UNIVERSIDAD NACIONAL DEL LITORAL

Facultad de Ingeniería y Ciencias Hídricas

Medición de Calibración en Modelos Convolucionales para la Clasificación de Melanomas.

Integrantes

- Emiliano E. Kalafatic
- Nicolás Giuliano
- Martín Córdoba

Introducción

Un melanoma maligno tiene la contrapartida de que, si es pronosticado como tal de forma tardía, será muy complicado su tratamiento. Además, la mayoría de los pacientes, concurren a un médico en las últimas etapas, cuando los melanomas ya evidencian características malignas como son el tamaño, color (más oscuro, variaciones, entre otros) y sí probablemente han llegado al sangrado. Es importante, entonces, que ante las mínimas cualidades sospechosas detectadas en algún lunar, se recurra a un especialista para un análisis.

Aun así, posterior a la visita al especialista en piel, se requerirá también hacer un seguimiento del estado del melanoma, se haya o no identificado el lunar en cuestión como maligno durante la primera visita, ya que los lunares malignos tienen la característica de evolucionar y acrecentar sus características negativas con el paso del tiempo. Cuando se da este periodo de tiempo para observar la evolución de dicha lesión, el médico no tiene acceso al paciente hasta la siguiente revisión.

Sin embargo, se puede utilizar una herramienta de control que detecte si la lesión ha tenido cambios significativos. De esta manera, este trabajo tiene el objetivo de desarrollar un método que le sirva de apoyo tanto a una persona para analizar el estado de un melanoma y acudir a un médico en caso de síntomas sospechosos, o bien, que sirva de apoyo al análisis rápido de melanomas por parte de un profesional en la salud.

En este trabajo, se planteó el uso de diferentes modelos de redes neuronales convolucionales (o CNN) con la finalidad de clasificar melanomas malignos provenientes de imágenes de la base de datos proporcionada por *The International Skin Imaging Collaboration (ISIC)*[1], disponible online.

Adicionalmente se obtuvieron y analizaron una serie de medidas que describen tanto las características de cada red como así también si su performance a la hora de entrenar y clasificar. Dichas medidas son: *accuracy, loss, número de parámetros de la red, tiempo de entrenamiento, y medida Brier Score*, que mide la precisión de las predicciones probabilísticas de una red.

Obtención y Particionado de Dataset

El dataset cuenta con 2000 imágenes en total, del cual se realizaron tres subdivisiones que consisten en las imágenes para entrenamiento, validación y testeo, como se aprecia en la Figura 1. En la Figura 2 se muestran algunos ejemplos de imágenes con melanomas etiquetados.

[1] ISIC Archive: https://www.isic-archive.com/#!/topWithHeader/onlyHeaderTop/gallery?filter=%5B%5D

