Universidad Internacional de Valencia

Aprendizaje Semisupervisado en imagen médica – Mean Teacher

Titulación:
Máster en Inteligencia Artificial

Curso académico
2022-2023

Alumno/a: Pérez de Mendiola Rubio, Mikel
D.N.I: 72747678N

Director/a de TFM: Karen López-Linares Román

Convocatoria:

Extraordinaria

De:
Planeta Formación y Universidades

4 de noviembre de 2023

Índice general

Índice de figuras

Índice de tablas

Resumen

En el ámbito de la medicina, la detección de enfermedades a partir de imágenes de radiológicas es esencial para una detección temprana y tratamiento efectivo. Con el avance en las técnicas de inteligencia artificial y aprendizaje profundo, se ha buscado mejorar la precisión de dicha detección.

Este Trabajo Fin de Máster explora la implementación y eficacia de la arquitectura de red neuronal Mean Teacher en la detección automática de enfermedades pulmonares utilizando imágenes médicas. Se llevó a cabo una revisión exhaustiva de las técnicas actuales basadas en modelos teacher-student y se implementó una arquitectura Mean Teacher, cuyos resultados se compararon con otra baseline puramente supervisada. Los hallazgos mostraron que Mean Teacher ofrece mejoras en la tarea de clasificación en escenarios donde se dispone de pocos ejemplos anotados por especialistas.

Se discuten las ventajas y limitaciones de esta arquitectura en el ámbito de diagnóstico usando imágenes médicas y se proponen recomendaciones para trabajos futuros.

Abstract

In the field of medicine, the detection of diseases from radiological images is essential for early diagnosis and effective treatment. With advancements in artificial intelligence and deep learning techniques, efforts have been made to improve the accuracy of such detections.

This Master's Thesis explores the implementation and efficacy of the Mean Teacher neural network architecture in the automatic detection of pulmonary diseases using medical images. An exhaustive review of current techniques based on teacher-student models was conducted, and a Mean Teacher architecture was implemented. Its results were compared with a purely supervised baseline. Findings showed that the Mean Teacher provides improvements in the classification task in scenarios with few examples annotated by specialists.

The advantages and limitations of this architecture in the field of medical imaging diagnosis are discussed, and recommendations for future work are proposed.

Introducción 1

1.1. Justificación

El gran avance tecnológico de los últimos años, junto con la creciente disponibilidad de datos, ha impulsado a científicos y grandes empresas a buscar soluciones más eficientes y precisas para abordar problemas complejos en diversos campos.

En la era de internet y con gran parte de nuestras labores digitalizadas, la cantidad de datos recopilados ha crecido de forma notable y, sin embargo, a pesar de la abundancia de datos disponibles una gran proporción de estos carece de etiquetas.

Esta situación presenta un desafío, ya que dificulta su uso en métodos tradicionales de aprendizaje supervisado. En este escenario, el aprendizaje semisupervisado se presenta como una solución prometedora, capaz de aprovechar tanto los datos etiquetados como los no etiquetados.

La eficiencia en el uso de recursos es una de las principales ventajas de este enfoque, ya que etiquetar grandes conjuntos de datos es una tarea que requiere tiempo, esfuerzo y, en muchos casos, un amplio conocimiento del campo o materia específica. Por lo tanto, reducir la dependencia de grandes conjuntos de datos etiquetados puede traducirse en ahorros significativos en términos de recursos y tiempo. Además, al combinar la información de datos etiquetados y no etiquetados, se abre la posibilidad de mejorar la precisión y robustez de los modelos, especialmente en escenarios donde los datos etiquetados son limitados. Esta mejora en la precisión es especialmente relevante en aplicaciones del mundo real.

En el ámbito de la imagen médica se dispone de grandes volúmenes de datos mayoritariamente no etiquetados, y la capacidad de utilizar estos datos para mejorar los modelos tiene un gran valor.

En este dominio concreto la dificultad de contar con radiólogos para realizar anotaciones precisas y confiables es una problemática considerable. El etiquetado de imágenes médicas no puede depender de anotaciones generadas por individuos no expertos o a través de métodos de “crowd annotation”, debido a la especialización y conocimiento específico requerido. Además, la variabilidad en la adquisición de imágenes médicas entre diferentes hospitales usando diferentes equipos radiológicos, así como la variabilidad demográfica presentan obstáculos adicionales.

Para mitigar este problema es esencial trabajar en la adaptación de dominio de los modelos. El aprendizaje semisupervisado surge como una estrategia valiosa en este contexto,

permitiendo que los modelos se beneficien de usar un conjunto más amplio de datos, que incluye información no etiquetada del dominio específico al que se desea adaptar el modelo. Este enfoque se presenta como una alternativa más viable y eficiente en comparación con el reentrenamiento completo de los modelos utilizando únicamente datos etiquetados.

Numerosos estudios descritos en la literatura han intentado mejorar la capacidad diagnóstica aplicando diferentes técnicas de aprendizaje semisupervisado, entre otros el teacher-student. Esto representa una dirección innovadora en el campo de la inteligencia artificial, y su investigación puede conducir a avances significativos en la disciplina.

Finalmente, cabe destacar la trascendencia social que conllevan los avances en campos como la medicina, donde las decisiones basadas en modelos de aprendizaje automático pueden tener un impacto directo en la vida de las personas. Por ello es esencial contar con modelos precisos y confiables que ayuden a los profesionales en el diagnóstico, redundando en un beneficio para los pacientes. El aprendizaje semisupervisado, al proporcionar una mayor precisión y confiabilidad, puede desempeñar un papel crucial en este aspecto.

1.2. Marco teórico

Para contextualizar el aprendizaje semisupervisado se pasa primero a explicar en qué consiste el aprendizaje automático, paradigma en el que se encuadra el mismo, y comentar brevemente qué tipos de aprendizaje automático se encuentran en la literatura. Estos conceptos se obtienen de Zhu y Goldberg (2009).

1.2.1. Aprendizaje automático

El aprendizaje automático es una rama de la inteligencia artificial cuyo fin es desarrollar algoritmos y modelos que permitan a los ordenadores realizar tareas para las que no han sido explícitamente programadas.

Esto se consigue haciendo que las máquinas aprendan a partir de los datos proporcionados, y vayan mejorando su rendimiento con la experiencia. Por tanto, el aprendizaje automático desarrolla técnicas que permiten a los ordenadores realizar generalizaciones a partir de datos, que permitan usarse en tareas como clasificación, inferencia, segmentación o clustering.

Se suelen diferenciar tres tipos dentro del aprendizaje automático, los cuales se describen a continuación.

Aprendizaje supervisado

En este caso, el modelo se entrena con un conjunto de datos para los que se conocen las etiquetas de éstos. El objetivo es aprender una función capaz de predecir el valor de salida o etiqueta para datos no vistos. Se suele usar para problemas de regresión, clasificación, segmentación y detección.

Aprendizaje no supervisado

En el aprendizaje no supervisado no se tienen las etiquetas de los datos, por lo que aquí el objetivo es encontrar las estructuras o patrones que subyacen en los datos, de forma que se pueda realizar agrupaciones de datos (clustering) o reducción de dimensionalidad.

Aprendizaje semisupervisado

El aprendizaje semisupervisado combina los anteriores enfoques, supervisado y no supervisado. Es especialmente útil cuando se dispone de un volumen de datos muy grande, donde muchos de estos no están etiquetados. El objetivo de este tipo de aprendizaje es poder hacer uso de este gran número de datos no anotados junto con los etiquetados para mejorar el rendimiento de los modelos.

En imagen médica es relevante aplicar este tipo de técnicas de aprendizaje, ya que en este dominio es habitual tener gran cantidad de imágenes y se dispone de muy pocas muestras anotadas por especialistas.

1.2.2. Enfoques dentro del aprendizaje semisupervisado

Dentro del aprendizaje semisupervisado se pueden diferenciar tres enfoques atendiendo a cómo se intenta aprovechar esa gran cantidad de datos no etiquetados para intentar mejorar el aprendizaje de los modelos semisupervisados.

Estos son el consistency regularization, el pseudo labeling y los métodos generativos.

1.2.2.1. Consistency regularization

El principio subyacente de la regularización de consistencia radica en la premisa esencial de que un modelo debería generar salidas coherentes y consistentes ante entradas idénticas o notablemente similares (Fan et al., 2021), incluso si estas últimas han sido sometidas a diversas transformaciones o perturbaciones mediante técnicas refinadas de aumento de datos (data augmentation). En este contexto, el objetivo principal de la regularización de consistencia es minimizar activamente las discrepancias y diferencias entre las predicciones dadas por el modelo para una entrada específica y aquellas correspondientes a versiones alteradas o perturbadas de la misma entrada.

En el terreno de la regularización de consistencia, se puede observar la emergencia y aplicación de múltiples métodos que incorporan este principio fundamental como base sólida para el entrenamiento de modelos. Dentro de este espectro, destaca el enfoque **Teacher-Student** (Yalniz et al., 2019). Este método peculiar opera mediante el mantenimiento de dos versiones distintas pero relacionadas del mismo modelo: el teacher (maestro) y el student (estudiante). Ambos modelos reciben la misma entrada, aunque cada una es perturbada de manera diferente. La meta central de este enfoque es lograr que las predicciones generadas por el modelo student se alineen y sean consistentes con aquellas producidas por el modelo teacher. En este escenario, el modelo student se somete a actualizaciones periódicas, guiadas por métodos

de descenso de gradiente aplicados a la función de pérdida, mientras que el modelo teacher permanece estático o experimenta actualizaciones con una frecuencia reducida.

Una variante importante y objeto de exploración en este trabajo es el método **Mean Teacher** (Tarvainen y Valpola, 2017). Esta técnica se presenta como una extensión innovadora del enfoque Teacher-Student. A diferencia de la versión original, el Mean Teacher no mantiene estático al modelo teacher. En cambio, este se actualiza continuamente mediante un promedio móvil de las actualizaciones realizadas al modelo student, manteniendo los pesos EMA (media móvil exponencial) de este último. Este proceso de actualización constante contribuye a suavizar las predicciones del modelo, lo cual, a su vez, tiene un impacto positivo y notable en el rendimiento general del sistema.

Otro método de interés es el **Noisy Student** Wang et al. (2020). En este enfoque, el modelo student se entrena con datos adicionales generados previamente por un modelo teacher de mayor envergadura, mientras que durante el entrenamiento se introduce ruido de manera intencional para amplificar la capacidad de generalización del modelo.

El método **Temporal Ensembling**, por su parte, considera las predicciones anteriores del modelo para cada entrada durante el proceso de entrenamiento. Este método calcula un promedio de las predicciones previas y genera un pseudolabel que sirve como base para la regularización de las predicciones actuales del modelo.

Finalmente, los métodos **MixMatch** y **ReMixMatch** representan estrategias híbridas que integran una variedad de técnicas, incluida la regularización de consistencia. MixMatch integra entradas etiquetadas y no etiquetadas para generar pseudolabels para estas últimas, mientras que ReMixMatch implementa la regularización de consistencia mientras realiza transformaciones específicas en los datos. Estas técnicas combinadas ofrecen un abanico de posibilidades y herramientas para mejorar la coherencia y precisión de los modelos en el aprendizaje semisupervisado.

