INSTITUTO POLITECNICO NACIONAL

CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN
LABORATORIO DE CIENCIA DE DATOS

GENERACIÓN DE CASCADAS ATMOSFÉRICAS USANDO REDES GENERATIVAS ADVERSARIAS

T E S I S

PARA OBTENER EL TÍTULO DE:

MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

JOSÉ DE JESUS DANIEL AGUIRRE ARZATE

TUTORES:

DR. RICARDO MENCHACA MÉNDEZ

DR. LUKAS NELLEN FILLA

México, CDMX
20 de mayo de 2022

Declaración de Autoría

I, José de Jesus Daniel AGUIRRE ARZATE, declare that this thesis titled, «Generación de cascadas atmosféricas usando redes generativas adversarias» and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

«*If I have seen further it is by standing on the shoulders of Giants.*»

Issac Newton

INSTITUTO POLITECNICO NACIONAL

Resumen

CIC

Centro de Investigación en Computación

Maestría en Ciencias de la Computación

Generación de cascadas atmosféricas usando redes generativas adversarias

por José de Jesus Daniel AGUIRRE ARZATE

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too…

Agradecimientos

The acknowledgments and the people to thank go here, don't forget to include your project advisor…

Índice general

Índice de figuras

Índice de cuadros

Índice de Abreviaturas

HEP	**H**igh **E**nergy **P**hysics
GEANT	**GE**oemtry **AN**d **T**racking
CORSIKA	**CO**smic **R**ay **SI**mulations for **KA**scade
GAN	**G**enerative **A**dversarial **N**etworks
DESY	**D**eutsches **E**lektronen-**SY**nchrotron
HAWC	**H**igh-**A**ltitude **W**ater **C**herenkov
SNOLAB	**S**udbury **N**eutrino **O**bservatory **LAB**oratory
CERN	**C**onseil **E**uropéen pour la **R**echerche **N**ucléaire
HL-LHC	**H**igh **L**uminosity **L**arge **H**adron **C**ollider
LAr	**L**iquid-**Ar**gon
EAS	**E**xtensive **A**ir **S**hower
ASICO	**A**ir shower **SI**mulation and **CO**rrelation

Constantes Físicas

Speed of Light $c_0 = 2.997\,924\,58 \times 10^8\,\mathrm{m\,s^{-1}}$ (exact)

Índice de Símbolos

eV	energy	J
a	distance	m
P	power	W ($\mathrm{J\,s^{-1}}$)
ω	angular frequency	rad

For/Dedicated to/To my…

Capítulo 1

Introduccion

En este capítulo se muestra la principal motivación para el desarrollo de esta tesis, así como una breve perspectiva de la problemática general. Como capítulo introductorio este contiene la introducción, el planteamiento del problema, el objetivo, las fronteras del estudio y la estructura del escrito.

1.1. Introducción

En el área de la física de altas energías (HEP) las técnicas de aprendizaje máquina siempre han estado presentes. Debido a la sorprendente efectividad de técnicas modernas del aprendizaje profundo, se comenzó a adaptar y desarrollar estos métodos en todos los rubros del campo. Algunas de las aplicaciones van desde los enfoques que se tienen en la parte experimental, la fenomenología o en el análisis teórico de los eventos.

En los experimentos más importantes del campo, el tratamiento y análisis de datos es una tarea fundamental. Técnicas como, árboles de decisión, máquinas de soporte vectorial, algoritmos genéticos, entre otras, fallan cuando la dimensionalidad de los datos aumenta. Como referencia de la alta dimensionalidad, en el gran colisionador de hadrones (LHC), las colisiones ocurren con una frecuencia de aproximadamente 40Mhz, además de que cada colisión genera un gran número de partículas y en particular el LHC tiene alrededor $O(10^8)$ sensores para su detección.