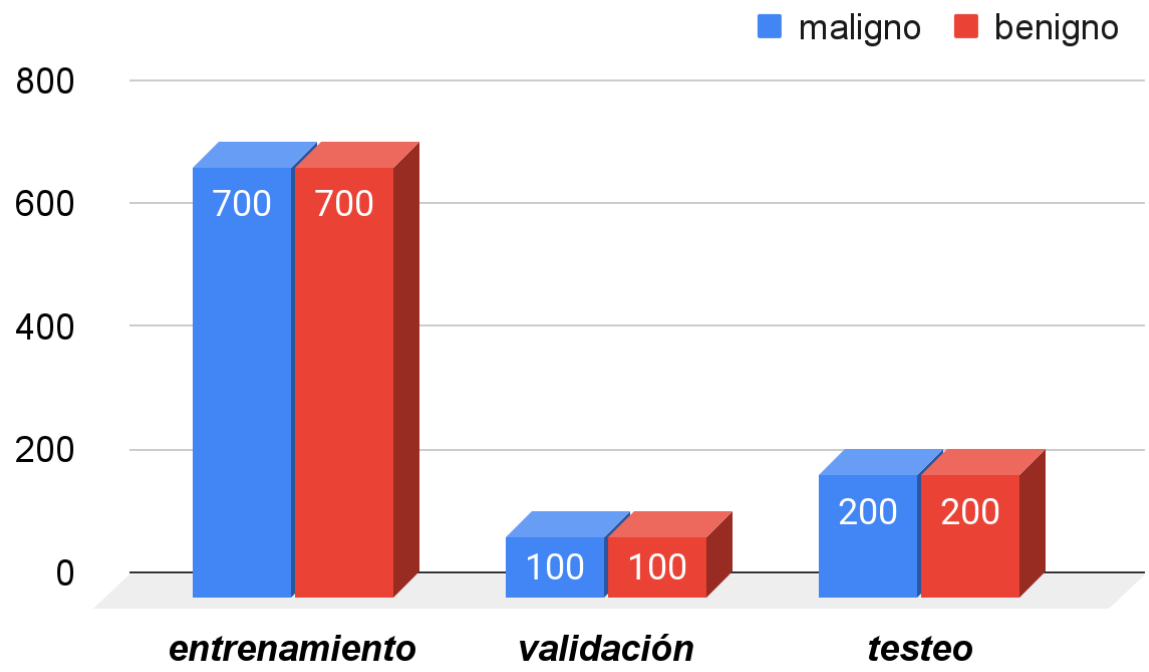

Figura 1: Cantidad de imágenes y distribución de clases en el dataset. *Azul*: Imágenes etiquetadas como benignos. *Rojo*: Imágenes etiquetadas como maligno.

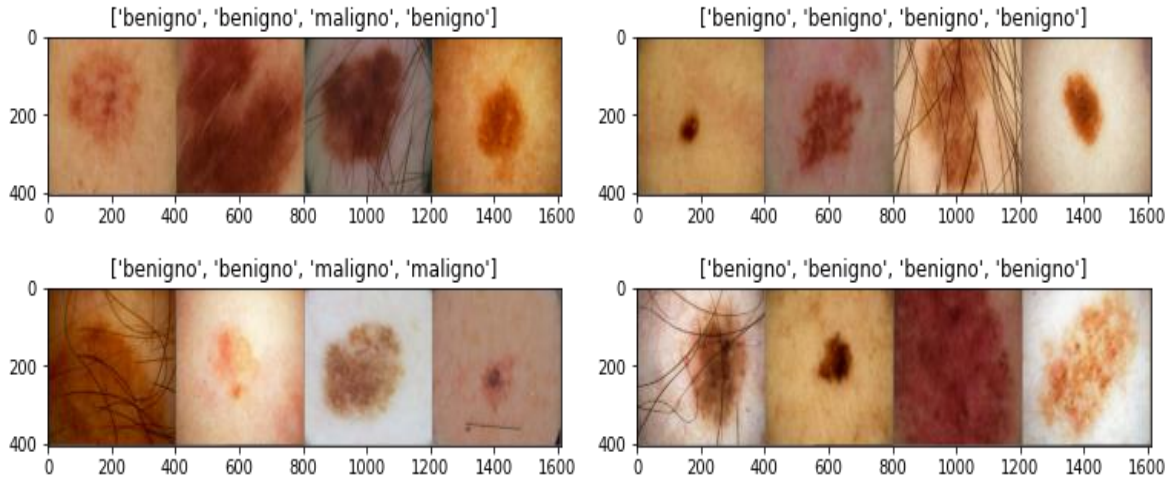

Figura 2: imágenes de ejemplo del dataset utilizado.

Experimentaciones

Los modelos de CNN que se analizaron y probaron se encuentran disponibles en las librerías de *PyTorch*[2]. Los mismos son: *ResNet18, ResNet152, GoogleNet, AlexNet, VGG11, VGG16, SqueezeNet, DenseNet, MobileNet v3 small, MobileNet v3 large,*

[2] Modelos de Torchvision: https://pytorch.org/vision/stable/models.html

EfficientNet b0 y EfficientNet b3. Para cada una se emplearon pesos pre-entrenados y todas fueron implementadas en la plataforma Google Colaboratory[3].