1.2.2.2. Pseudo labeling

El enfoque de Pseudo labeling (Lee, 2013) constituye una estrategia en la cual se utiliza un modelo previamente entrenado con el propósito de generar etiquetas, conocidas como pseudoetiquetas, para aquellos datos que originalmente carecen de ellas. Estas pseudoetiquetas generadas son luego implementadas como herramienta fundamental para el reentrenamiento del modelo en cuestión. En este proceso iterativo y refinado de reentrenamiento, el modelo se ve progresivamente beneficiado al tener a su disposición un volumen creciente de datos etiquetados, optimizando así su desempeño y precisión a medida que avanza en las fases del entrenamiento.

En el universo de estrategias que incorporan el pseudolabeling, emergen con particular relevancia algunas variantes de modelos, siendo cada uno de estos distintivos por sus características y métodos de implementación. Se exponen algunas a continuación (Livieris et al., 2018).

El modelo de **Self-training** es una de estas variantes prominentes. En este enfoque, inicial-

mente, se lleva a cabo el entrenamiento del modelo utilizando exclusivamente los datos que ya poseen etiquetas. Posteriormente, el modelo ya entrenado se emplea como instrumento para asignar etiquetas a aquellos datos que no las tienen. Finalmente, se procede a reentrenar el modelo, esta vez utilizando un conjunto de datos ampliado que incluye tanto los datos originalmente etiquetados como aquellos que recibieron pseudoetiquetas a través del proceso anterior.

Por otro lado, el enfoque **Co-training** presenta una dinámica distinta. Este método se fundamenta en el entrenamiento paralelo y coordinado de dos modelos distintos. En este proceso de aprendizaje colaborativo, cada modelo asume la responsabilidad de generar pseudoetiquetas para los datos no etiquetados del otro, creando un ciclo de retroalimentación y aprendizaje mutuo entre ambos.

Una extensión natural y avanzada del Co-training es el enfoque conocido como **Tri-training**. Este método involucra no dos, sino tres modelos trabajando en conjunto. En esta configuración, cuando un dato no etiquetado recibe la misma etiqueta predicha por al menos dos de los tres modelos involucrados, dicha etiqueta es entonces validada y adoptada como pseudoetiqueta del dato en cuestión. Este mecanismo introduce una capa adicional de verificación y consenso en la generación de pseudoetiquetas, fortaleciendo la confiabilidad y precisión de las etiquetas generadas en el proceso.

1.2.2.3. Modelos generativos

En el ámbito de los modelos generativos, se encuentra una especialización que busca meticulosamente aprender la distribución de probabilidad inherente a los datos con los que se está entrenando. Este aprendizaje profundo e intrínseco de la distribución probabilística tiene como finalidad principal la generación de datos sintéticos novedosos. Estos datos recién creados y sintéticos, cuando se integran en el proceso de entrenamiento, potencian significativamente el desempeño del modelo, optimizando su capacidad de análisis y predicción.

En el contexto particular del aprendizaje semisupervisado, diversos modelos generativos se despliegan con usos y aplicaciones específicas. Se exponen a continuación los modelos más destacados (Gm et al., 2020).

Las **Redes Generativas Antagónicas, o GANs** (por sus siglas en inglés, Generative Adversarial Networks), ocupan un lugar destacado. Las GANs se han adaptado meticulosamente para que, durante el proceso de entrenamiento, el discriminador no solo distinga entre muestras reales y generadas (falsas), sino que también emprenda tareas de clasificación de manera efectiva y precisa. Esta dualidad en las funciones del discriminador en las GANs refuerza su aplicabilidad y eficacia en problemas de aprendizaje semisupervisado.

Del mismo modo, los **Autoencoders Variacionales (VAE, por sus siglas en inglés, Variational Auto Encoders)** representan otra herramienta vital en este dominio. Estos se emplean con el objetivo preciso de aprender representaciones que sean no solo robustas, sino también útiles de los datos no etiquetados disponibles. Posteriormente, estas representaciones aprendidas se aplican de manera directa y eficiente en diversas tareas de clasificación, sirviendo

como un puente entre los datos no estructurados y los algoritmos clasificatorios.

Adicionalmente, en el amplio y cambiante panorama de los modelos generativos, se encuentran otras redes que, aunque menos conocidas, juegan un papel crucial en el aprendizaje semisupervisado. Ejemplo de estas son las Redes de Creencia Profunda (Deep Belief Networks) y las Máquinas de Bolzmann Restringidas (RBM, por sus siglas en inglés, Restricted Boltzmann Machines). Estos modelos, con sus características y particularidades únicas, también se incorporan activamente en la resolución de problemas relacionados con el aprendizaje semisupervisado, contribuyendo al avance y la innovación constantes en este campo de estudio esencial y dinámico.

Estos tres enfoques difieren en el concepto empleado, pero tienen como objetivo común aprovechar al máximo los datos disponibles, tanto etiquetados como sin etiquetar, para construir modelos robustos y con buen desempeño.

Dentro de los modelos de aprendizaje semisupervisado, este trabajo se centra en el enfoque de regularización de consistencia, en concreto en los basados en el modelo teacher-student, donde se tienen por lo general dos modelos, un student y un modelo teacher.

La idea central detrás de este paradigma es que el modelo teacher actúa como una referencia para el modelo student, proporcionando objetivos suaves o pseudo-etiquetas basadas en sus propias predicciones. Estos objetivos suaves, que se obtienen por lo general a partir de una versión promediada o más suavizada del modelo teacher, se utilizan para entrenar al modelo student. A medida que avanza el entrenamiento, el modelo student se beneficia del aprendizaje acumulado del modelo teacher, mientras que el modelo teacher se actualiza periódicamente para reflejar los avances del modelo student.

De esta forma se consigue aprovechar la información latente en los datos no etiquetados en escenarios donde no se disponen de suficientes datos etiquetados. Al requerir que las predicciones del modelo student sean consistentes con las del modelo teacher en datos no etiquetados, se introduce una forma de regularización que ayuda a prevenir el sobreajuste y mejora la generalización del modelo.

1.2.3. Mean teacher

El método mean teacher se presenta en 2017 en NeurIPS en el paper *Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results* (Tarvainen y Valpola, 2017) como propuesta de mejora del modelo *temporal ensembling* publicado en *Temporal Ensembling for Semi-Supervised Learning* (Laine y Aila, 2017).

Se trata de un método de *deep learning* usando aprendizaje semisupervisado que permite usar datos tanto etiquetados como no etiquetados para mejorar el aprendizaje de la red y poder hacer mejores predicciones que mejoran los resultados obtenidos solo con datos anotados en casos en los que se tienen pocos datos anotados, como suele ser el caso en imagen médica.

Conceptualmente, mean teacher consiste en dos redes, una red student y otra teacher. La red student se entrena con datos tanto etiquetados como no etiquetados, mientras que la

red teacher se va actualizando con el promedio móvil de los pesos de la red student. De esta forma, la red teacher proporciona objetivos suaves para los datos no etiquetados, ayudando as student a aprender.

Esta es la esencia del método Mean Teacher, donde el modelo teacher (promedio de los pesos del modelo a lo largo del tiempo) proporciona una referencia para el modelo student en entrenamiento.

Objetivos

2

El objetivo de este trabajo es investigar y evaluar la eficacia de la arquitectura de red neuronal Mean Teacher para la detección de neumonía en imágenes radiológicas de pulmones.

Para esto, se definen como objetivos parciales:

1. **Realizar una revisión bibliográfica exhaustiva sobre las técnicas actuales que emplean arquitecturas teacher-student.**

2. **Implementar una arquitectura Mean Teacher para clasificación de la enfermedad de neumonía en imagen radiológica médica.**

3. **Comparar el rendimiento de la arquitectura Mean Teacher con un enfoque puramente supervisado en términos de precisión, sensibilidad, ROC-AUC y falsos negativos.**

Estado del arte

3

Dentro de los modelos de aprendizaje semisupervisado, este trabajo se centra en el enfoque de regularización de consistencia, en concreto en los basados en el modelo teacher-student, donde se tienen por lo general dos modelos, un student y un modelo teacher.

La idea central detrás de este paradigma es que el modelo teacher actúa como una referencia para el modelo student, proporcionando objetivos suaves o pseudo-etiquetas basadas en sus propias predicciones. Estos objetivos suaves, que se obtienen por lo general a partir de una versión promediada o más suavizada del modelo teacher, se utilizan para entrenar al modelo student.

A medida que avanza el entrenamiento, el modelo student se beneficia del aprendizaje acumulado del modelo teacher, mientras que el modelo teacher se actualiza periódicamente para reflejar los avances del modelo student.

De esta forma se consigue aprovechar la información latente en los datos no etiquetados en escenarios donde no se disponen de suficientes datos etiquetados. Al requerir que las predicciones del modelo student sean consistentes con las del modelo teacher en datos no etiquetados, se introduce una forma de regularización que ayuda a prevenir el sobreajuste y mejora la generalización del modelo.

3.1. Aprendizaje semisupervisado en imagen médica

En el campo de la medicina está tomando especial fuerza el aprendizaje semisupervisado. La imagen médica es un dominio donde se dispone habitualmente de pocos datos correctamente etiquetados por especialistas. Se trata de un proceso costoso en cuanto a tiempo y esfuerzo, por lo que aplicar aprendizaje semisupervisado a datasets con solo unos pocos datos anotados junto con gran volumen de datos no etiquetados permite mejorar mucho el rendimiento de los modelos, y se trata de un campo de investigación muy activo, como se verá más adelante.

Combinando datos etiquetados y no etiquetados se obtiene una precisión en el diagnóstico equiparable al que se obtiene con modelos supervisados con volumen suficiente de datos, y a veces llega a superarla.

También es útil a la hora de integrar datos provenientes de diversas fuentes, como pueden ser imágenes obtenidas por resonancia magnética, tomografía computerizada, etc.

Principalmente las aplicaciones que se le dan es la de segmentación de lesiones o tumores, donde se ha visto que consigue mejorar la precisión de segmentación al aprovechas cantidades

grandes de imágenes y en la clasificación para diagnóstico de enfermedades.

En la literatura podemos encontrar varios usos del método Mean Teacher y de otros métodos basados en arquitecturas teacher-student en el dominio de la imagen médica, usados sobre todo para tareas de segmentación y clasificación.

El paper *Uncertainty-aware Self-ensembling Model for Semi-supervised 3D Left Atrium Segmentation* (Yu et al., 2019) propone el método Mean Teacher en segmentación del atrio izquierdo del corazón a partir de imágenes de resonancia magnética en 3D. Además, se emplea un esquema de incertidumbre para que el student aprenda de forma gradual de los objetivos significativos y confiables, de forma que no todas las predicciones del modelo teacher se tratan por igual: aquellas que tienen baja incertidumbre se consideran más confiables y por tanto se usan para que aprenda el student. En los indicadores de la calidad de segmentación se obtiene que el modelo semisupervisado alcanza un índice Dice de 88.88 frente al 86.03 obtenido usando V-Net Bayesianas.

