

Informe de pruebas SISO-VID

Elaboración:	Equipo técnico
Entidad Ejecutora:	X-TRA PLUS SOLUCIONES DE ENERGÍA S.A.C
Proyecto:	Desarrollo de un Sistema Adaptativo para la Detección de Somnolencia en Conductores de Transporte Interprovincial idóneo para las características únicas de las Carreteras del Perú mediante Sensado Híbrido utilizando Técnicas de Deep Learning.
Periodo:	Marzo 2021
Fecha:	15 de marzo de 2021

1. Objetivo

Realizar las pruebas del modulo SISO-VID. Este componente de software es el encargado de detectar somnolencia a partir de una secuencia de video utilizando aprendizaje profundo.

2. Introducción

Según *World Health Organization* WHO (2013), 1.24 millones de accidentes de tráfico ocurren cada día. Además, *The National Highway Traffic Safety Administration* (NHTSA), menciona que en USA han ocurrido 153,297 accidentes de tránsito entre el 2011 al 2015, y de estos el 2.4 % fueron causados por conductores con somnolencia. Incluso, 1.25 millones de personas mueren cada año en accidentes de tránsito, 20 a 50 millones han sido heridos o están discapacitados y todo esto ha llegado a costar 518 billones de dólares. Mas alarmante, se predice que los accidentes de tránsito serán la quinta causa mas frecuente de muertes para el 2030 (ASIRT, 2020).

SISO-VID toma como entrada una secuencia de video y detecta si alguna escena de la secuencia de video presenta somnolencia. En resumen, se ha utilizado detección de rostros con una red neuronal *Single Shot Detector* (SSD), luego sobre el rostro detectado se evaluaron tres modelos de redes neuronales (VGG, ResNet e Inception) para determinar la de mejor desempeño, de estas se escogió la red neuronal Inception.

3. SISO-VID

SISO-VID es un componente de software basado en aprendizaje profundo para la detección de somnolencia tomando como entrada una secuencia de video. En la Figura 1, presentamos la metodología utilizada de SISO-VID. El método propuesto toma una secuencia de video como entrada, luego extrae los fotogramas, por cada fotograma se detecta el rostro utilizando la red neuronal SSD, para finalmente utilizar la red Inception sobre el rostro detectado para determinar si exista somnolencia. En la ultima etapa, se utilizó Inception al ser la de mejor acierto comparado con VGG y ResNet.

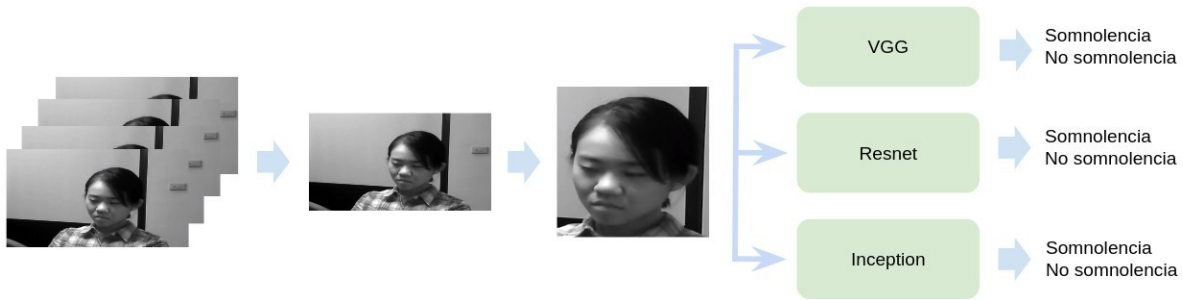


Figura 1: Metodología utilizada por SISO-VID.

3.1. VGG

Propuesta por Simonyan and Zisserman (2014) en el 2014, es una de las primeras redes profundas más conocida. VGG tiene varias versiones diferenciadas por el número de capas, entre esta tenemos VGG-16 y VGG-19. En la Figura 2 y 3, detallamos la arquitectura de VGG-16 y VGG-19 respectivamente. Nosotros evaluamos el desempeño de VGG-16 para la detección de somnolencia.

