Detección automática de la calidad del lavado de manos empleando aprendizaje profundo

Diana Cabrera, Miguel Ortiz, Daniel Garavito

https://youtu.be/TjLdP1-_4Io

Antecedentes

- El lavado de manos es el mecanismo más costo - eficiente para la prevención no solo del COVID-19 sino múltiples enfermedades respiratorias e intestinales.
- Implementar un modelo de reconocimiento de acciones requiere 1) Preprocesamiento 2) Muestreo adecuado de videos 3) Modelos basados en redes convolucionales como I3D, ResNet 3D, ResNet(2+1)D, NGM, modelos de dos flujos como SlowFast, entre otros.

Definición del Problema

Problema: Construir un modelo que pueda identificar los pasos del lavado de manos recomendados por la OMS.
Enfoque: Redes convolucionales residuales 3D pre-entrenadas para clasificación de acciones en videos.
Métricas de evaluación: F1, Exactitud

Datos

- Se emplearon modelos pre-entrenados en el Kinetics-400, un dataset masivo de acciones humanas.
- Se usó un dataset pequeño disponible en Kaggle
- HandWash: 25 grabaciones de los 7 pasos recomendados por la OMS particionados en 12 sub-pasos con longitud promedio de 12.9 segundos.

Datasets	Año	Acciones	Videos	Total
Kinetics-400	2017	400	min 400	306,245
HandWash	2020	12	25	300

Tabla 1. Estadísticas de los datasets
"Acciones", especifica el número de clases. "Videos" es el número de videos por clase.

Figura 1.
Ejemplo del primer cuadro de un batch de clips

- Los clips son redimensionados a 171x128 y luego se hace un crop de 112x112
- Se realiza data augmentation con crops aleatorios y flips horizontales aleatorios en entrenamiento
- Se aplica center crop durante validación y prueba

Modelos

Empleamos redes residuales 3D basadas en ResNet-18:

- **R3D**: ResNet 3D
- **MC3**: Convoluciones Mixtas 3D primero y luego 2D
- **R(2+1)D**: Descomposición en convoluciones espaciales (2D) y convoluciones temporales (1D).

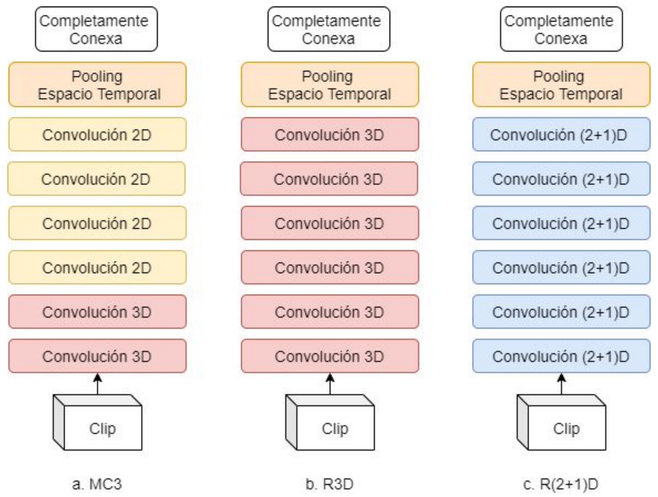

Figura 2.
Configuraciones de los modelos implementados

Experimentos

- Los cuadros por segundo y el muestreo de cuadros son factores muy importantes para el análisis de videos.
- Probamos post-procesamiento de las predicciones de clips para mejorar el rendimiento en prueba

	Stride Temporal			
	1		2	
	Clip			
	1	2	1	2
R(2+1)D	0.87	0.90	0.88	0.92
R3D	0.88	0.93	0.92	0.92
MC3	0.95	**0.97**	0.95	0.95

Tabla 2. F1 Resultados en la partición de prueba
"Clip": número de clips por video con 2 la predicción final es promediada. "Stride Temporal" controla los cuadros por segundo, 2 equivale a 15 fps.

Discusión

- El hecho de que la implementación más parsimoniosa mejore el desempeño sugiere que es posible simplificar aún más la arquitectura.
- Los datos presentados a los modelos son distintos en cada epoch, lo que puede ocasionar un entrenamiento caótico y resultados con alta varianza.

Conclusiones

- La MC3 además de presentar un mayor desempeño tiene una tercera parte de los parámetros.
- Modelos relativamente sencillos consiguen muy buenos resultados.
- No usar características "hechas a mano" es un gran beneficio.
- El conjunto de datos no presenta la variabilidad necesaria para ponerlo en producción. No obstante muestra que las arquitecturas empleadas son buenas candidatas para atender el problema.