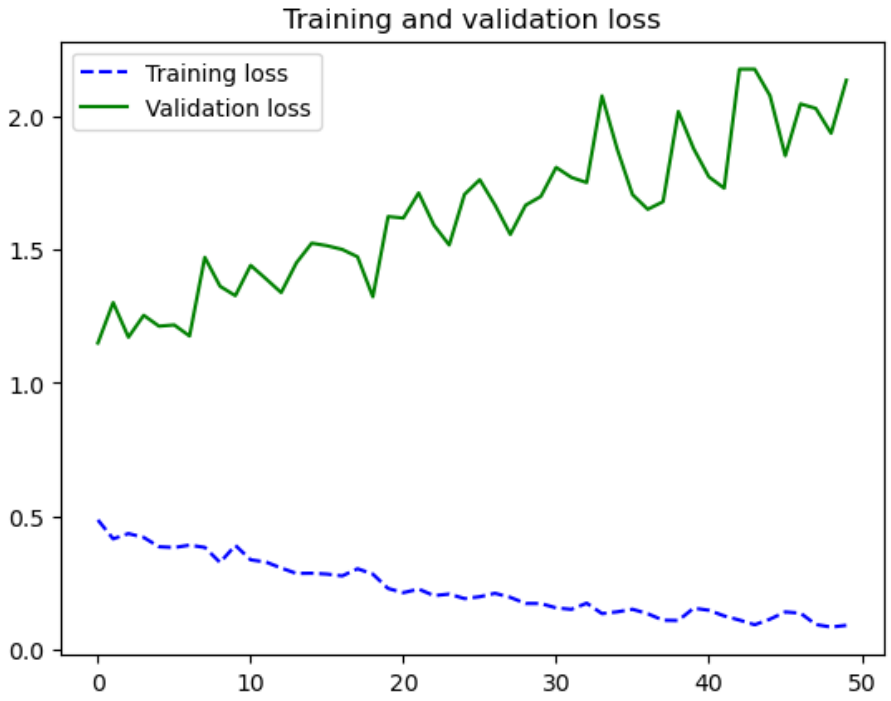

FIGURE 20 – Courbe du loss du modèle vgg16 avec fine tuning

5.4.3 Matrice de confusion

La matrice de confusion révèle que le modèle a une bonne capacité à prédire les émotions de bonheur (0.82), de peur (0.79) et de dégoût (0.71), avec des scores élevés de précision. Cependant, il semble moins performant pour les émotions de colère (0.31) et de surprise (0.64), où les scores de précision sont relativement plus bas. Ces résultats suggèrent que le modèle peut avoir des difficultés à distinguer efficacement les émotions de colère et de surprise par rapport aux autres catégories.