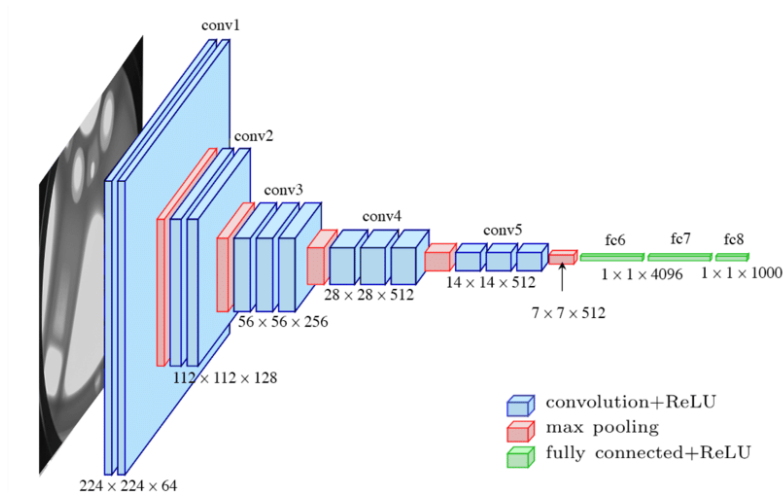


Figura 2: Arquitectura de la red neuronal VGG-16.

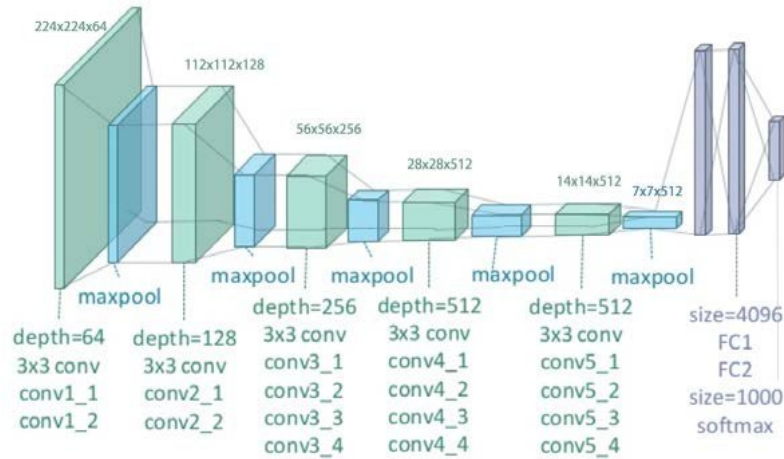


Figura 3: Arquitectura de la red neuronal VGG-19.

Los parametros e hyper parametros utilizados para esta red neuronal VGG-16 se presentan en la Tabla 1.

Tabla 1: Parametros utilizados con VGG-16

Parametro	Valor
batch	32
input	224 x 224 x 3
momentum	0.9
decay	0.0005
learning rate	0.001
epochs	100

3.2. ResNet

Propuesta por He et al. (2016), representa un gran avance en computación gráfica en los ultimos años. ResNet, hizo posible entrenar cientos de capas de neuronas y lograr buenos resultados. La idea principal de ResNet es la introducción del concepto *identity shortcut connection*, que salta una o mas capas (ver Figura 4). Con este concepto, los autores solucionaron el problema de *vanish gradient* que se tenia en arquitecturas grandes de redes. En la Figura 5, mostramos una comparación de ResNet con VGG-16 y otra red con 34 capas (SISO-VID utilizo dicha arquitectura de ResNet).

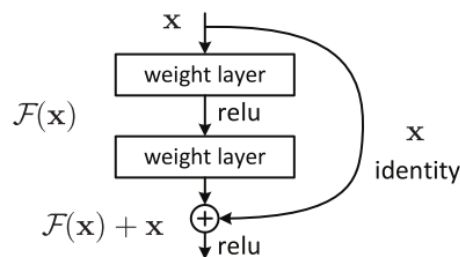


Figura 4: Bloque residual utilizado por ResNet.