En *Transformation-consistent Self-ensembling Model for Semi-supervised Medical Image Segmentation* (Li et al., 2020) se utiliza el método Mean Teacher para mejorar aún más la regularización a la hora de segmentar lesiones cutáneas (dermoscopias), discos ópticos en imágenes de fondo de ojo y tomografías de hígado. Se extiende la transformación y se optimiza la pérdida de consistencia con un modelo teacher. Este modelo teacher es un promedio de los pesos del modelo student. El método propuesto en el paper muestra una mejora en todas las métricas evaluadas en comparación tanto con el método supervisado simple como con el supervisado con regularización. Las mejoras son especialmente notables en la Exactitud de Jaccard, el Índice de Dice y la Sensibilidad, que son métricas críticas para evaluar la calidad de la segmentación en imágenes médicas.

En segmentación de lesiones cerebrales, en *Semi-supervised Brain Lesion Segmentation with an Adapted Mean Teacher Model* (Cui et al., 2019) se adapta el Mean Teacher para mejorar el rendimiento en la segmentación de las CNN cuando hay datos etiquetados limitados. El método propuesto muestra el mejor rendimiento con un coeficiente de Dice de 0.6676 ± 0.2392, que es estadísticamente significativamente superior a los otros métodos (0.6236 ± 0.2577), como lo indica la prueba t de Student emparejada ($p < 0.05$). Esto sugiere que el método propuesto ofrece una mejora significativa en la calidad de la segmentación en comparación con los otros métodos evaluados en este experimento particular.

En esta misma linea MTANS: Multi-Scale Mean Teacher Combined Adversarial Network with Shape-Aware Embedding for Semi-Supervised Brain Lesion Segmentation (Chen et al., 2021) usa Mean Teacher para segmentación de lesiones de esclerosis múltiple, de accidente cerebrovascular isquémico y de tumores cerebrales. Combina una versión mejorada de Mean Teacher y una red adversarial. Además, se ha introducido una nueva función de pérdida que alienta a los modelos student y teacher a tener características consistentes en diferentes niveles de resolución para una misma entrada. Esta pérdida de consistencia a múltiples escalas ayuda a asegurar que el modelo sea robusto y produzca predicciones coherentes en diferentes niveles de detalle. MTANS muestra mejoras usando solo el 20 % de datos etiquetados en ciertas métricas clave como el coeficiente de Dice (88.48 a 90.86) y la tasa de verdaderos

positivos (TPR) en comparación con el método baseline. Sin embargo, también muestran una disminución en otras métricas como la tasa de verdaderos positivos de localización.

Otro paper en el que se usa un método teacher-student para segmentación de lesiones de neumonía en tomografía computerizada en pacientes con COVID-19 es *A Noise-Robust Framework for Automatic Segmentation of COVID-19 Pneumonia Lesions From CT Images* (Wang et al., 2020). En este caso, el modelo teacher se actualiza de forma adaptativa de forma que se suprime la contribución del student cuando comete errores significativos. En cuanto al modelo student, también es adaptativo y aprende del teacher solo cuando el modelo teacher supera al student en rendimiento.

En *Semi-supervised Medical Image Classification with Relation-driven Self-ensembling Model* (Liu et al., 2020) se aplica a clasificación de imágenes médicas de lesiones cutáneas y de radiografías de tórax. se hace uso de un modelo de auto-ensamblado (*self-ensembling*) para producir objetivos de consistencia de alta calidad para los datos no etiquetados. Este es un concepto central del método Mean Teacher. En el dataset ISIC 2018 se compara el método propuesto (SRC-MT) y un método baseline. Ambos métodos utilizan el 20 % de las imágenes etiquetadas, pero el baseline no utiliza imágenes no etiquetadas, mientras que el método propuesto utiliza el 80 % de las imágenes no etiquetadas. La precisión no mejora significativamente (92.54 vs. 92.17), pero la sensibilidad tiene mejora de 5.97 y la AUC una mejora de 3.43 con respecto al baseline.

Para la clasificación de imágenes de cáncer colorrectal, el paper *Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images* (Yu et al., 2021) emplea Mean Teacher con imágenes WSI (whole slide images), evidenciando que con el uso de aprendizaje semisupervisado y un número limitado de imágenes etiquetadas se puede llegar al rendimiento de aprendizaje supervisado tradicional usando un número mucho mayor de etiquetas, y estando el rendimiento cercano al de patólogos humanos en la clasificación. Siendo su dataset de 5000 clases balanceadas se evaluaron escenarios donde solo una pequeña fracción de las imágenes estaban etiquetadas (5 % y 20 %). En ambos casos, los modelos SSL (Mean Teacher) superaron a los modelos SL en términos de precisión (0.96 vs. 0.91 y 0.98 vs. 0.96) , demostrando que la incorporación de datos no etiquetados en SSL puede mejorar la precisión del modelo. Incluso con una mayor proporción de datos etiquetados (80 %), los modelos SSL y SL mostraron una precisión comparable, indicando que SSL puede alcanzar un rendimiento similar a SL con menos datos etiquetados. En cuanto a la robustez, se concluye que el modelo Mean Teacher empleado en el estudio usando el 10 % de los datos etiquetados, al incluir conjuntos de datos de imágenes diferentes centros usando distinto equipamiento radiológico tiene un rendimiento similar al obtenido por otros puramente supervisados.

En clasificación de defectos del cartílago de la rodilla, el artículo *Automatic Grading Assessments for Knee MRI Cartilage Defects via Self-ensembling Semi-supervised Learning with Dual-Consistency* (Huo et al., 2022) usa Mean Teacher para detectar defectos causados por la osteoartritis. Como particularidad, este trabajo diseña una estrategia de doble consistencia para mejorar los modelos del student y el teacher haciendo que ambos modelos se centren en las mismas regiones del cartílago.

En *Teaching Semi-Supervised Classifier via Generalized Distillation* (Gong et al., 2018) se toma un enfoque de aprendizaje semi-supervisado (SSL) que formula el SSL como un problema de Destilación Generalizada (GD). En este enfoque, se introduce un "maestro"(teacher) para guiar el proceso de entrenamiento de un aprendiz (learner). El maestro posee un conocimiento privilegiado que explica los datos de entrenamiento pero que el aprendiz no conoce. El maestro transmite su conocimiento al aprendiz a través de una función de enseñanza específica. Luego, el aprendiz adquiere conocimiento ïmitando"la salida de la función de enseñanza bajo un marco de optimización. Los resultados obtenidos sobre el dataset ORL usando el 25 % de las imágenes etiquetadas mejoran la precisión en 0.02 puntos respecto al baseline supervisado usando solo los datos etiquetados.

Otro paper sugiere una mezcla de aprendizaje autosupervisado con semisupervisado usando Mean Teacher. En *Self-supervised Mean Teacher for Semi-supervised Chest X-Ray Classification* (Liu et al., 2021) se usa el modelo Mean Teacher en el *fine-tuning* para proporcionar etiquetas suaves para el caso de los ejemplos sin etiquetar, y el modelo se entrena para predecir dichas etiquetas suaves. En esta ocasión se consiguen superar los resultados del modelo completamente supervisado entrenado en el dataset Chest X-Ray14 (100 % de los datos supervisados) contra el modelo semisupervisado propuesto usando solo un 20 % de datos etiquetados.

Una propuesta para mejorar Mean Teacher, haciendo que no sea sensible a la elección de hiperparámetros se encuentra en *Semi-supervised Classification of Diagnostic Radiographs with NoTeacher: A Teacher that is Not Mean* (Unnikrishnan et al., 2020). Aquí se elimina la necesidad de una usar una red maestra y utiliza en su lugar un modelo gráfico probabilístico para mejorar la consistencia y el rendimiento. El modelo NoTeacher propuesto logra un 90 % de AUROC con respecto al baseline completamente supervisado haciendo uso de tan solo un 5 % de los datos etiquetados.

Otro ejemplo de uso de Mean Teacher en imagen médica, en este caso para clasificación de núcleos en el diagnóstico patológico de cánceres se encuentra en *Local and Global Consistency Regularized Mean Teacher for Semi-supervised Nuclei Classification* (Su et al., 2019). Los autores refieren resultados equivalentes a los obtenidos con aprendizaje supervisado usando solo entre el 5 % y el 25 % de los datos etiquetados.

Para clasificación de enfermedades cutáneas para detección de melanoma el artículo *Skin lesion classification with ensemble of squeeze-and-excitation networks and semi-supervised learning* (Kitada y Iyatomi, 2018) hace uso de Mean Teacher consiguiendo muy buenos resultados con un accuracy del 87.2 % en el dataset de validación de ISIC2018.

Enfocado en el uso de grandes redes convolucionales, en *Billion-scale semi-supervised learning for image classification* (Yalniz et al., 2019) se propone un pipeline basado en teacher-student que es capaz de aprovechar colecciones de datos no etiquetados (hasta mil millones de imágenes), haciendo que el teacher guíe al student durante el entrenamiento proporcionándole etiquetas suaves.

Un uso de la técnica anterior se da en *Teacher-Student chain for efficient semi-supervised histology image classification* (Shaw et al., 2020), donde se aplica la técnica propuesta por (Yal-

niz et al., 2019) para la clasificación en imágenes de cáncer colorrectal. Se consigue mejorar la precisión mediante el uso de cadena de estudiantes, donde cada estudiante se usa para entrenar al siguiente estudiante, haciendo así las veces de maestro. Cabe destacar los buenos resultados, consiguiendo con tan solo un 0.5 % de datos etiquetados igualar la precisión del aprendizaje supervisado usando el 100 % de datos etiquetados.

Centrado en mejorar el aprendizaje semisupervisado en clasificación de imágenes médicas, *Reliability-Aware Contrastive Self-ensembling for Semi-supervised Medical Image Classification* (Hang et al., 2022) utiliza Mean Teacher junto con un método de autoensamblaje contrastivo (self-ensembling) para que los datos no etiquetados obtenidos de diferentes poblaciones y entornos no afecten negativamente al rendimiento en la clasificación.

Materiales

4

Con el fin de comprobar el desempeño de un modelo semisupervisado usando Mean Teacher, se ha obtenido un dataset con imágenes médicas (noa) y se han creado a partir de este varios subconjuntos de datasets con distinto número de imágenes etiquetadas y sin etiquetar.

El dataset utilizado consiste en imágenes de rayos X de tórax (anteroposteriores) seleccionadas de cohortes retrospectivas de pacientes pediátricos de uno a cinco años del Guangzhou Women and Children's Medical Center, Guangzhou. Este dataset está organizado en tres carpetas: entrenamiento (train), prueba (test) y validación (val), conteniendo subcarpetas para cada categoría de imagen (Neumonía/Normal). En las imágenes de neumonía se encuentran ejemplos de casos provocados por virus o por bacterias. No obstante, estos dos tipos de neumonía son considerados simplemente como caso positivo de neumonía.

En total, el dataset incluye 5,863 imágenes en formato JPEG en escala de grises. Dentro de este conjunto, 1341 imágenes pertenecen a la categoría Normal y las 3875 restantes están categorizadas como Pneumonia.

