Evaluación de desempeño al entrenamiento de una red neuronal profunda usando precisión mixta

Alfredo Morales H.

Instituto de Informática, Universidad Austral de Chile, Valdivia, Chile

Resumen En este trabajo se comparan desempeños de entrenamiento con y sin uso de Precisión Mixta para datos flotantes en el entrenamiento por GPU de una red neuronal profunda que resuelve un problema de clasificación de curvas de luz de estrellas pulsantes tipo RR Lyrae del proyecto *Vista Variables in the Via Lactea* (VVV). Se realizan distintas pruebas para mayor confiabilidad de comparación. Los resultados obtenidos son analizados y cuestionados.

Keywords: Light Curves · Convolutional Neural Networks · High Performance Computing

1. Introduccion

1.1. Contexto y Motivación

Los proyectos de sondaje astronómico actuales y por venir buscan mapear el cielo nocturno en busca de objetos o cuerpos celestes interesantes. Las imágenes digitales capturadas por estos instrumentos son procesadas usando técnicas de fotometría [11], resultando en extensos catálogos de "curvas de luz". Una curva de luz es una serie de tiempo del brillo aparente de una estrella en particular. El tiempo suele medirse en días Julianos mientras que el brillo aparente se mide en magnitudes, una escala logarítmica y relativa. El análisis de curvas de luz le permite astrónomo identificar a qué clase de estrella corresponde y también estimar sus parámetros fundamentales tales como su masa, composición e incluso su distancia real hasta nosotros. Dicho análisis es fundamental para el estudio de estrellas variables, objetos cuyo brillo percibido desde la Tierra varía considerablemente ya sea por razones intrínsecas, es decir asociadas a la física de la estrella, o por razones extrínsecas, como por ejemplo oclusiones periódicas o eclipses causados por otra estrella o planeta. Un tipo particularmente interesante de estrella variable intrínseca son las RR Lyrae (RRL) [3]. Estas estrellas son muy antiguas, con edades comparables a la Vía Láctea. Otra característica que las distingue de otros objetos astronómicos es que estas estrellas pulsan de forma regular, es decir que el tiempo entre mínimos y máximos sucesivos se mantiene constante. En el caso de las RRL este tiempo, denominado período de pulsación, se encuentran entre los 0,3 y 1 día. La Figura 1 muestra un ejemplo de curva

de luz de una estrella RR Lyrae con un período fundamental de 0,557 días. La subfigura izquierda corresponde a la curva de luz original. La subfigura derecha es la curva de luz luego del proceso conocido como *epoch folding*, donde el eje de tiempo se transforma usando

$$\phi_i = \text{modulo}(t_i, P)/P,\ i = 1, 2, \ldots, N$$

donde $\{t_i\}$ son los tiempos de observación y P es el periodo de pulsación. El resultado de esta transformación se conoce como diagrama de fase, y como se muestra en la figura es muy útil para revelar el comportamiento periódico de la estrella, siempre que se conozca el valor correcto de P.

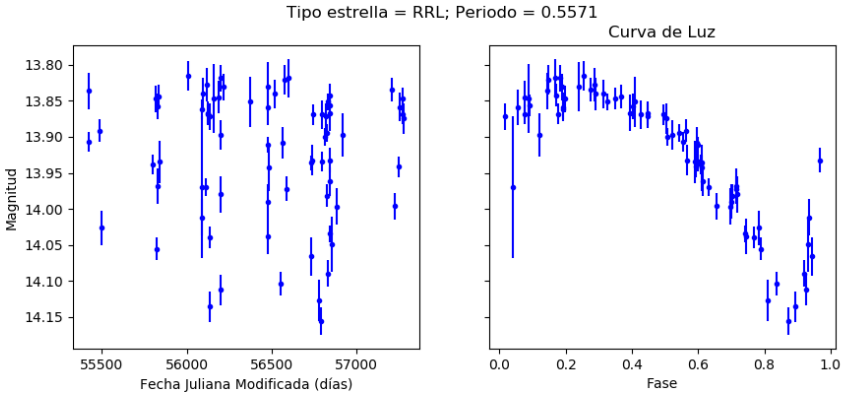

Figura 1. Curva de Luz de estrella RR Lyrae

Usando la magnitud aparente que se estima a partir de las imágenes astronómicas y la magnitud absoluta que es calculada a partir del periodo, es posible deducir la distancia entre una RRL en particular y la Tierra. Esto las convierte en un instrumento fundamental para la medición de distancias en la escala de nuestro vecindario galáctico y en particular para analizar la estructura topológica de la Vía Láctea [10]. Sin embargo, se requiere detectar una gran cantidad de estos objetos estelares para que las mapas sean confiables.