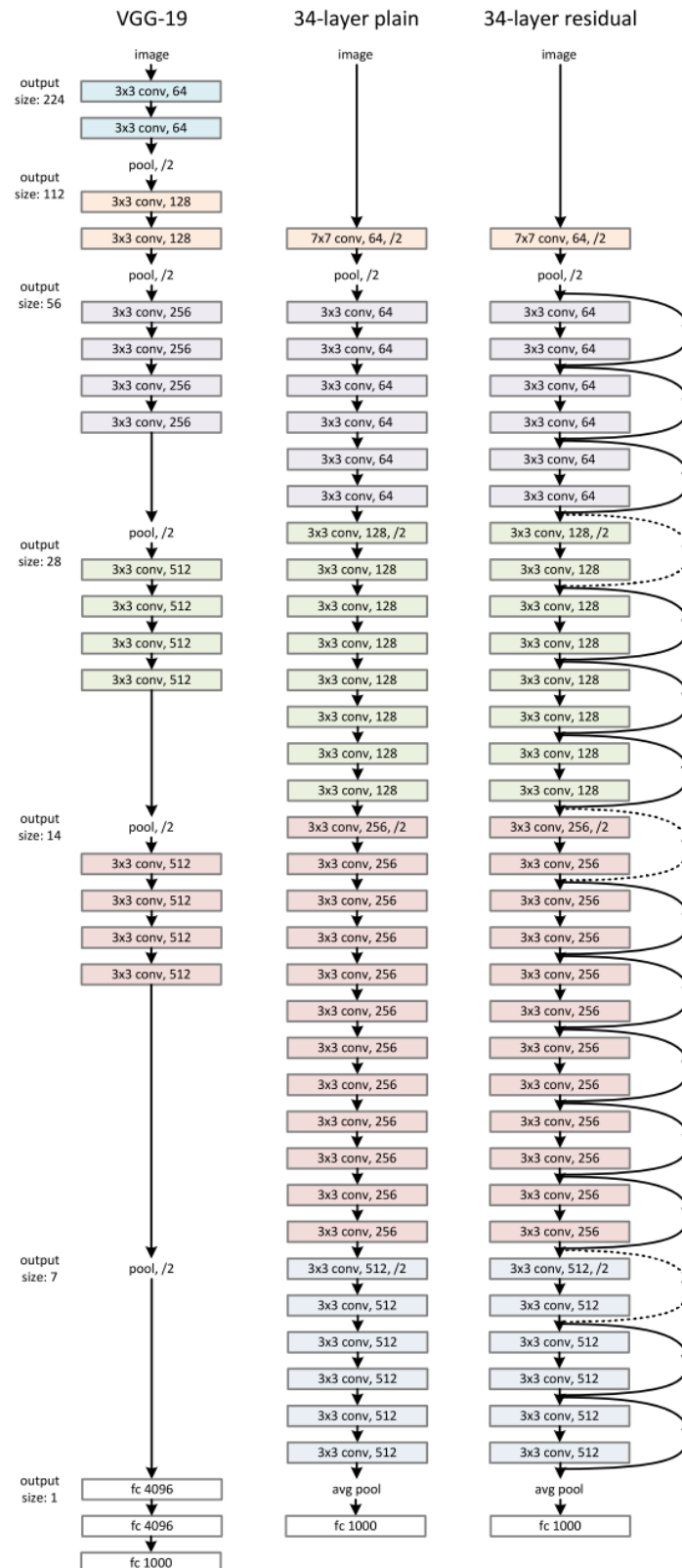


Figura 5: Arquitectura de ResNet utilizada y una comparación con VGG-16 y una red con 34 capas.

Los parametros e hyper parametros utilizados para esta red neuronal ResNet se presentan en la Tabla 2.

Tabla 2: Parametros utilizados con ResNet

Parametro	Valor
batch	32
input	224 x 224 x 3
momentum	0.9
decay	0.0005
learning rate	0.001
epochs	100

3.3. Inception

La red neuronal Inception Szegedy et al. (2016) en su tercera versión, es un modelo de reconocimiento de imágenes muy utilizado, está logró una exactitud de 78.1 % en la base de datos ImageNet. Debido a esto, se utilizo esta red neuronal para detectar somnolencia.

La red neuronal Inception-v3 hace uso de unos módulos llamados Inception. Estos actúan como filtros aplicados a un mismo valor de entrada mediante capas convolucionales y de *pooling*. Esto permite sacar provecho de la extracción de patrones que brindan diferentes tamaños en los filtros. Luego, el resultado de estos filtros es concatenado y utilizado como el valor de salida del módulo. Este modelo aumenta el número de parámetros entrenables y la computación requerida, pero mejora considerablemente la exactitud. En la Figura 6, mostramos la arquitectura de esta red.