Procedimiento

Para cada modelo se realizó un entrenamiento con 20 épocas. Adicionalmente se utilizó un *Batch Size* de 8 y un *Learning Rate* de 0.001. Por otro lado, se eligió a la *entropía cruzada* como función de pérdida, y a la función de *Gradiente Descendiente* como función de optimización del error. Estos valores de hiperparámetros y funciones fueron elegidas debido a que son las recomendadas por cada modelo según la documentación de PyTorch.

Calibración Probabilística de la clasificación

Al realizar la clasificación, a menudo se desea no solo predecir la etiqueta de cada clase, sino también obtener una probabilidad de la etiqueta respectiva. Esta probabilidad provee algún tipo de confianza en la predicción. De esta manera, este proceso permite calibrar mejor las probabilidades de un modelo dado, o agregar soporte para la predicción de probabilidad.

Los clasificadores bien calibrados son clasificadores probabilísticos para los cuales la salida de un modelo puede interpretarse directamente como un nivel de confianza. Por ejemplo, un clasificador binario bien calibrado debe clasificar las muestras de tal manera que entre las muestras a las que dio un valor de clasificación cercano a 0,8, aproximadamente el 80% pertenezca realmente a la clase positiva.

En este trabajo, para analizar la calibración de cada modelo se utilizaron librerías de *sklearn*[4], que son compatibles con los modelos de PyTorch utilizados.

[3] Google Colaboratory:https://colab.research.google.com/

[4] Librerias para Calibración probabilística: https://scikit-learn.org/stable/modules/calibration.html

Gráficas de Calibración y Brier Scores

En las Figuras 3, 4 y 5 se esquematizan cada una de las redes evaluadas, agrupadas en tres gráficas con cuatro modelos cada una y su correspondiente medición del Brier Scores[5].