Las imágenes en la categoría Normal (ver figura 1) muestran pulmones claros sin áreas anormales de opacidad.

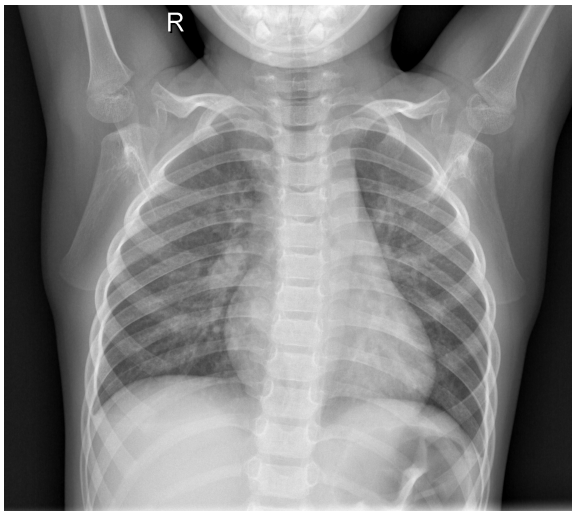

Figura 1: *Imagen de radiografía normal*

Las imágenes categorizadas como neumonía bacteriana (ver figura 2) típicamente exhiben una consolidación focal lobular, mientras que las imágenes Virales (ver figura 3) manifiestan

un patrón “intersticial” más difuso en ambos pulmones.

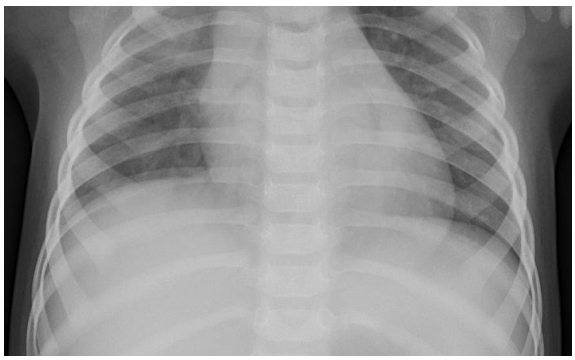

Figura 2: *Imagen de neumonía provocada por bacteria*

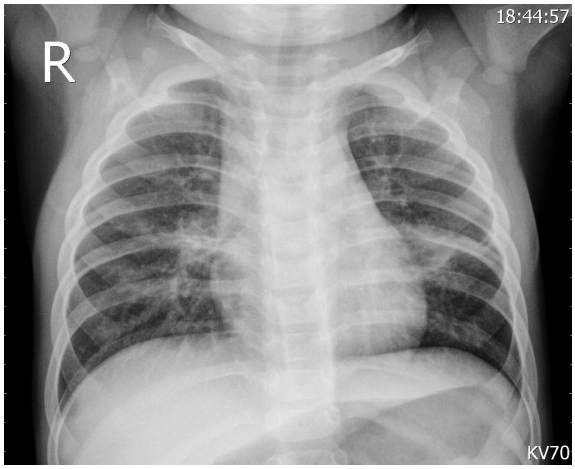

Figura 3: *Imagen de neumonía provocada por virus*

Las imágenes de rayos X se seleccionaron meticulosamente, eliminando aquellas de baja calidad o que no se podían leer correctamente y fueron evaluadas y calificadas por dos médicos expertos como parte del proceso de control de calidad antes de ser utilizadas para el entrenamiento del sistema de inteligencia artificial. Además, para compensar cualquier error en la calificación, un tercer experto revisó también el conjunto de evaluación.

Este proceso de selección y revisión asegura que solo las imágenes de la más alta calidad y relevancia clínica se incluyan en el dataset, y proporciona una base sólida y confiable para el entrenamiento y evaluación de modelos de aprendizaje automático en el diagnóstico de neumonía a partir de imágenes de rayos X de tórax.

4.1. Datasets para entrenamiento de los modelos

El presente apartado proporciona detalles sobre los conjuntos de datos empleados para los experimentos, con el objetivo de validar el modelo Mean Teacher en contextos donde hay una escasez de imágenes etiquetadas.

A fin de generar conjuntos de datos con variadas proporciones de imágenes etiquetadas y no etiquetadas, se han diseñado subconjuntos derivados del dataset original. Para crear estos subconjuntos, se seleccionaron imágenes etiquetadas de manera aleatoria, manteniendo una distribución controlada, mientras que las imágenes restantes se utilizaron como datos sin etiquetar.

El subconjunto denominado “Dataset 200” integra 75 imágenes “Normal”, 125 “Pneumonia” y 5016 sin etiquetar. En el caso del “Dataset 500”, se incluyen 187 imágenes “Normal”, 313 “Pneumonia” y 4716 sin etiquetar. Por otro lado, el “Dataset 1000” comprende 375 imágenes “Normal”, 625 “Pneumonia” y 4216 sin etiquetar. Finalmente, “Dataset 2500” cuenta con 937 imágenes “Normal”, 1563 “Pneumonia” y 2716 sin etiquetar.

La tabla 1 muestra el resumen de la composición de los distintos datasets usados para entrenar y validar los modelos.

Dataset	Normal	Pneumonia	Sin etiquetar
Dataset 200	75	125	5016
Dataset 500	187	313	4716
Dataset 1000	375	625	4216
Dataset 1500	937	1563	2716

Tabla 1: ***Tamaño de los datasets.***

Además, se han preparado conjuntos específicos para validación y prueba. El conjunto de validación consiste en 48 imágenes normales y 78 imágenes de neumonía, mientras que el conjunto de prueba comprende 234 imágenes normales y 390 de neumonía.

Es crucial mencionar que se ha mantenido una proporción de clases similar en todos los conjuntos de datos. Esta decisión metodológica se ha tomado con el propósito de facilitar y hacer válida la comparación de resultados entre los diferentes conjuntos de datos y modelos evaluados.

4.2. Verificación de los Conjuntos de Datos de Entrenamiento

Para garantizar la integridad y la validez de los procesos de entrenamiento y evaluación de los modelos de aprendizaje automático, es importante llevar a cabo una verificación meticulosa de los conjuntos de datos de entrenamiento. Dicha verificación se enfoca en asegurar que los conjuntos de datos generados para el entrenamiento estén libres de contaminación. En este contexto, un conjunto de datos “contaminado” se refiere a aquel que podría estar utilizando, erróneamente, imágenes del conjunto de pruebas (test) durante la fase de entrenamiento.

El propósito de esta precaución es fundamental para el éxito de los sistemas de aprendizaje automático. Estos sistemas están diseñados para aprender y generalizar conceptos complejos y patrones a partir de sus conjuntos de datos de entrenamiento, para luego aplicar de manera eficaz y precisa este conocimiento adquirido a datos nuevos que se les presenten posteriormente. Por ende, asegurar que los datos de prueba no se infiltren en la fase de entrenamiento

es esencial para preservar la capacidad de generalización de los modelos y garantizar que su desempeño sea auténtico y no producto de un sobreajuste a datos específicos.

El protocolo de verificación implementado para lograr este objetivo ha sido el siguiente. Inicialmente, los datos designados para entrenamiento (train) y prueba (test) son cargados y organizados en listas separadas. Posteriormente, se realiza una evaluación para identificar la intersección entre estas dos listas, es decir, se busca cualquier elemento que podría estar presente en ambas simultáneamente. Esta intersección debe resultar en un conjunto vacío, lo que confirmaría que no hay imágenes compartidas o duplicadas entre los conjuntos de datos de entrenamiento y prueba. Esta verificación es aplicada a cada uno de los conjuntos de datos creados, los cuales varían en tamaño y están compuestos por 200, 500, 1000 y 2500 imágenes respectivamente.

Una vez que se concluye satisfactoriamente esta fase de verificación y se confirma que los conjuntos de datos han sido configurados y segmentados correctamente, se da inicio al siguiente paso del proceso. En esta etapa subsiguiente, se procede a la construcción y definición de las arquitecturas de los modelos de aprendizaje automático que serán entrenados con los datos previamente validados. Este procedimiento asegura que la base sobre la cual se entrenan los modelos es sólida y confiable, estableciendo una buena base para el desarrollo y evaluación de sistemas de aprendizaje automático eficaces y precisos.

Metodología 5

5.1. Entorno y librerías

Para el desarrollo y ejecución de este Trabajo Fin de Máster, se ha optado por utilizar **Google Colab**, una plataforma de desarrollo que ofrece la capacidad de escribir y ejecutar código en la nube de forma gratuita. Google Colab proporciona un entorno interactivo basado en **Jupyter Notebooks**, lo que resulta muy conveniente para proyectos que involucran análisis de datos, aprendizaje automático y otras aplicaciones de investigación.

La elección de Google Colab se basa en sus ventajas inherentes, como el acceso gratuito a GPUs, la facilidad de compartir el trabajo y la integración con Google Drive para alojar el dataset de imágenes necesarias para el entrenamiento y evaluación de los modelos. Estas características facilitan la implementación y prueba de modelos de aprendizaje profundo de manera eficiente.

En cuanto a las librerías, se ha utilizado principalmente **PyTorch**, una de las bibliotecas de aprendizaje profundo más populares y potentes disponibles en la actualidad. PyTorch no solo ofrece una amplia variedad de herramientas y funciones para entrenar y evaluar modelos de redes neuronales, sino que también es conocido por su flexibilidad y eficiencia, lo que lo hace adecuado para proyectos de investigación en inteligencia artificial.

5.2. Preprocesamiento de los datos

Se realiza un preprocesamiento de las imágenes antes de alimentar a los modelos con ellas. Este preprocesamiento es aplicado de diferente forma según se trate del conjunto de imágenes de entrenamiento o de validación y test.

5.2.1. Datos de entrenamiento

Para las imágenes de entrenamiento se aplican una serie de transformaciones para ayudar a que los modelos aprendan a generalizar mejor, aplicando pequeñas modificaciones en forma de recortado, rotaciones, y normalizaciones. Estas transformaciones se hacen para cada imagen usada en el entrenamiento. Los rangos de aplicación hacen que no se alteren siempre todos los parámetros de la imagen, ya de forma aleatoria puede que una característica permanezca invariable. Esta idea de implementación se basa en lo realizado en la propuesta de

implementación en Yu et al. (2021).

Se han aplicado concretamente las siguientes transformaciones:

- Redimensionado de las imágenes a 224x224 píxeles: esto es debido a que se realiza un *fine tuning* sobre una arquitectura ResNet50 ya entrenada, y esta requiere que las imágenes empleadas tengan estas dimensiones.
- Recorte de imágenes: se aplica un recorte aleatorio con un rango de 0.8 a 1 que posteriormente se redimensiona a 224x224 para mantener el tamaño requerido. Esto hace que los modelos no sobre ajusten y sean más robustos ante variaciones en la posición y escala de las distintas imágenes.
- Cambios aleatorios en el brillo: de forma aleatoria se cambia el brillo de la imagen, pudiendo variar entre el 80 % y el 120 % del brillo original.
- Variación en el contraste de forma aleatoria entre un 80 % y un 120 % del contraste original.
- Variación de la saturación: de igual forma se modifica la saturación de la imagen entre un 80 % y un 120 % de los valores originales.
- Modificación del tono: se aplican modificaciones en el tono de la imagen en un rango del 90 % al 110 % del tono original.
- Normalización: se normalizan los valores de los pixeles con media y desviación estándar adecuada para hacer *fine tuning* con la ResNet50 previamente entrenada en ImageNet.