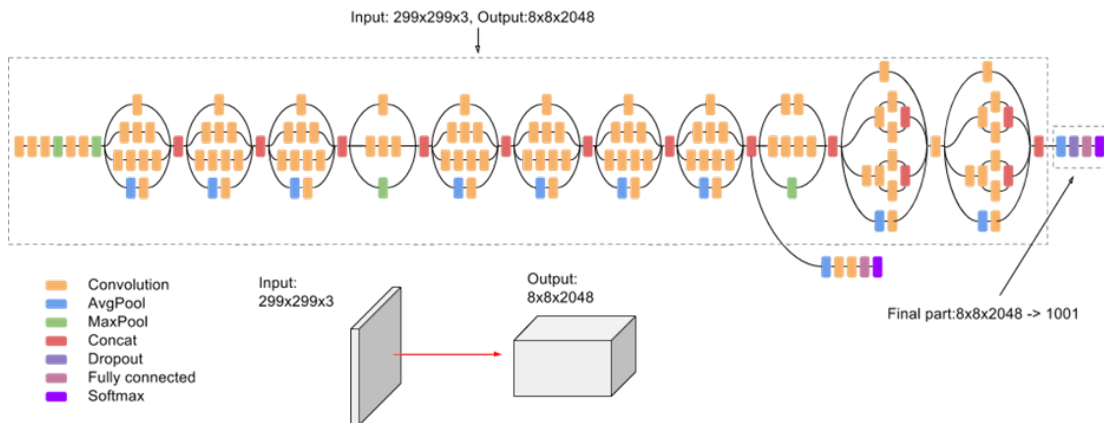


Figura 6: Detección de somnolencia a partir de un rostro utilizando VGG, Resnet e Inception.

Los parametros e hyper parametros utilizados para esta red neuronal se presentan en la Tabla 3.

Tabla 3: Parametros utilizados con Inception-v3

Parametro	Valor
batch	32
input	224 x 224 x 3
momentum	0.9
decay	0.0005
learning rate	0.001
epochs	100

4. Pruebas

Para el entrenamiento de las tres redes neuronales se utilizó tres bases de datos:

- **NTHU.-** Propuesta por Weng et al. (2016), esta base de datos contiene videos de personas con somnolencia. Lamentablemente, los videos no son de conductores reales, al contrario, son actores en una oficina con buenas condiciones de iluminación.
- **SISO-IMG.-** En el proyecto se construyo esta base de datos de varios videos recolectados de una empresa de transporte. A diferencia de NTHU, esta base de datos contiene videos de conductores reales durante su jornada laboral.
- **SISO-NTHU.-** Esta es la unión de la base de datos NTHU y SISO-IMG.

En la Tabla 4, detallamos la cantidad de muestras por clase de cada base de datos. En este caso, la base de datos SISO-IMG, tiene pocas muestras debido a las pocas grabaciones que se hizo en la empresa de transporte (debido a la pandemia causada por el COVID-19).

Tabla 4: Cantidad de muestras por clase y base de datos utilizadas en los experimentos.

Base de datos	Dormido	Despierto
NTHU	8995	8995
SISO-IMG	933	933
SISO-NTHU	9928	9928

4.1. Métricas

Para medir el desempeño de cada red neuronal se utilizo el acierto (*accuracy*) y la matriz de confusión. El acierto es definido por la Ecuación 1 y la matriz de confusión representa los valores de la Tabla 5.

$$acc = \frac{TN + TP}{TP + FP + TN + FN} \quad (1)$$

donde, TN = *True Negative*, TP = *True Positive*, FP = *False Positive* y FN = *False Negative*.

Tabla 5: Matriz de confusión.

	Actual Positive(1)	Actual Negative(0)
Predicted Positive(1)	TP	FP
Predicted Negative(0)	FN	TN

Una comparativa del acierto obtenido por cada red neuronal es detallado en la Tabla 6, en este caso, la red neuronal Inception-v3, obtuvo los mejores resultados. Luego, en las Tablas 7, 8 y 9, mostramos

la matriz de confusión de la red neuronal VGG-16, ResNet e Inception-v3 respectivamente. De estos resultados, verificamos que Inception-v3 supera a VGG-16 y ResNet. También, vemos que el acierto para la base de datos NTHU es bajo con respecto a las otras bases de datos.