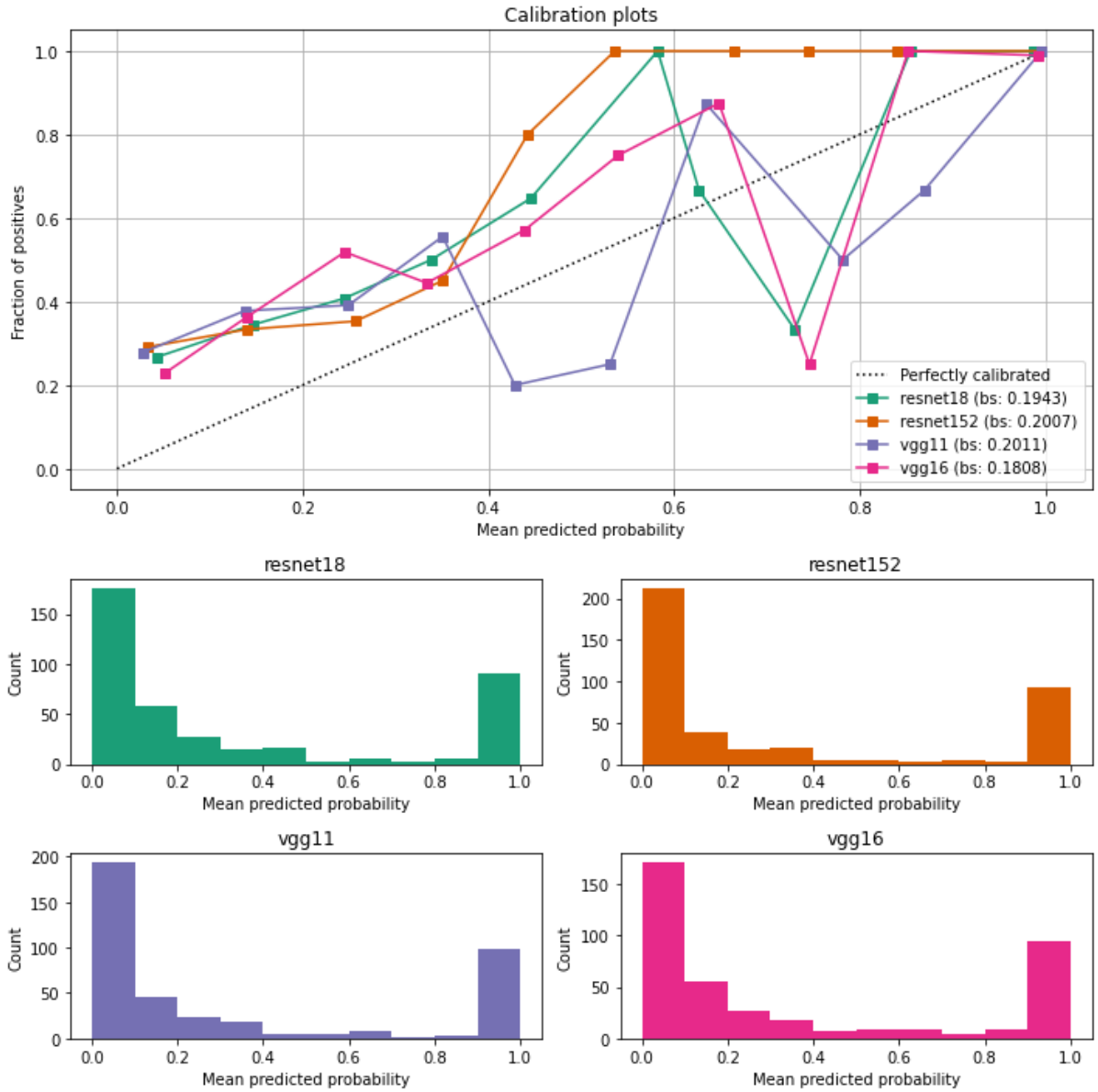

Figura 3: Gráficas de calibración de las redes ResNet 18, ResNet152, VGG11 y VGG16.

[5] Brier Score: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.brier_score_loss.html