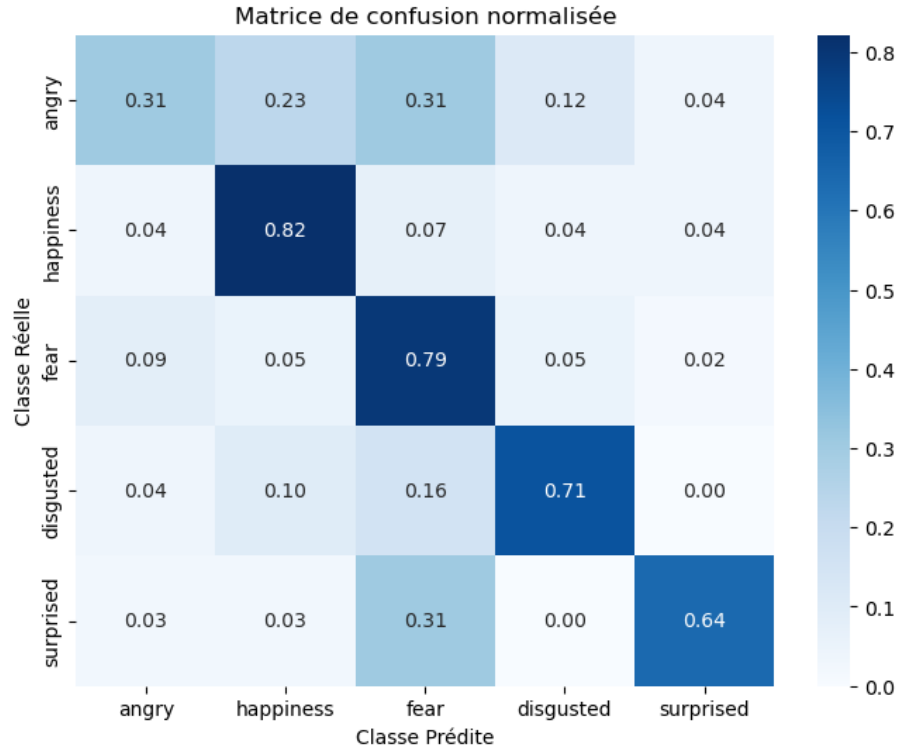

FIGURE 21 – Matrice de confusion du modèle vgg16 avec fine tuning

Malgré toutes les approches explorées, la détection de l'expression faciale de la colère reste un défi en raison des similitudes avec d'autres expressions faciales. Par exemple, une ouverture de bouche peut être confondue avec la joie, tandis que les gestes oculaires peuvent être interprétés à tort comme de la peur.

5.4.4 Précision globale et par classe

Il est notable que le fine-tuning a été bénéfique pour le modèle VGG16. En ajustant les poids pré-entraînés du modèle VGG16 aux caractéristiques spécifiques de notre tâche de classification d'émotions, nous avons pu obtenir une amélioration significative des performances par rapport à un modèle VGG16 simple. Cela suggère que le processus de fine-tuning a permis au modèle de mieux capturer les nuances des expressions faciales associées à différentes émotions, contribuant ainsi à une meilleure précision globale et par classe.

FIGURE 22 – La précision globale et par classe du modèle vgg16 avec fine tuning

5.5 Comparaison entre tous les modèles

En comparant nos trois modèles, nous observons les différences suivantes :

- Le modèle simple a un temps d'exécution rapide, mais une précision globale de 0.59.

- Le modèle VGG16 simple offre un temps d'exécution moyen avec une précision globale légèrement supérieure de 0.61.

-le modèle VGG16 avec fine-tuning nécessite un temps d'exécution plus long, mais atteint une précision globale plus élevée de 0.65.

En considérant ces aspects, le modèle VGG16 avec fine-tuning semble être le meilleur choix, car il offre à la fois une précision globale supérieure et une meilleure capacité à capturer les caractéristiques complexes des expressions faciales. Bien que son temps d'exécution soit plus long, les améliorations significatives de performance en valent la peine pour notre tâche de classification d'émotions.

En augmentant davantage les données d'entraînement, nous aurions pu obtenir de meilleurs résultats et potentiellement atteindre la prédiction initiale que nous avions établie dans le premier rapport, qui visait une précision de 0.7. En enrichissant notre ensemble de données avec plus d'exemples, nous aurions pu améliorer la capacité de nos modèles à généraliser et à reconnaître les différentes expressions faciales avec une plus grande précision. Cela aurait également pu réduire les effets du surapprentissage et renforcer la robustesse de nos modèles vis-à-vis de la variabilité des données réelles. Ainsi, l'augmentation de la taille de l'ensemble de données aurait pu constituer une étape importante vers l'atteinte de nos objectifs de performance initiaux.

6 Méthodes agiles

Dans notre projet de reconnaissance des émotions faciales, nous avons opté pour une approche agile afin de favoriser la collaboration et l'efficacité dans un environnement de travail en binôme. Les méthodes agiles nous ont permis de maintenir une organisation rigoureuse, d'adapter notre planification aux changements et de livrer des fonctionnalités de manière itérative et incrémentielle.

6.1 Cérémonies Hebdomadaires

Chaque semaine, notre binôme organisait des cérémonies agiles pour évaluer les progrès, identifier les défis et planifier les prochaines étapes. Ces réunions nous ont permis de maintenir une communication claire et de prendre des décisions informées pour avancer dans le projet.