5.2.2. Datos de validación y test

Para el caso de las imágenes de validación y pruebas, tan solo se realizan las transformaciones necesarias para poder usarlas en nuestro modelo, sin aplicar ninguna modificación adicional. Estas son:

- Redimensionado a 224x224 píxeles: esto es debido a que se realiza un *fine tuning* sobre una arquitectura ResNet50 ya entrenada, y esta requiere que las imágenes empleadas tengan estas dimensiones.
- Normalización: se normalizan los valores de los pixeles con media y desviación estándar adecuada para hacer *fine tuning* con la ResNet50 previamente entrenada en ImageNet.

5.3. Concepto tras el Mean Teacher

El concepto central del modelo Mean Teacher es una estructura que comprende dos componentes: el modelo student y el modelo teacher.

El **modelo student** representa la red neural principal y actúa como la pieza fundamental del proceso de aprendizaje. Este modelo se somete a un proceso de entrenamiento intensivo y

dinámico mediante la técnica de retropropagación. Durante este proceso, el modelo student se actualiza y refina continuamente, haciendo uso tanto de datos etiquetados como de aquellos que no poseen etiquetas. En el caso de los datos no etiquetados, el modelo student no opera de manera aislada; en cambio, toma las predicciones generadas por el modelo teacher, utilizando estas como objetivos pseudo-etiquetados. Esta interacción entre los dos modelos facilita un aprendizaje más profundo y controlado del modelo student.

Por otro lado, el **modelo teacher** desempeña otro papel. A diferencia del modelo student, el teacher no se entrena directamente mediante retropropagación. Su función principal es proporcionar una guía y referencia para el modelo student. Los pesos del modelo teacher se actualizan mediante un promedio móvil de los pesos del student, realizado al final de cada época de entrenamiento. Este método de actualización proporciona predicciones suavizadas y estabilizadas que sirven como objetivos precisos para el student en el caso de datos no etiquetados.

Un aspecto fundamental del modelo Mean Teacher es el énfasis en la consistencia. El modelo está diseñado para asegurar coherencia entre las predicciones del modelo student y del modelo teacher, especialmente cuando se trata de datos no etiquetados. Para mantener y fortalecer esta consistencia, se implementa una función de pérdida de consistencia con el propósito de minimizar cualquier discrepancia entre las predicciones de ambos modelos.

En cuanto a la **actualización de los pesos**, después de cada iteración, los pesos del modelo teacher se recalibran tomando un promedio móvil exponencial (EMA) de los pesos del modelo student. Esta técnica de actualización garantiza una transición suave y gradual en los cambios del modelo teacher, proporcionando así objetivos estables y confiables para el modelo student.

El objetivo final del Mean Teacher va más allá de simplemente minimizar la pérdida en los datos etiquetados. Busca también asegurar una correspondencia y coherencia entre las predicciones del modelo student y del modelo teacher en el contexto de datos no etiquetados. La arquitectura del Mean Teacher está diseñada para maximizar la utilidad de los datos no etiquetados. Se va alineando a las predicciones del modelo student con las de un modelo teacher más estable y confiable.

Este enfoque coordinado resulta en una mejora significativa en la capacidad de generalización del sistema, optimizando su rendimiento especialmente en situaciones donde los datos etiquetados son limitados o escasos.

5.4. Arquitectura del Mean Teacher

Como se ha descrito anteriormente, el modelo Mean Teacher consta de dos redes neuronales: student y teacher.

Ambas comparten la misma estructura. En este esquema, la red student aprende de los datos etiquetados, mientras que la teacher funciona como una versión promediada del student con el paso del tiempo.

En la figura 4 se muestra el esquema de una arquitectura Mean Teacher usando EMA, tal como se publicó en el paper (Tarvainen y Valpola, 2017).

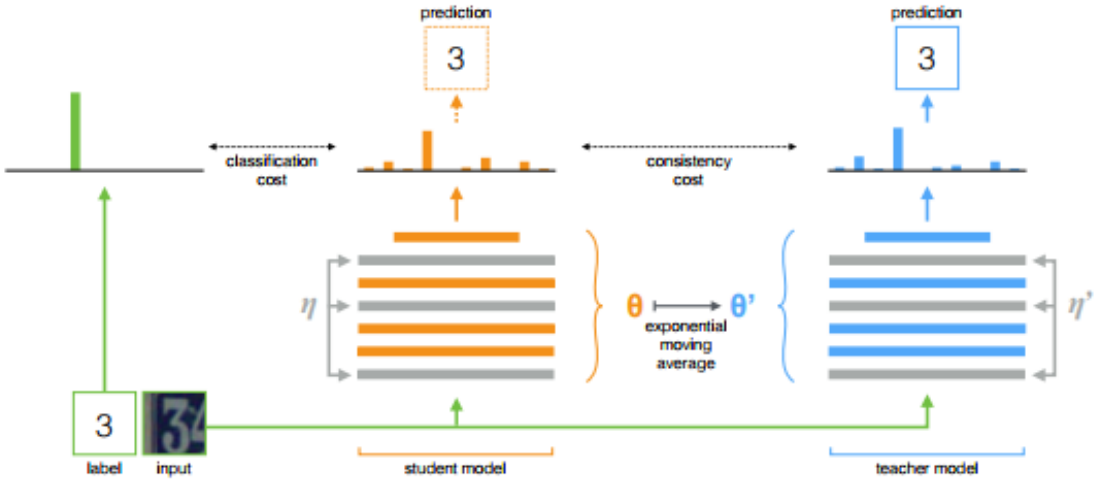

Figura 4: *Arquitectura Mean Teacher*

Para la construcción del modelo se ha utilizado una arquitectura **ResNet50** preentrenada en ImageNet. Se trata de una red neuronal del tipo CNN muy utilizada en tareas de clasificación de imágenes. Esta arquitectura se ha modificado para adaptarla a la necesidad específica de este trabajo: clasificar imágenes en dos clases, a saber, neumonía y normal.

Los cambios realizados a la ResNet50 original son los siguientes:

- Se eliminó la última capa completamente conectada (FC) de la ResNet50 original.
- Se añadió una capa de Global Average Pooling (GAP) para reducir la dimensionalidad espacial.
- Se introdujo una capa densa, compuesta por un aplanamiento, una capa lineal de 2048 a 512 neuronas, una función de activación ReLU y un dropout con una tasa del 50 % para evitar el sobreajuste.
- Finalmente, se añadió una capa clasificadora lineal que tiene 512 entradas y 2 salidas, correspondientes a las clases “neumonía” y “normal”.

La arquitectura resultante se define mediante la clase `CustomResNet50`. Esta clase define la estructura del modelo, así como su comportamiento de propagación hacia adelante (forward).

Se empieza cargando el ResNet50 preentrenado, una arquitectura profunda de 50 capas ampliamente usada para tareas de clasificación. Esto se debe a su habilidad para manejar el desvanecimiento del gradiente mediante conexiones residuales. Al usar el modelo con pesos preentrenados en ImageNet, se facilita un entrenamiento más ágil y una convergencia más efectiva.

El modelo toma todas las capas de ResNet50, excepto la última. Estas capas funcionan como extractores de características, convirtiendo la imagen de entrada en un conjunto de características de alto nivel. Luego, la salida de estas capas de características pasa por una

operación de Global Average Pooling (GAP). Esta operación reduce las dimensiones espaciales de la salida a 1x1 mientras mantiene la profundidad, lo que es útil para reducir el número de parámetros y prevenir el sobreajuste.

La salida del GAP es aplanada y dirigida a través de una capa densa. Aquí, la salida se aplana, se aplica un dropout de 0.2 para evitar el sobreajuste, y se introduce una capa lineal que reduce las dimensiones de 2048 a 512. Luego, se aplica una función de activación ReLU, seguida por otra capa de dropout con un valor de 0.2 antes del clasificador.

Finalmente, la salida pasa a través de un clasificador con dos neuronas, correspondientes a las dos clases de interés. La salida de este clasificador puede ser procesada por una función softmax para asignar probabilidades a cada clase. Todo el proceso desde la entrada hasta la salida está bien definido, con la entrada moviéndose sucesivamente a través de las capas de características, el GAP, la capa densa y finalmente el clasificador para generar la salida.

5.5. Entrenamiento

Antes de iniciar el proceso de entrenamiento del modelo Mean Teacher, es necesario realizar varios ajustes y tener en cuenta ciertos aspectos cruciales debido a las características únicas del método empleado.

En la fase inicial, se crean dos instancias del modelo CustomResNet50. Una de estas instancias se asigna al modelo student y la otra al modelo teacher. A continuación, todas las capas del modelo student se congelan. Esto se logra estableciendo el parámetro “requires_grad” en False, lo que significa que, en las primeras etapas del entrenamiento, los pesos de estas capas permanecerán inalterados.

Después de congelar las capas del student, se procede a descongelar específicamente las capas densas y la capa de clasificación que fueron añadidas a la CustomResNet50. Al hacerlo, estas capas se vuelven entrenables. Los parámetros de estas capas descongeladas se añaden a una lista denominada “params_to_optimize”, que será esencial en etapas posteriores del proceso para la optimización.

A continuación, se inicia el modelo teacher con los pesos del modelo student. Para lograr esto, se utiliza la función “load_state_dict” para copiar los pesos del student al teacher, garantizando que ambos modelos inicien el proceso de entrenamiento con pesos idénticos. Después de esta inicialización, se desactiva el cálculo del gradiente para el modelo teacher. Esto se realiza estableciendo “requires_grad” en False para todos los parámetros del teacher, asegurando que no se realicen actualizaciones directas a este modelo durante el entrenamiento.

Finalmente, se verifica la disponibilidad de una GPU. En caso de que esté disponible, se configura el dispositivo para utilizarla; de lo contrario, el proceso de entrenamiento se llevará a cabo en la CPU. Una vez decidido el dispositivo que se utilizará, tanto el modelo student como el teacher se trasladan a dicho dispositivo (ya sea GPU o CPU) utilizando el método “.to(device)”. Esto asegura que ambos modelos estén en el lugar correcto para comenzar el entrenamiento.

5.6. Hiperparámetros

Para el adecuado entrenamiento del modelo Mean Teacher, es necesario ajustar correctamente diversos hiperparámetros, ya que cada uno de estos influye directamente en la eficacia y optimización del rendimiento del modelo.

Primero, consideremos la **Tasa de Aprendizaje o Learning Rate**, un hiperparámetro crucial que representa el tamaño de los pasos que el optimizador dará en cada iteración para actualizar los pesos del modelo. Se requiere un equilibrio en su selección: una tasa muy alta podría provocar oscilaciones y falta de convergencia, mientras que una tasa demasiado baja podría alentar excesivamente el proceso de entrenamiento. Después de realizar pruebas con diferentes valores, se determinó que una tasa de aprendizaje de 0.0001 es la más apropiada para este proyecto.