El sondeo VISTA Variables in the Vía Láctea (VVV) [9] es un proyecto de sondaje público financiado por la Unión Europea que observa continuamente nuestra galaxia en el espectro infrarrojo cercano (banda K) usando el instrumento VISTA, localizado en el observatorio Paranal, Región de Antofagasta, Chile. El proyecto VVV busca explicar como se formó la Vía Láctea por medio del análisis de su estructura actual e histórica. Para revelar esta estructura es fundamental la detección de un gran número de estrellas RR Lyrae que habiten en las distintas estructuras de nuestra galaxia. Es por esto que se busca resol-

ver el problema de detección y clasificación de RRL a partir de datos de VVV usando modelos clasificadores basados en redes neuronales artificiales.

1.2. Desafíos

Sondeos astronómicos modernos como VVV producen datos a una tasa que vuelve infactible el análisis manual de la información por parte de los astrónomos. Se reconoce entonces una necesidad por métodos y algoritmos automáticos que apoyen al astrónomo en las tareas de análisis y clasificación de datos [2, 6, 4]. Para enfrentar estos desafíos es necesaria la colaboración y el trabajo interdisciplinar entre astrónomos, matemáticos, científicos de datos e ingenieros. En este contexto un campo de las ciencias de la computación que ha recibido mucha atención es Machine Learning (ML) [7, 1]. Una estrategia de ML consiste en utilizar métodos que extraigan características o atributos de forma automática a partir de los datos. Un ejemplo de este paradigma son las redes neuronales artificiales profundas [5]. En el ámbito de la clasificación de imágenes y series de tiempo, estos métodos representan el estado del arte, alcanzando precisiones más altas que los métodos clásicos. Al no necesitar atributos diseñados a priori se reduce la posibilidad de añadir involuntariamente sesgos al modelo. Una desventaja de este paradigma es que requieren una mayor cantidad de datos para ser entrenados adecuadamente, lo cual a su vez necesita de mayor capacidad de cómputo.

Es aquí donde entramos al campo de la computación de alto rendimiento (HPC por sus siglas en inglés *High Performance Computing*), dado que el entrenamiento de una red profunda con grandes cantidades de datos puede verse beneficiado de gran manera si se realiza sobre tarjetas gráficas con estructuras y arquitecturas dedicadas para este tipo de situaciones como lo son los TensorCores en las tarjetas NVIDIA [8].

En este trabajo se propone entrenar y evaluar una arquitecturaa de red neuronal para resolver el problema de clasificación automática de datos de VVV, haciendo uso de librerías de alto nivel que optimizan de manera automática el tipo de datos según sea requerido para las distintas operaciones de entrenamiento, ya sean las capas convolucionales o lineales, como también en operaciones de reducción, por ejemplo. El principal desafío consiste en adaptar los modelos y estrategias de entrenamiento considerando las características particulares del hardware para entrenamiento.

1.3. Pregunta de investigación

Se quiere resolver la siguiente pregunta de investigación: ¿Se lograrán mejores tiempos de entrenamiento usando Automatic Mixed Precision?

2. Metodología

2.1. Tensor Cores y Automatic Mixed Precision

Las generaciones recientes de GPU NVIDIA, desde Volta en adelante, vienen cargadas con estructuras diseñados especialmente para operaciones matriciales rápidas con FP16 llamadas Tensor Cores. Dichas estructuras aceleran las operaciones comunes de aprendizaje profundo, específicamente las tareas de computación intensiva, como capas convolucionales y completamente conectadas. Sin embargo, hasta ahora estos núcleos tensoriales han seguido siendo difíciles de usar, ya que para ciertas operaciones dinámicas como la reducción, requieren otro formato de datos. Aquí es donde entra el entrenamiento .ªutomático"de precisión mixta o AMP.
AMP es una librería de alto nivel introducida en el núcleo de Pytorch en Marzo de 2020, la cual mediante sencillas instrucciones y ajustes dentro del loop de entrenamiento, permite mezclar múltiples tipos de baja precisión para adaptarse al rango dinámico de las diferentes capas de una red neuronal profunda. Los Tensor Cores se activan cuando ciertos parámetros de una capa son divisibles entre 8 (para datos FP16) o 16 (para datos INT8)