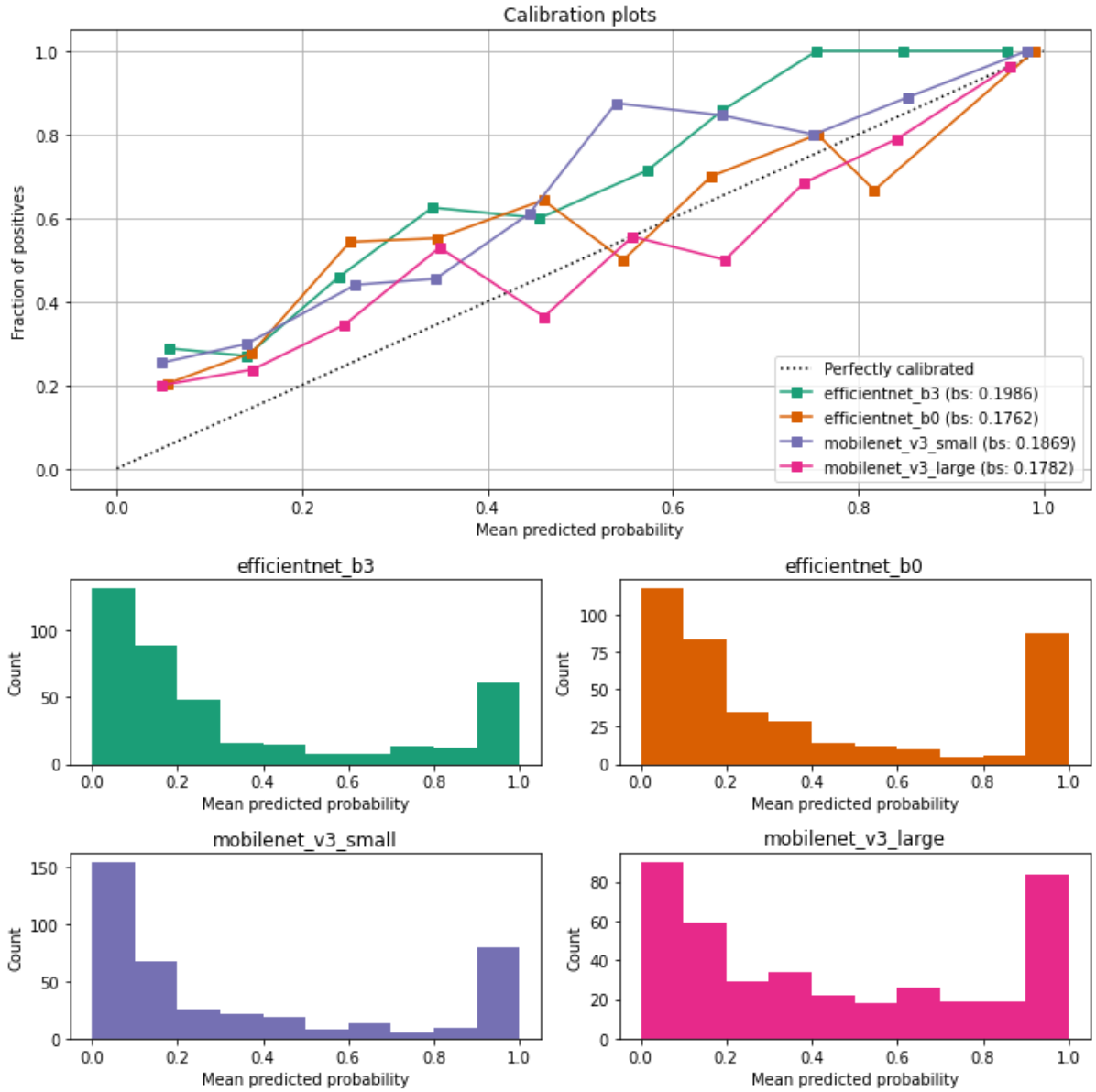

Figura 4: Gráficas de calibración de las redes EfficientNet b0 y b1 y MobileNet Small y Large.