En cuanto al **optimizador**, este hiperparámetro es responsable de actualizar los pesos del modelo basándose en el gradiente de la pérdida. Para este proyecto, se ha seleccionado el optimizador Adam. Este optimizador es preferido porque integra las fortalezas de otros dos algoritmos de optimización importantes: AdaGrad y RMSProp.

La **Función de Pérdida**, por otro lado, es la métrica que el optimizador busca minimizar durante el proceso de entrenamiento. En este proyecto, se ha elegido utilizar CrossEntropyLoss, una función de pérdida que es especialmente efectiva para tareas de clasificación.

El **Tamaño del Lote**, o **Batch Size**, es otro hiperparámetro significativo. Este representa la cantidad de ejemplos utilizados en cada iteración para calcular el gradiente y actualizar los pesos. Un tamaño de lote más grande puede acelerar el entrenamiento, aunque requiere más memoria. Para este proyecto, se ha decidido usar un tamaño de lote de 64.

El **Número de Epochs** determina la cantidad de veces que el modelo se expondrá al conjunto de datos completo durante el entrenamiento. Se debe evitar un número excesivo de epochs para prevenir el sobreajuste y un número demasiado pequeño para evitar el subajuste. En este proyecto, se ha establecido el número de epochs en 400. Aun así, se ha implementado una técnica de Early Stopping con una paciencia de 40 epochs. Esto significa que el entrenamiento se detendrá si no se observa mejora en el accuracy en el conjunto de validación después de 40 epochs.

La **Pérdida de Consistencia o Consistency Loss** se calcula entre las predicciones del modelo student y el teacher para los datos no etiquetados. Se utiliza una temperatura de 1.2 para suavizar las probabilidades, y al ser este valor mayor que 1, resulta en que la pérdida de consistencia sea menos sensible a diferencias pequeñas entre las predicciones del student y las del teacher.

$$\text{Consistency Loss} = \text{mean}\left(\left(\frac{\text{softmax}(\text{preds}_a)}{\text{temperature}} - \frac{\text{softmax}(\text{preds}_b)}{\text{temperature}}\right)^2\right)$$

Se utiliza también una **Ponderación Dinámica**. Este método asigna un peso variable a la pérdida de consistencia en la pérdida total, con el peso calculado mediante una función específica.

$$\text{loss} = \text{supervised\_loss} + \text{weight} \times \text{consistency\_loss\_value}$$

El peso se calcula con la función:

$$w(t) = \frac{1}{1 + e^{-k(t-t_0)}}$$

Donde:

- $w(t)$ es el peso en el tiempo/época t.
- k es una constante que controla la pendiente de la función sigmoide.
- t_0 es el punto de inflexión de la sigmoide, es decir, el punto en el que la función comienza a aumentar rápidamente.

Finalmente, se utiliza un coeficiente **Exponential Moving Average (EMA)** para actualizar los pesos del modelo teacher con los pesos del student al final de cada epoch. Este coeficiente depende de los hiperparámetros decay y constant. El parámetro decay pondera la importancia entre el valor promedio anterior y el nuevo valor observado, y el valor de decay utilizado es 0.9. Por otro lado, el parámetro constant controla el suavizado de la EMA a lo largo de las epochs, con un valor fijado en 0.7.

El coeficiente EMA se define como:

$$ema\_coefficient = \frac{(1 - \text{decay}) \times (1 - \text{constant}^{\text{epoch}+1})}{(1 - \text{constant})}$$

Con este coeficiente, se actualizan los parámetros del teacher siguiendo la ecuación:

$$\begin{aligned} teacher\_params\_data = {} & (1 - \text{ema\_coefficient}) \times \text{student\_params\_data} \\ & + \text{ema\_coefficient} \times \text{teacher\_params\_data} \end{aligned}$$

- **Parámetro decay**: Este parámetro determina la velocidad a la que el modelo actualiza sus pesos en función de los nuevos valores. Su función radica en ponderar la importancia entre el valor promedio anterior y el nuevo valor observado. En caso de la implementación de Mean Teacher, el EMA se usa para actualizar los pesos del modelo del teacher basándose en los del student, y el objetivo es que el modelo del teacher tenga una versión suavizada y estable de los pesos del modelo student.

  Valores cercanos a 1 hacen que el teacher cambie de forma muy lenta y de más importancia a sus pesos anteriores. Por contra, valores cercanos a 0 hacen que el modelo cambie más rápidamente, dando más importancia a los valores recientes del student.

  El valor de decay que se usa en la implementación de este modelo es 0.9. Esto significa que el modelo del teacher da una mayor importancia a sus pesos anteriores y se actualiza de una forma más gradual con respecto a los cambios en el student.

- **Parámetro constant**: se trata de un parámetro que controla el suavizado de la EMA a lo largo de las épocas, de forma que adapta dinámicamente el coeficiente EMA según va progresando el entrenamiento. **El valor que se fija para constant es de 0.7.**

En cuanto a la **regularización**, y como se mencionó anteriormente, se utiliza la técnica de dropout para prevenir el sobreajuste. Este método aleatoriamente desactiva un conjunto de neuronas durante el entrenamiento, lo que puede ayudar a mejorar la generalización del modelo, con un **dropout de 0.2**, basado en estudios previos y experimentación propia.

5.7. Baseline con modelo supervisado

La implementación del modelo baseline supervisado tiene como principal objetivo proporcionar un punto de referencia claro para evaluar el desempeño del modelo Mean Teacher, especialmente en cuanto a su capacidad de aprovechar los datos no supervisados para mejorar su rendimiento. Este modelo baseline supervisado se configura y entrena exclusivamente con datos etiquetados derivados de los conjuntos de datos previamente creados.

En cuanto a la arquitectura seleccionada para el modelo baseline, esta se mantiene idéntica a la empleada tanto para el modelo teacher como para el student dentro del marco del Mean Teacher, utilizando también la CustomResNet50. Esta elección intencional de arquitectura idéntica busca establecer un terreno de comparación homogéneo y válido, aplicando técnicas de preprocesamiento de imagen y métricas de validación similares en todos los casos.

El proceso de entrenamiento y optimización del modelo baseline supervisado sigue una metodología consciente y detallada. Este modelo se somete a un proceso de entrenamiento utilizando el optimizador Adam, estableciendo una tasa de aprendizaje de 0.0001, en concordancia con los parámetros seleccionados para el modelo Mean Teacher. Este enfoque asegura que cualquier diferencia de rendimiento observable entre los modelos pueda atribuirse con confianza a la aplicación del enfoque semisupervisado, evitando interferencias por variaciones en parámetros o técnicas de entrenamiento.

En la fase de evaluación y medición de métricas, el modelo baseline se somete a pruebas con los mismos conjuntos de validación y test que se utilizan para el modelo Mean Teacher. Se emplean varias métricas cruciales para facilitar una comparación comprensiva y reveladora entre los modelos. Estas métricas incluyen precisión (accuracy), sensibilidad (recall), puntuación F1 (F1-score), área bajo la curva ROC (AUC-ROC) y el conteo de falsos negativos.

5.8. Métricas

La **precisión (accuracy)** sirve como un indicador inicial del rendimiento general, mostrando el porcentaje de predicciones correctas del modelo en relación con el total de predicciones realizadas. Aunque útil, esta métrica puede no ser suficiente para evaluaciones detalladas, especialmente en situaciones donde las clases están desequilibradas.

La **sensibilidad (recall)** es una métrica vital, destacando la proporción de positivos reales correctamente identificados por el modelo. Es calculada dividiendo el número de verdaderos positivos entre la suma de verdaderos positivos y falsos negativos. Esta métrica es esencial en contextos donde los falsos negativos pueden tener implicaciones significativas, como en la detección de enfermedades.

La **puntuación F1 (F1-score)** proporciona una medida que combina la precisión y la sensibilidad en un solo valor, calculado como el promedio armónico de ambas métricas. Este indicador es particularmente valioso cuando se busca equilibrar precisión y sensibilidad, especialmente en datasets con clases desequilibradas.

El **área bajo la curva ROC (AUC-ROC)** y la propia curva ROC son instrumentos analíticos cruciales. La curva ROC ilustra la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos, mientras que el AUC-ROC cuantifica el área bajo esta curva. Un AUC-ROC de 1 simboliza una predicción perfecta, y un valor de 0.5 indica un rendimiento equivalente al azar, siendo esta métrica vital para evaluar la capacidad distintiva del modelo entre clases.

Finalmente, el conteo de **Falsos Negativos** se enfoca en identificar las observaciones positivas que fueron incorrectamente clasificadas como negativas, siendo una métrica de suma importancia en contextos donde los falsos negativos pueden tener consecuencias severas, dado que señalan casos no detectados o pasados por alto.

Resultados y Discusión

6

Se presentan los resultados de los distintos experimentos usando cantidades variables de datos anotados y contrastando los resultados con el modelo baseline.

6.1. Resultados usando 200 imágenes etiquetadas

En este segmento se exponen los resultados derivados del uso de 200 imágenes etiquetadas, destacando los hallazgos obtenidos con el modelo supervisado y el modelo Mean Teacher, evaluando ambos a través de distintas métricas críticas.

El modelo supervisado alcanza una precisión del 81.09 %, logrando clasificar correctamente esta proporción de instancias. Mientras tanto, el modelo Mean Teacher muestra una clara superioridad en este aspecto, con una precisión del 86.38 %. Este incremento, que se traduce en aproximadamente 5.3 puntos porcentuales, pone de manifiesto que el modelo Mean Teacher es más eficaz en términos de precisión.

En cuanto a la sensibilidad, que se refiere a la proporción de casos positivos reales identificados correctamente, el modelo supervisado reconoce el 77 % de estos. En cambio, el modelo Mean Teacher demuestra una capacidad superior, identificando el 84 % de los casos positivos reales, lo que se traduce en una mejora de 7 puntos porcentuales en comparación con el modelo supervisado.

El valor F1 del modelo supervisado es de 0.79, indicando un balance aceptable entre precisión y sensibilidad. Sin embargo, el modelo Mean Teacher muestra un mejor desempeño en este ámbito también, registrando un valor F1 de 0.85 y ofreciendo así un equilibrio entre estas métricas.

En relación con el área bajo la curva ROC, el modelo supervisado exhibe un rendimiento de clasificación notable, con un valor de 0.89. Pero una vez más, el modelo Mean Teacher se destaca con un valor superior, alcanzando un 0.92 en el área bajo la curva ROC y demostrando, por ende, una discriminación entre clases más precisa.

En lo que respecta a los falsos negativos, el modelo supervisado y el Mean Teacher generan 31 y 30 respectivamente, mostrando un comportamiento bastante similar en esta métrica.

Concluyendo, es evidente que, en la mayoría de las métricas evaluadas, el modelo Mean Teacher presenta un rendimiento superior al modelo supervisado. Esta superioridad sugiere que la adición de datos no etiquetados mediante la estrategia Mean Teacher mejora notablemente la capacidad del modelo en la tarea de clasificación en cuestión.

En la tabla 2 se presenta a modo de resumen los resultados de las distintas métricas evaluadas en el dataset 200.

