Universidad del Valle de Guatemala

Facultad de Ingeniería

Deep Learning

Proyecto

Integrantes

Cesar López 22535

Descripcion del problema

En entornos interactivos modernos, como interfaces de usuario sin contacto, videojuegos o sistemas de asistencia, el reconocimiento de gestos con la mano se ha convertido en una herramienta esencial. Sin embargo, la implementación de un sistema robusto y en tiempo real presenta desafíos como la variabilidad de iluminación, la posición de la mano, el fondo y la diversidad de usuarios.

El problema principal que se busca resolver es la clasificación automática de gestos de la mano capturados por una cámara en tiempo real, utilizando un modelo de aprendizaje profundo que logre identificar correctamente distintas posiciones o movimientos (por ejemplo: “open”, “thumbs up”, “peace”, etc.).

Análisis

Para abordar este problema, se analizaron los siguientes aspectos:

- Entrada del sistema: Imágenes RGB capturadas desde una cámara web en tiempo real.
- Preprocesamiento: Recorte de la región de la mano utilizando MediaPipe Hands, conversión a RGB y redimensionamiento a 224x224 píxeles.
- Modelo base: Se seleccionó MobileNetV2, una red convolucional ligera y eficiente, ideal para dispositivos con recursos limitados.
- Clases de salida: Un conjunto de gestos definidos en labels.json, cada uno con un conjunto de imágenes recolectadas manualmente.
- Dificultades encontradas:
    - Confusión entre gestos visualmente similares.
    - Desequilibrio entre clases (algunas con más ejemplos que otras).
    - Ruido visual (fondos, sombras, posiciones parciales de la mano).

Propuesta de solución

Se propuso un sistema de reconocimiento de gestos basado en aprendizaje profundo, compuesto por las siguientes etapas:

1. Recolección de datos: Captura de imágenes de la mano usando una interfaz en OpenCV y MediaPipe para segmentar automáticamente la mano.
2. Preprocesamiento y etiquetado: Conversión a RGB, recorte de la mano y almacenamiento por clase.
3. Entrenamiento del modelo: Uso de MobileNetV2 preentrenado con ImageNet como extractor de características y una capa densa personalizada para clasificación de gestos.
4. Inferencia en tiempo real: Implementación de una aplicación que captura video, procesa la mano en cada frame y muestra la predicción sobre la imagen.
5. Optimización: Aplicación de técnicas de fine-tuning, normalización y suavizado temporal para mejorar la estabilidad de predicción.

Descripcion de Solución

El sistema fue implementado en Python utilizando TensorFlow y MediaPipe. Se compone de tres scripts principales:

- collect_dataset.py: Permite capturar imágenes para cada clase. Se pueden grabar ejemplos variando el ángulo, la iluminación y la distancia.
- train.py: Entrena el modelo MobileNetV2 con los datos preprocesados, genera reportes de desempeño y guarda el modelo final en formato .keras.
- inference_webcam.py: Ejecuta el modelo en tiempo real, detecta la mano, la recorta, predice el gesto y muestra el resultado en pantalla.

Además, se añadieron mejoras:

- Suavizado temporal: media móvil sobre las últimas predicciones para reducir fluctuaciones.
- Umbral de confianza: evita falsas clasificaciones cuando el modelo no está seguro.

Herramientas Aplicadas

- TensorFlow 2.12+: Framework principal para construir y entrenar el modelo de deep learning.

- Keras: API de alto nivel para definir y compilar el modelo.
- OpenCV: Captura y manipulación de imágenes en tiempo real.
- MediaPipe: Detección de la mano y extracción de puntos clave.
- NumPy y Matplotlib: Procesamiento de datos y visualización de resultados.
- Scikit-learn: Generación de métricas como matriz de confusión y clasificación.
- MobileNetV2: Arquitectura CNN eficiente utilizada como base.
- Data augmentation (rotación, flip, contraste, zoom): mejora la generalización del modelo.

Resultados (Métricas)

```
Saved model to models/mobilenetv2.keras
              precision    recall  f1-score   support

        open     1.0000    1.0000    1.0000       113
        fist     1.0000    1.0000    1.0000        19
   thumbs_up     1.0000    1.0000    1.0000        26
 thumbs_down     1.0000    1.0000    1.0000        33
       heart     1.0000    1.0000    1.0000        48
       peace     1.0000    1.0000    1.0000        24

    accuracy                         1.0000       263
   macro avg     1.0000    1.0000    1.0000       263
weighted avg     1.0000    1.0000    1.0000       263
```

Conclusiones

- El sistema propuesto logró reconocer gestos de la mano con alta precisión, manteniendo una ejecución fluida en tiempo real.

- El uso de MobileNetV2 permitió un balance adecuado entre velocidad y precisión, y las técnicas de suavizado temporal mejoraron la estabilidad de las predicciones.

Anexos

- Presentación: https://www.canva.com/design/DAG4WA-OoHY/zS2THlWGI0A5jeXGuqE_DQ/edit?utm_content=DAG4WA-OoHY&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton

- Repositorio: https://github.com/Czar272/Proyecto_DL.git