6.2 Collaboration et échange constant

Les méthodes agiles ont stimulé notre binôme à échanger régulièrement des idées et à se soutenir mutuellement dans les tâches. Cette dynamique de travail a favorisé un environnement collaboratif où nous avons pu surmonter les obstacles ensemble et atteindre nos objectifs.

6.3 Travail en sprint

Nous avons également adopté un modèle de travail en sprint pour une gestion efficace du temps et des ressources. Chaque sprint, d'une durée d'une semaine, était dédié à des objectifs spécifiques tels que l'optimisation des hyperparamètres, l'amélioration continue de la base de données et les tests avec différentes configurations de classes. Cette approche nous a permis de maintenir un rythme de travail constant et de produire des résultats concrets à la fin de chaque itération.

7 Problèmes rencontrés

La construction d'une base de données robuste et diversifiée s'est avérée être l'un des défis majeurs. Bien que nous ayons utilisé un scraper d'images pour collecter des données, la tâche de trouver des images de haute qualité et représentatives de différentes expressions émotionnelles s'est révélée ardue. De plus, la quantité d'images disponible était souvent limitée, ce qui a restreint la taille et la diversité de notre ensemble de données.

Une autre difficulté majeure a été les limitations des ressources informatiques, en particulier lors de l'utilisation de Google Colab. Les ressources disponibles étaient souvent insuffisantes pour traiter de grands ensembles de données ou pour entraîner des modèles complexes comme VGG16. Cette contrainte nous a obligés à utiliser les ressources de nos propres ordinateurs personnels, ce qui a parfois entraîné des temps de calcul considérablement plus longs et des retards dans nos expériences.

En dépit de nos efforts pour améliorer les performances du modèle, en utilisant des techniques telles que le mécanisme d'attention et le prétraitement des images pour détecter les régions pertinentes comme la bouche et le visage, les résultats n'ont pas été aussi significatifs que nous l'espérions. Nous avons poussé nos tentatives au maximum, mais il est clair que des avancées supplémentaires nécessiteront une exploration approfondie et peut-être l'adoption de nouvelles approches ou techniques.

Dans l'ensemble, ces défis nous ont offert une perspective importante sur les complexités de la classification des émotions à partir d'images et soulignent l'importance de la recherche continue pour surmonter ces obstacles.

8 Conclusion et perspectives

En conclusion, nos expériences avec les modèles de classification d'émotions ont donné des résultats mitigés. Bien que le modèle VGG16 ait montré une précision globale légèrement supérieure à celle du modèle simple, les performances n'ont pas été aussi élevées que prévu. De plus, le temps de calcul et d'entraînement du modèle VGG16 a été considérablement plus long, nécessitant même l'utilisation de ressources GPU externes en raison des limitations de puissance de calcul.

Malgré cela, ces résultats ne sont pas décourageants. Ils soulignent plutôt la complexité de la tâche de classification d'émotions à partir d'images, qui peut être influencée par une variété de facteurs tels que la qualité des données, la taille de l'ensemble de données et la complexité des émotions elles-mêmes.

Pour l'avenir, il serait intéressant d'explorer plusieurs pistes pour améliorer les performances des modèles. Cela pourrait inclure l'augmentation de la taille de l'ensemble de données avec des images de meilleure qualité et une plus grande diversité d'expressions émotionnelles. De plus, l'utilisation de techniques avancées telles que le transfert d'apprentissage avec des modèles pré-entraînés sur des ensembles de données plus vastes comme ImageNet pourrait être bénéfique. Et peut etre, on pourrait penser au Fine-tuning des couches supérieures c'est à dire plutôt que de geler complètement les poids des couches pré-entrainées, nous pourrions permettre à certaines couches supérieures d'être ré-entrainées avec un taux d'apprentissage plus faible. Cela permettrait au modèle de s'adapter davantage aux caractéristiques spécifiques de notre ensemble de données.

En outre, il serait utile d'explorer d'autres architectures de modèles plus complexes ou des techniques de traitement d'image avancées pour extraire des caractéristiques plus discriminantes des images d'émotions.

En fin de compte, bien que nous n'ayons pas atteint les performances souhaitées, ces résultats fournissent des indications précieuses pour orienter les efforts futurs dans la recherche sur la classification des émotions à partir d'images.