Métrica	**Supervisado**	**Mean Teacher**
Accuracy	81.09	86.38
Recall	0.77	0.84
F1	0.79	0.85
ROC - AUC	0.89	0.92
False Negatives	31	30

Tabla 2: ***Resumen de métricas de los modelos en dataset 200.***

6.2. Resultados usando 500 imágenes etiquetadas

Los resultados obtenidos con 500 imágenes etiquetadas revelan notables diferencias en el desempeño entre el modelo supervisado y el modelo Mean Teacher, siendo este último superior en diversas métricas.

En términos de precisión, el modelo supervisado logra clasificar de manera acertada el 83.65 % de las instancias, mientras que el modelo Mean Teacher mejora esta métrica al alcanzar un 88.62 % de precisión. Esta diferencia, equivalente a 5 puntos porcentuales, evidencia la mayor eficacia del modelo Mean Teacher.

Con respecto a la sensibilidad, el modelo supervisado identifica correctamente el 81 % de los casos positivos reales. Por otro lado, el Mean Teacher supera esta métrica con un 87 % de casos positivos reales identificados correctamente, lo que representa una ventaja de 6 puntos porcentuales frente al modelo supervisado.

El valor F1 de ambos modelos también muestra diferencias significativas. Mientras que el modelo supervisado obtiene un valor de 0.82, reflejando un balance adecuado entre precisión y sensibilidad, el modelo Mean Teacher alcanza un valor de 0.88, indicando un equilibrio óptimo entre estas dos métricas.

En lo concerniente al área bajo la curva ROC, el modelo supervisado muestra un buen desempeño con un valor de 0.92. Sin embargo, el modelo Mean Teacher sobresale con un valor de 0.95, denotando una capacidad superior en la discriminación entre clases.

En cuanto a los falsos negativos, el modelo supervisado produce 29, mientras que el modelo Mean Teacher reduce este número a 25, cometiendo menos errores de este tipo en comparación con el modelo supervisado.

La conclusión general de estos hallazgos resalta que, en todas las métricas analizadas, el modelo Mean Teacher demuestra un rendimiento superior al del modelo supervisado. Esta superioridad en el desempeño sugiere que la integración de datos no etiquetados mediante la metodología Mean Teacher potencia de manera significativa la eficacia del modelo en la tarea de clasificación abordada.

En la tabla 3 se muestran los resultados de las distintas métricas evaluadas en el dataset 500.

Métrica	**Supervisado**	**Mean Teacher**
Accuracy	83.65	88.62
Recall	0.81	0.87
F1	0.82	0.88
ROC - AUC	0.92	0.95
False Negatives	29	25

Tabla 3: ***Resumen de métricas de los modelos en dataset 500.***

6.3. Resultados usando 1000 imágenes etiquetadas

Con el análisis de 1000 imágenes etiquetadas, se observan diferencias en los resultados obtenidos por el modelo supervisado y el modelo Mean Teacher. En este conjunto de datos, el modelo Mean Teacher exhibe una eficacia superior en todas las métricas evaluadas.

El modelo supervisado logra una precisión del 83.97 %, clasificando correctamente esta proporción de instancias. En cambio, el modelo Mean Teacher muestra una precisión del 88.14 %, superando al modelo supervisado por un margen aproximado de 4.17 puntos porcentuales.

En cuanto a la sensibilidad, el modelo supervisado identifica correctamente el 81 % de los casos positivos reales. De manera contrastante, el modelo Mean Teacher incrementa esta proporción al 86 %, exhibiendo una mejora de 5 puntos porcentuales en comparación con el modelo supervisado.

Al analizar el valor F1, el modelo supervisado obtiene un 0.82, reflejando un adecuado equilibrio entre precisión y sensibilidad. Por su parte, el modelo Mean Teacher alcanza un valor F1 de 0.87, lo cual indica un equilibrio más destacado entre estas métricas.

En relación al área bajo la curva ROC, el modelo supervisado alcanza un valor notable de 0.93, señalando un rendimiento muy positivo en las tareas de clasificación. Sin embargo, el modelo Mean Teacher logra superar ligeramente este rendimiento con un área bajo la curva de 0.95, lo cual demuestra su ligera superioridad en la discriminación entre clases respecto al modelo supervisado.

En lo que respecta a los falsos negativos, el modelo supervisado genera 25, mientras que el modelo Mean Teacher reduce esta cifra a 22, cometiendo tres errores menos de este tipo.

En conclusión, el modelo Mean Teacher supera al modelo supervisado en todas las métricas analizadas en este conjunto de 1000 imágenes etiquetadas. Este rendimiento superior sugiere que el uso estratégico de datos no etiquetados, mediante el enfoque del modelo Mean Teacher, potencia significativamente las capacidades de clasificación del modelo, optimizando su rendimiento en esta tarea específica.

En la tabla 4 se muestran los resultados de las distintas métricas evaluadas en el dataset 1000.

Métrica	**Supervisado**	**Mean Teacher**
Accuracy	83.97	88.14
Recall	0.81	0.86
F1	0.82	0.87
ROC - AUC	0.93	0.95
False Negatives	25	22

Tabla 4: ***Resumen de métricas de los modelos en dataset 1000.***

6.4. Resultados usando 2500 imágenes etiquetadas

Al analizar los resultados obtenidos con 2500 imágenes etiquetadas, se observa un comportamiento distinto en los modelos Supervisado y Mean Teacher respecto a los conjuntos previos. En este conjunto de datos más extenso, el modelo supervisado arroja un rendimiento superior en la mayoría de las métricas evaluadas, con una precisión del 87.18 %, mientras que el modelo Mean Teacher alcanza una precisión del 84.78 %. Esta diferencia refleja una ventaja de aproximadamente 2.4 puntos porcentuales para el modelo supervisado en términos de precisión.

En cuanto a la sensibilidad, el modelo supervisado identifica correctamente el 85 % de los casos positivos reales, superando en 3 puntos porcentuales al modelo Mean Teacher, que identifica correctamente el 82 % de dichos casos. Esta métrica refuerza la tendencia observada en la precisión, con el modelo supervisado mostrando una capacidad ligeramente superior para identificar los casos positivos correctamente.

Al considerar el valor F1, que es un indicador del equilibrio entre precisión y sensibilidad, el modelo supervisado también supera al modelo Mean Teacher con un valor de 0.86 contra 0.83, respectivamente. Esta diferencia en el valor F1, aunque pequeña, indica un mejor equilibrio en el rendimiento del modelo supervisado.

Con respecto al área bajo la curva ROC, una métrica crucial para evaluar la efectividad en la clasificación, el modelo supervisado logra un rendimiento notable de 0.94, mientras que el modelo Mean Teacher alcanza un valor de 0.92. Aunque ambos valores son altos, reflejando una capacidad de discriminación efectiva entre clases para ambos modelos, el modelo supervisado muestra un rendimiento ligeramente superior.

Sin embargo, es crucial notar que, a pesar de la superioridad del modelo supervisado en las métricas mencionadas, el modelo Mean Teacher registra un menor número de falsos negativos, con 22 frente a los 26 del modelo supervisado. Dado que los falsos negativos pueden tener implicaciones significativas en distintas aplicaciones, esta reducción en el error podría ser un factor determinante en la selección del modelo para tareas específicas.

En conclusión, aunque el modelo supervisado muestra un rendimiento global superior en este conjunto de 2500 imágenes etiquetadas, el modelo Mean Teacher presenta una ventaja significativa en la reducción de falsos negativos. Esta característica podría hacer que el modelo Mean Teacher sea una opción preferible en situaciones donde minimizar los falsos negativos sea crucial, a pesar de sus desventajas ligeras en otras métricas evaluadas.

En la tabla 5 se presenta a modo de resumen los resultados de las distintas métricas evaluadas en el dataset 2500.

Métrica	**Supervisado**	**Mean Teacher**
Accuracy	87.18	84.78
Recall	0.85	0.82
F1	0.86	0.83
ROC - AUC	0.94	0.92
False Negatives	26	22

Tabla 5: ***Resumen de métricas de los modelos en dataset 2500.***

6.4.1. Resumen de los resultados

Se presenta a modo de resumen la tabla 6 con el *accuracy* obtenido por ambos modelos, supervisado y mean teacher, evaluados en los distintos datasets de prueba. En ella se puede apreciar la mejoría obtenida al usar el enfoque semisupervisado en los datasets con menos imágenes etiquetadas, y cómo el rendimiento empeora cuando se disponen de más datos etiquetados al evaluarlo en el dataset con 2500 imágenes etiquetadas.

Accuracy	**Dataset 200**	**Dataset 500**	**Dataset 1000**	**Dataset 2500**
Supervisado	81.09	83.65	83.97	87.18
Mean Teacher	86.38	88.62	88.14	84.78

Tabla 6: ***Tabla resumen de comparación de los modelos***

6.5. Discusión

Durante el desarrollo de este trabajo, se han ejecutado una serie de experimentos utilizando diversos conjuntos de datos que contienen diferentes cantidades de imágenes etiquetadas. El propósito principal de estos experimentos fue evaluar y comparar el rendimiento entre un modelo supervisado tradicional y el modelo Mean Teacher, que incorpora datos no etiquetados durante el proceso de entrenamiento.

Al analizar los resultados con un conjunto de 200 imágenes etiquetadas, se destaca que el modelo Mean Teacher tiene un rendimiento superior al modelo supervisado en todas las métricas evaluadas, evidenciando una precisión de aproximadamente 5.3 puntos porcentuales

más alta, una sensibilidad superior en 7 puntos porcentuales y un valor F1 y área bajo la curva ROC igualmente superiores. A pesar de presentar un número similar de falsos negativos, el modelo Mean Teacher logra un falso negativo menos.

En el siguiente escenario, con 500 imágenes etiquetadas, el modelo Mean Teacher persiste en mostrar un rendimiento superior, manteniendo una diferencia en precisión de aproximadamente 5 puntos porcentuales y una sensibilidad 6 puntos porcentuales más alta, además de registrar cuatro falsos negativos menos que el modelo supervisado.

Al incrementar la cantidad de imágenes etiquetadas a 1000, el Mean Teacher sigue la tendencia de superar al modelo supervisado en todas las métricas, aunque con una diferencia en precisión reducida a 4.17 puntos porcentuales. A pesar de esto, sigue teniendo una ventaja en términos de sensibilidad y registra tres falsos negativos menos.

Con un conjunto más amplio de 2500 imágenes etiquetadas, los resultados cambian. En general, un modelo supervisado tiende a mostrar un rendimiento superior al de un modelo semi-supervisado cuando se cuenta con una amplia cantidad de datos etiquetados. En este estudio, con 2500 imágenes etiquetadas, el modelo supervisado supera al Mean Teacher en la mayoría de las métricas evaluadas.

No obstante, es fundamental resaltar el valor inherente de los modelos semi-supervisados, en especial en contextos donde la anotación de datos implica un esfuerzo significativo. Un modelo como Mean Teacher demuestra su potencial al necesitar un esfuerzo de anotación considerablemente menor para alcanzar un rendimiento similar al modelo supervisado. En este estudio, el modelo semi-supervisado podría brindar resultados comparables usando solo 500 imágenes etiquetadas, a diferencia de las 2500 requeridas para el supervisado. Esta consideración es esencial en campos como la radiología, donde el etiquetado de imágenes demanda tiempo y especialización. Por ende, técnicas semi-supervisadas emergen como alternativas valiosas y eficientes, optimizando recursos valiosos sin sacrificar significativamente la calidad del rendimiento del modelo.