2.2. Dispositivos GPU para entrenar

Se contó con la disposición de dos tarjetas gráficas muy distintas para efectuar pruebas de entrenamiento, si bien los resultados no puedens er comparables entre dichas tarjetas, sí pueden serlo con respecto a los resultados de sí misma para las distintas pruebas.
La primera tarjeta graáfica es una NVIDIA GeForce GTX 1050 ti, la cual Pertenece a un computador de uso doméstico apodado 'Frank'.
La segunda tarjeta gráfica es una NVIDIA GeForce RTX 2080 ti, la cual está diseñada para la computación de alto rendimiento y está montada en el servidor de cómputo de la Universidad Austral de Chile 'Guanaco'.

Specs	**GTX 1050 ti**	**RTX 2080 ti**
Arquitectura	Pascal	Turing
Memoria total [Gb]	4	11
Ancho de banda [GBps]	112	616
Interfaz de memoria	GDDR5	GDDR6
Base Clock [MHz]	1392	1545
Tensor Cores	–	544
Desempeño teórico		
FP16	33.4 GFLOPS	26.9 TFLOPS
FP32	2.1 TFLOPS	13.4 TFLOPS
FP64	66.8 GFLOPS	420 GFLOPS

Cuadro 1. Comparación de especificaciones

2.3. Pruebas

Para la ejecución y posterior comparación de los entrenamientos, se calculan y almacenan dos criterios:

1. Tiempo promedio por época de entrenamiento: se tomó un tiempo inicial al momento de iniciar el loop de entramiento antes de leer el batch de datos del train loader. Luego de todo el loop de train y el de test, se toma un segundo tiempo haciendo la resta con el primero. Ésto se acumulú para todas las épocas y posteriormente se dividía por este número para obtener el tiempo promedio.
2. Puntaje F1 Score de test: este criterio se calcula en base a la exactitud y precisión obtenida de la clasificación de los datos según el modelo versus las etiquetas reales. Es una medida que toma en cuenta las clases que estén representadas en menor cantidad. Al igual que con el tiempo, se guardaba el mejor puntaje por época para luego dividirlo por el total.

Para cada entrenamiento se decidió realizar un total de cinco pruebas con cada configuración, obteniendo los promedios de ellas para una mejor confiabilidad. Lo anterior se realizó para tres tamaños distintos de batch de entrenamiento $[16, 32, 64]$, es decir, del lote de datos que entran a la red. Además, toda la configuración nombrada se ejecutó para una versión con y sin AMP. Por lo tanto, hubo un total de 30 pruebas para cada dispositivo, con 10 épocas cada una.

3. Resultados

Los resultados obtenidos se aprecian en las figuras 2 y 3. en donde en cada una se muestran dos gráficos: el de la izquierda corresponde al tiempo promedio por época de entrenamiento versus el tamaño del batch usado en esas pruebas. En el gráfico de la derecha se muestran los resultados de *aprendizaje* en base al valor de F1 Score versus los tamaños de batch. Cabe mencionar que cada punto del gráfico representa el promedio de cinco entrenamientos con semillas distintas para el muestreo del dataset, cuyas barras de error representan la desviación estándar de dichos promedios.