Tabla 6: Comparativa del acierto de cada red neuronal por base de datos.

Base de datos	VGG-16	ResNet	Inception-v3
NTHU	0.78	0.62	0.88
SISO-IMG	0.90	0.59	0.96
SISO-NTHU	0.81	0.65	0.87

Tabla 7: Matriz de confusión de la red neuronal VGG-16 en las bases de datos NTHU, SISO-IMG y SISO-NTHU.

	NTHU		SISO-IMG		SISO-NTHU	
	Somnolencia	Despierto	Somnolencia	Despierto	Somnolencia	Despierto
Somonolencia	1211	588	179	8	1518	468
Despierto	195	1604	26	161	290	1696

Tabla 8: Matriz de confusión de la red neuronal ResNet en las bases de datos NTHU, SISO-IMG y SISO-NTHU.

	NTHU		SISO-IMG		SISO-NTHU	
	Somnolencia	Despierto	Somnolencia	Despierto	Somnolencia	Despierto
Somonolencia	1473	326	112	75	1640	346
Despierto	1050	749	78	109	1026	960

Tabla 9: Matriz de confusión de la red neuronal Inception-v3 en las bases de datos NTHU, SISO-IMG y SISO-NTHU.

	NTHU		SISO-IMG		SISO-NTHU	
	Somnolencia	Despierto	Somnolencia	Despierto	Somnolencia	Despierto
Somonolencia	1498	301	179	8	1608	378
Despierto	119	1680	8	179	124	1862

4.2. Resultados

En la Figura 7, mostramos como es el funcionamiento de SISO-VID evaluado en la base de datos creada para este proyecto. Esta base de datos tiene muestras de conductores reales en su rutina normal de trabajo de una empresa de transporte. Adicionalmente, en la Figura 8, mostramos como es el funcionamiento de SISO-VID, pero esta vez evaluado en la base de datos NTHU. La base de datos NTHU, tiene muestras de videos de personas con somnolencia en una oficina con buenas condiciones de iluminación y calidad de imagen.



Figura 7: SISO-VID en ejecución en la base de datos creada para el proyecto.







Figura 8: SISO-VID en ejecución en la base de datos NTHU.

5. Conclusiones

En este informe se a detallado las pruebas realizadas para la detección de somnolencia a partir de la imagen del rostro del conductor. Las pruebas consistieron en evaluar el desempeño de tres redes neuronales: VGG-16, ResNet e Inception-v3. Se evaluó el desempeño de cada red en tres bases de datos: NTHU, SISO-IMG y SISO-NTHU.

En todas las bases de datos, la red neuronal Inception-v3 obtuvo el mejor desempeño, con un acierto de 87 % en la base de datos SISO-NTHU, 88 % en la base de datos NTHU y 96 % en la base de datos SISO-IMG. Esto demostro, que es factible detectar somnolencia utilizando aprendizaje profundo, incluso en entornos reales (SISO-IMG).

	Nombre	Cargo	Firma	Fecha
Elaborado por:	Jason Paul Cahua-na Nina	Equipo técnico - Tesis- ta La Salle		12/03/2021
	Karla Mariel Fernández Fabián	Investigador y desarro- llador - Área de inno- vación y desarrollo tec- nológico		12/03/2021
	Vicente Enrique Machaca Arceda	Investigador y desarro- llador - Área de inno- vación y desarrollo tec- nológico		12/03/2021
Revisado por:	Medardo Delgado Paredes	Investigador La Salle		12/03/2021
	Antonio Simon Bo- livar Paredes	Coordinador adminis- trativo del proyecto		12/03/2021
Aprobado por:	Elvis Diego Supo Colquehuanca	Coorinador general del proyecto		12/03/2021

Referencias

- ASIRT (2020). Annual global road crash statistics. <https://www.asirt.org/safe-travel/road-safety-facts/>.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.
- Weng, C.-H., Lai, Y.-H., and Lai, S.-H. (2016). Driver drowsiness detection via a hierarchical temporal deep belief network. In *Asian Conference on Computer Vision*, pages 117–133. Springer.
- WHO (2013). World health organization - infographics on global road safety 2013. https://www.who.int/violence_injury_prevention/road_safety_status/2013/facts/en/.