Los resultados sugieren que el Mean Teacher se beneficia significativamente de la incorporación de datos no etiquetados, particularmente cuando los conjuntos de datos son más pequeños. Sin embargo, conforme aumenta el número de imágenes etiquetadas, la ventaja del Mean Teacher parece disminuir. Es plausible que, con conjuntos de datos más grandes, el modelo supervisado pueda aprender adecuadamente sin necesidad de datos no etiquetados.

Aunque el modelo supervisado superó al Mean Teacher con el conjunto de datos más grande, el Mean Teacher aún registró menos falsos negativos. Dependiendo de la aplicación específica, minimizar los falsos negativos podría ser más crucial que otras métricas, lo que podría hacer del Mean Teacher una opción preferible en ciertos contextos.

En resumen, estos resultados subrayan la necesidad de considerar tanto el tamaño del conjunto de datos como las métricas específicas de interés al seleccionar un enfoque de modelado. Mientras que el Mean Teacher ofrece ventajas claras con conjuntos de datos más pequeños, estos beneficios pueden no ser tan pronunciados conforme aumenta la cantidad de datos etiquetados.

Conclusiones

7

Este trabajo ha proporcionado una evaluación exhaustiva del rendimiento de dos enfoques distintos de modelado: el supervisado tradicional y el semi-supervisado Mean Teacher, este último incorporando datos no etiquetados durante el entrenamiento. A través de los experimentos realizados se pueden analizar las ventajas y limitaciones de cada enfoque.

1. **Superioridad del Modelo Mean Teacher en Conjuntos de Datos Pequeños:** El modelo Mean Teacher mostró una superioridad notable en la precisión, sensibilidad, valor F1 y área bajo la curva ROC en comparación con el modelo supervisado, especialmente en conjuntos de datos con un número limitado de imágenes etiquetadas. Esta superioridad indica el potencial de los modelos semi-supervisados en escenarios donde los datos etiquetados son limitados o costosos de obtener.

2. **Disminución de la Ventaja con Más Datos Etiquetados:** A medida que aumentaba la cantidad de datos etiquetados disponibles, la ventaja del modelo Mean Teacher disminuyó. En el conjunto de datos más grande con 2500 imágenes etiquetadas, el modelo supervisado superó al modelo Mean Teacher en la mayoría de las métricas evaluadas, resaltando la eficacia del aprendizaje supervisado en escenarios con abundantes datos etiquetados.

Limitaciones y Perspectivas de Futuro

8

Este capítulo tiene como objetivo discutir las limitaciones observadas en el presente trabajo, así como identificar diversas perspectivas y líneas de investigación para futuros desarrollos. Las limitaciones identificadas pueden ofrecer una guía para dirigir futuras investigaciones y mejorar los modelos y técnicas utilizados.

8.1. Limitaciones

A continuación, se discuten algunas limitaciones observadas en el presente estudio:

- **Variedad de Datos:** La limitada variedad de datos podría afectar la generalización de los modelos en entornos del mundo real.
- **Hiperparámetros:** Los hiperparámetros utilizados en los modelos fueron seleccionados mediante una búsqueda preliminar, pero no exhaustiva, lo que podría haber impactado en el rendimiento.
- **Tamaño del Dataset:** Los datasets utilizados son relativamente pequeños, lo que podría no reflejar de forma precisa el rendimiento del modelo en datasets más extensos y variados.

8.2. Perspectivas de Futuro

Para futuros trabajos, se proponen las siguientes líneas de investigación y desarrollo:

- **Optimización de Modelos:** Explorar y experimentar con otros modelos semi-supervisados, como el modelo MixMatch o Noisy Student, y técnicas de aprendizaje profundo, como las redes neuronales convolucionales avanzadas, para optimizar el rendimiento en tareas de clasificación de imágenes.
- **Análisis de Sensibilidad:** Implementar un análisis de sensibilidad exhaustivo a los hiperparámetros de ambos modelos, incluyendo la tasa de aprendizaje, la profundidad de la red, y otros, para entender mejor cómo afectan al rendimiento y cómo pueden ser ajustados para mejorar los resultados.

- **Expansión del Dataset:** Ampliar el conjunto de datos con imágenes más variadas y complejas, e inclusión de datos de fuentes externas, para proporcionar una evaluación más robusta y generalizable de los modelos.
- **Evaluación en Contextos Diferentes:** Aplicar y evaluar estos modelos en diferentes contextos y dominios, como el reconocimiento de patrones en imágenes médicas o la clasificación de objetos en entornos industriales, para verificar su eficacia y adaptabilidad a diversos tipos de datos y problemáticas.
- **Reducción de Falsos Negativos:** Desarrollar y explorar técnicas y métodos específicos, como técnicas de re-muestreo o métodos de costo-sensible, que puedan reducir aún más los falsos negativos sin comprometer otras métricas de rendimiento.

Bibliografía

Chest X-Ray Images (Pneumonia).

Chen, G., Ru, J., Zhou, Y., Rekik, I., Pan, Z., Liu, X., Lin, Y., Lu, B., y Shi, J. (2021). MTANS: Multi-Scale Mean Teacher Combined Adversarial Network with Shape-Aware Embedding for Semi-Supervised Brain Lesion Segmentation. *NeuroImage*, 244:118568.

Cui, W., Liu, Y., Li, Y., Guo, M., Li, Y., Li, X., Wang, T., Zeng, X., y Ye, C. (2019). Semi-supervised Brain Lesion Segmentation with an Adapted Mean Teacher Model. In Chung, A. C. S., Gee, J. C., Yushkevich, P. A., y Bao, S., editors, *Information Processing in Medical Imaging*, Lecture Notes in Computer Science, pages 554–565, Cham. Springer International Publishing.

Fan, Y., Kukleva, A., y Schiele, B. (2021). Revisiting Consistency Regularization for Semi-Supervised Learning. arXiv:2112.05825 [cs].

Gm, H., Gourisaria, M., Pandey, M., y Rautaray, S. (2020). A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review*, 38:100285.

Gong, C., Chang, X., Fang, M., y Yang, J. (2018). Teaching Semi-Supervised Classifier via Generalized Distillation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 2156–2162, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization.

Hang, W., Huang, Y., Liang, S., Lei, B., Choi, K.-S., y Qin, J. (2022). Reliability-Aware Contrastive Self-ensembling for Semi-supervised Medical Image Classification. In Wang, L., Dou, Q., Fletcher, P. T., Speidel, S., y Li, S., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, Lecture Notes in Computer Science, pages 754–763, Cham. Springer Nature Switzerland.

Huo, J., Ouyang, X., Si, L., Xuan, K., Wang, S., Yao, W., Liu, Y., Xu, J., Qian, D., Xue, Z., Wang, Q., Shen, D., y Zhang, L. (2022). Automatic Grading Assessments for Knee MRI Cartilage Defects via Self-ensembling Semi-supervised Learning with Dual-Consistency. *Medical Image Analysis*, 80:102508.

Kitada, S. y Iyatomi, H. (2018). Skin lesion classification with ensemble of squeeze-and-excitation networks and semi-supervised learning. arXiv:1809.02568 [cs].

Laine, S. y Aila, T. (2017). Temporal Ensembling for Semi-Supervised Learning. arXiv:1610.02242 [cs] version: 3.

Lee, D.-H. (2013). Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*.

Li, X., Yu, L., Chen, H., Fu, C.-W., Xing, L., y Heng, P.-A. (2020). Transformation Consistent Self-ensembling Model for Semi-supervised Medical Image Segmentation. arXiv:1903.00348 [cs].

Liu, F., Tian, Y., Cordeiro, F. R., Belagiannis, V., Reid, I., y Carneiro, G. (2021). Self-supervised Mean Teacher for Semi-supervised Chest X-Ray Classification. In Lian, C., Cao, X., Rekik, I., Xu, X., y Yan, P., editors, *Machine Learning in Medical Imaging*, Lecture Notes in Computer Science, pages 426–436, Cham. Springer International Publishing.

Liu, Q., Yu, L., Luo, L., Dou, Q., y Heng, P. A. (2020). Semi-supervised Medical Image Classification with Relation-driven Self-ensembling Model. *IEEE Transactions on Medical Imaging*, 39(11):3429–3440. arXiv:2005.07377 [cs].

Livieris, I., Kiriakidou, N., Kanavos, A., Tampakas, V., y Pintelas, P. (2018). On Ensemble SSL Algorithms for Credit Scoring Problem. *Informatics*, 5.

Shaw, S., Pajak, M., Lisowska, A., Tsaftaris, S. A., y O'Neil, A. Q. (2020). Teacher-Student chain for efficient semi-supervised histology image classification. arXiv:2003.08797 [cs, eess, stat].

Su, H., Shi, X., Cai, J., y Yang, L. (2019). Local and Global Consistency Regularized Mean Teacher for Semi-supervised Nuclei Classification. In Shen, D., Liu, T., Peters, T. M., Staib, L. H., Essert, C., Zhou, S., Yap, P.-T., y Khan, A., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Lecture Notes in Computer Science, pages 559–567, Cham. Springer International Publishing.

Tarvainen, A. y Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv:1703.01780 [cs, stat] version: 1.

Unnikrishnan, B., Nguyen, C. M., Balaram, S., Foo, C. S., y Krishnaswamy, P. (2020). Semi-supervised Classification of Diagnostic Radiographs with NoTeacher: A Teacher that is Not Mean. In Martel, A. L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M. A., Zhou, S. K., Racoceanu, D., y Joskowicz, L., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, Lecture Notes in Computer Science, pages 624–634, Cham. Springer International Publishing.

Wang, G., Liu, X., Li, C., Xu, Z., Ruan, J., Zhu, H., Meng, T., Li, K., Huang, N., y Zhang, S. (2020). A Noise-Robust Framework for Automatic Segmentation of COVID-19 Pneumonia Lesions From CT Images. *IEEE transactions on medical imaging*, 39(8):2653–2663.

Yalniz, I. Z., Jégou, H., Chen, K., Paluri, M., y Mahajan, D. (2019). Billion-scale semi-supervised learning for image classification. arXiv:1905.00546 [cs].

Yu, G., Sun, K., Xu, C., Shi, X.-H., Wu, C., Xie, T., Meng, R.-Q., Meng, X.-H., Wang, K.-S., Xiao, H.-M., y Deng, H.-W. (2021). Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images. *Nature Communications*, 12(1):6311.

Yu, L., Wang, S., Li, X., Fu, C.-W., y Heng, P.-A. (2019). Uncertainty-aware Self-ensembling Model for Semi-supervised 3D Left Atrium Segmentation. arXiv:1907.07034 [cs].

Zhu, X. y Goldberg, A. B. (2009). Introduction to semi-supervised learning. In *Introduction to Semi-Supervised Learning*.