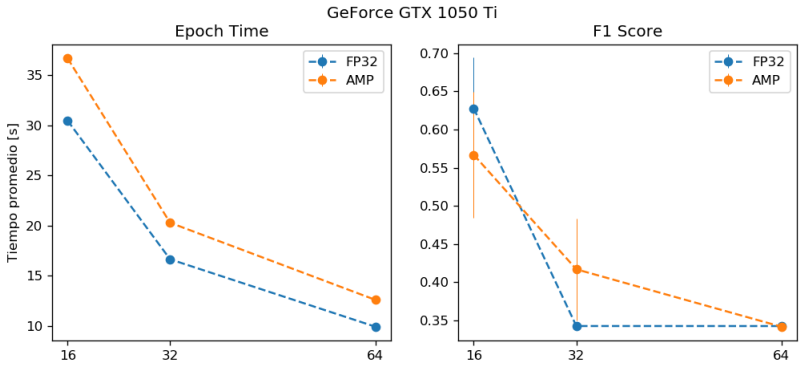

Figura 2. Resultados obtenidos en Frank

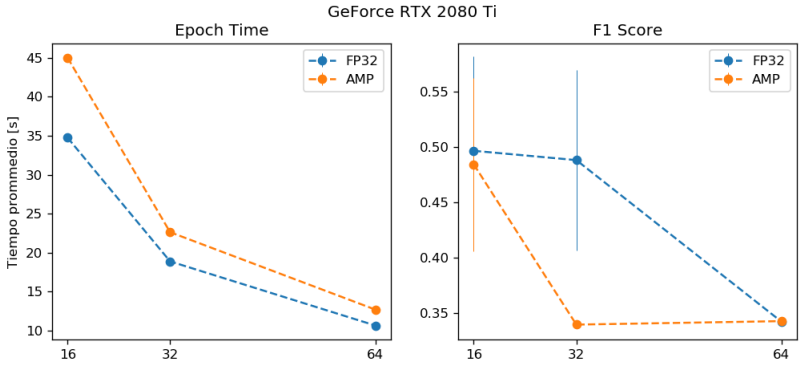

Figura 3. Resultados obtenidos en Guanaco

4. Discusión y Conclusiones

En este trabajo se intentó poner a prueba el uso de precision mixta para el entenamiento de una red neuronal profunda, con el fin de mejorar los tiempos de entrenamiento de la misma gracias al uso de Tensor Cores dentro de una de las tarjetas GPU utilizadas para dicho entrenamiento. Sin embargo, los resultados obtenidos no fueron del todo los esperados. En ambas figuras se aprecian resultados similares, lo cual causa sorpresa dado las características mencionadas de cada dispositivo. Se cree que pueda deberse tanto a una implementación ineficiente de código, como a la estructura de los dataloaders, lo que haya causado que los Tensor Cores no se activaran de manera correcta.
Por otra parte, se sospecha que exista una penalización o demora en los cómputos y sistema de colas en los procesos de Guanaco, dado que el entrenamiento fue ejecutado a través de Jupyter Hub, el cual implementa Python interactivo.

Como trabajo futuro, se espera poder implementar más pruebas dentro de distintos contextos, es decir, con otros datos y también con distintas arquitecturas de modelos de red neuronal, asegurando las configuraciones para un correcto uso y optimización de los equipos disponibles.

Bibliografía

[1] Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg (2006)

[2] Borne, K.D.: Astroinformatics: data-oriented astronomy research and education. Earth Science Informatics **3**(1-2), 5–17 (2010). https://doi.org/10.1007/s12145-010-0055-2

[3] Catelan, M., Smith, H.A.: Pulsating Stars (2015)

[4] Feigelson, E.D.: The changing landscape of astrostatistics and astroinformatics. Proceedings of the International Astronomical Union **12**(S325), 3–9 (2016). https://doi.org/10.1017/S1743921317003453

[5] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), http://www.deeplearningbook.org

[6] Huijse, P., Estevez, P.A., Protopapas, P., Principe, J.C., Zegers, P.: Computational intelligence challenges and applications on large-scale astronomical time series databases. IEEE Computational Intelligence Magazine **9**(3), 27–39 (2014)

[7] MacKay, D.J.C.: Information Theory, Inference & Learning Algorithms. Cambridge University Press, USA (2002)

[8] Markidis, S., Chien, S.W.D., Laure, E., Peng, I.B., Vetter, J.S.: Nvidia tensor core programmability, performance precision. In: 2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). pp. 522–531 (2018)

[9] Minniti, D., Lucas, P., Emerson, J., Saito, R., Hempel, M., Pietrukowicz, P., Ahumada, A., Alonso, M., Alonso-Garcia, J., Arias, J.: Vista variables in the via lactea (vvv): The public eso near-ir variability survey of the milky way. New Astronomy **15**(5), 433–443 (Jul 2010). https://doi.org/10.1016/j.newast.2009.12.002, http://dx.doi.org/10.1016/j.newast.2009.12.002

[10] Saito, R.K., Zoccali, M., McWilliam, A., Minniti, D., Gonzalez, O.A., Hill, V.: Mapping the x-shaped milky way bulge. The Astronomical Journal **142**(3), 76 (2011)

[11] Warner, B.D., et al.: A practical guide to lightcurve photometry and analysis, vol. 300. Springer (2006)