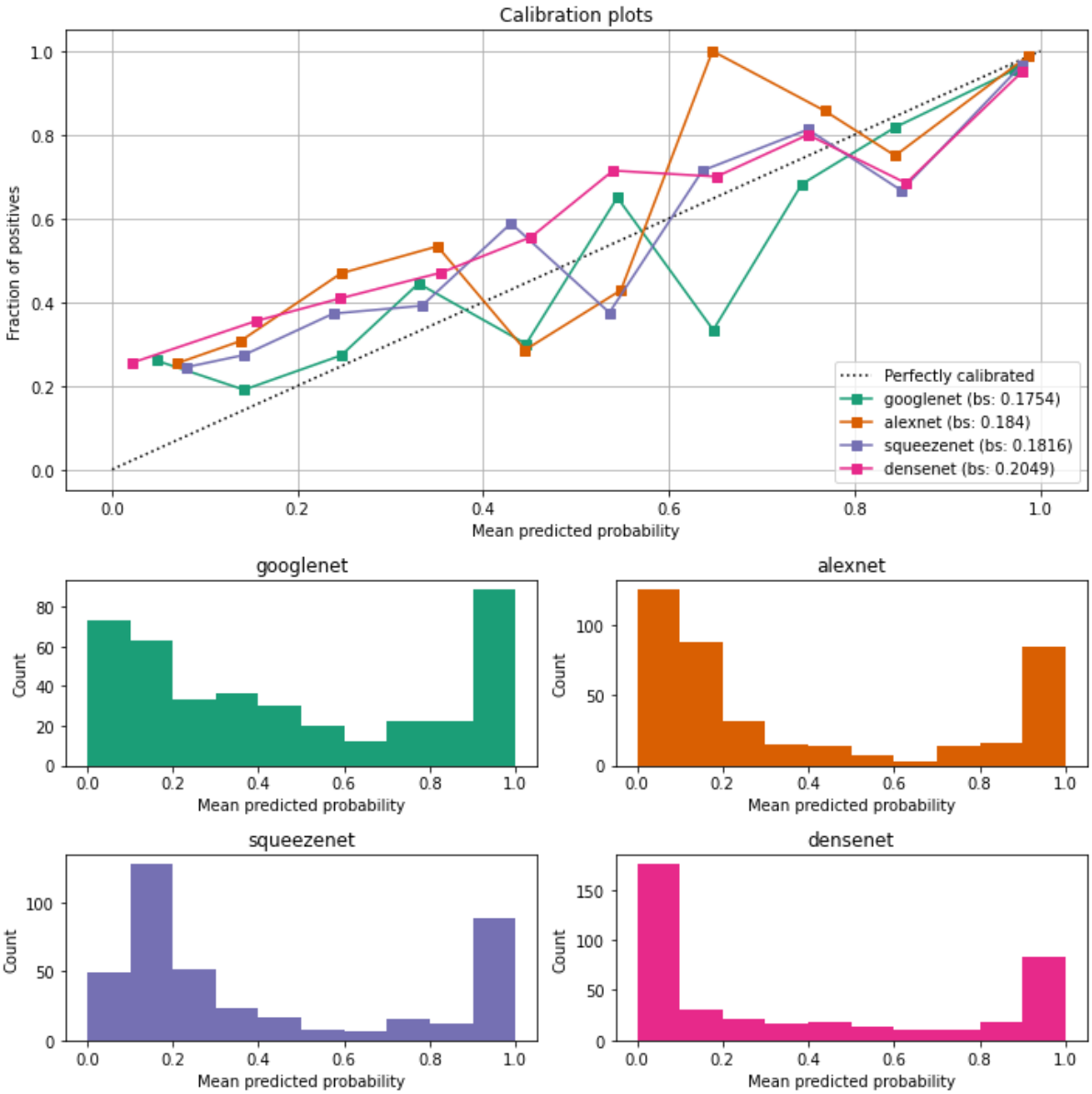

Figura 5: Gráficas de calibración de las redes GoogleNet, AlexNet, SqueezeNet y DenseNet.

Resultados

A continuación, en la Tabla 1 se muestran los resultados obtenidos tras los entrenamientos durante 20 épocas de las 12 redes convolucionales.

Red	Accuracy	Loss	Brier Scores	Training Time	Params
SqueezeNet	74.5	0.6142	0.1816	26 min	1.2 M
MobileNet v3 Small	76.0	0.6769	0.1869	24 min	2.5 M
EfficientNet b0	74.2	0.6606	0.1762	30 min	5.3 M
MobileNet v3 Large	74.5	0.5970	0.1782	26 min	5.4 M
GoogleNet	76.2	0.6418	0.1754	29 min	7.0 M
DenseNet	74.5	0.6774	0.2049	46 min	7.2 M
Resnet 18	75.0	0.7218	0.1943	28 min	11.5 M
EfficientNet b3	73.5	0.6972	0.1986	44 min	12.1 M
Resnet 152	76.5	0.8225	0.2007	114 min	60.3 M
AlexNet	75.7	0.6871	0.1840	23 min	62.3 M
VGG11	75.7	0.7059	0.2011	50 min	133.0 M
VGG16	77.0	0.6468	0.1808	91 min	138.4 M

Tabla 1: Resultados Experimentales

Conclusiones

Tras las experimentaciones se pudo observar que es posible clasificar las imágenes con un accuracy de entre 0.74 y 0.77. Posiblemente mediante la aplicación de ajustes finos se puede mejorar aún más estos resultados.

Por otro lado, a priori se consideró que las redes de mayor tamaño (las que poseen mayor número parámetros) serían las que posean una mayor descalibración, respecto a redes más pequeñas. Tras observar los resultados esta hipótesis no se cumplió completamente (solo en algunos casos), pero se estima que esto se debe a la baja cantidad de épocas de entrenamiento.

Trabajos Futuros

Se plantearon como trabajos futuros utilizar un dataset con mayor cantidad de imágenes, entrenar las redes con mayor cantidad de épocas, probar diferentes valores de los hiperparámetros definidos y experimentar con los que no fueron analizados, como así también entrenar los modelos desde cero (sin usar pre-entrenamiento). Además, se podría probar resolver el problema mediante otros

modelos y arquitecturas diferentes. Todos con la misma finalidad de mejorar la calidad de clasificación y la calibración probabilística de los modelos estudiados.