Debido a que las observaciones son fundamentalmente probabilísticas se tiene un modelo estadístico que describe la probabilidad de observar un evento dado los parámetros de una teoría. Pero la alta dimensionalidad, junto con los grandes volúmenes de datos generan un problema, ya que el modelo de los datos experimentales no se conoce explícitamente. Sin embargo, si se tiene acceso a muestras de datos generados por simuladores estocásticos que modelan la física de las interacciones.

Herramientas como PYTHIA, HERWING, GEANT, CORSIKA, se les suele denominar como simuladores de Monte Carlo, los cuales cumplen con dos necesidades. La primera es aproximar el modelo estadístico al mostrar de un espacio enorme de procesos no observados o latentes y la segunda es generar una base de datos.

Entre las tareas de bajo nivel se tiene la identificación de partículas y la reconstrucción de la energía/momento de la partícula. Debido a que los simuladores completos que describen las interacciones de las partículas con la materia, son computacionalmente pesados y se llevan gran parte del presupuesto computacional de las colaboraciones, los simuladores rápidos son esenciales.

Simuladores como GEANT y CORSIKA que generan una excelente descripción de interacciones hadrónicas son lentos para eventos de altas energías. En los últimos años ha nacido

un gran interés por usar modelos generativos para aumentar la velocidad de las simulaciones y tal vez llegar a usar estos métodos directamente en datos generados por colisiones reales y hacer ajustes en el momento.

El presente trabajo está fundamentado en el desarrollo de algoritmos de aprendizaje máquina profundo, específicamente el uso de modelos implícitos como arquitecturas adversarias (GAN) para la generación de cascadas atmosféricas. Esto debido a la necesidad de generar simulaciones precisas de una manera más rápida, ya que actualizaciones a dichos experimentos como el de alta luminosidad del LHC (HL-LHC), exigirán una mayor capacidad computacional que no se tiene con la proyección de presupuestos actuales.

Existe un gran interés por parte de la comunidad en usar métodos de aprendizaje no supervisado como GANs o VAEs para generar espacios de características con una alta dimensionalidad. Uno de los mayores desafíos que hay al usar estos métodos, es el de cómo cuantificar su desempeño.

1.2. Planteamiento del problema

El planteamiento se fundamenta en lo siguiente:

Será posible diseñar un método que utilice redes neuronales generativas que logre simular cascadas atmosféricas precisas y así reducir el tiempo computacional de la generación de simulaciones mediante métodos tradicionales.

1.3. Objetivo de la tesis

Objetivo principal:

- Diseñar y implementar una red generativa adversaria para generar cascadas atmosféricas acordes a simulaciones de detectores de rayos cósmicos.

Objetivos particulares:

I. Obtener datos de simulaciones acordes a cascadas atmosféricas.

II. Diseñar y entrenar una red generativa adversaria para la simulación de cascadas atmosféricas.

III. Comparar el tiempo de simulación de la red contra el tiempo que les toma a simuladores tradicionales.

1.4. Delimitación del tema

El presente trabajo se limita a utilizar al menos dos arquitecturas generativas en forma funcional para lograr la generación de respuestas de detectores a una cascada atmosférica extensa. Enfatizando la rapidez de generación de nuevas muestras.

Este trabajo se enfoca en la parte electromagnética de una cascada atmosférica pero no se limita para el análisis de otro tipo de cascadas de partículas o en su defecto, simulaciones de otro tipo de fenómenos físicos que involucran una multitud de procesos probabilísticos.

Aunque no se puede asegurar que sea el mejor método para la aceleración de simulaciones, los próximos estudios deben de intentar desenredar los parámetros latentes y probar la efectividad de otro tipo de arquitecturas.

En este trabajo sólo se utilizarán arquitecturas adversarias (GAN) para acelerar las simulaciones.

1.5. Organización de la tesis

El **Capítulo 1** presenta una breve y concisa introducción a la problemática principal, el planteamiento del problema, el objetivo principal y objetivos particulares del trabajo.

El **Capítulo 2** muestra el estado del arte, así como una breve explicación fenomenológica de las cascadas atmosféricas y la correspondiente operación básica de simuladores modernos. También se introducen las problemáticas principales relacionadas con los simuladores.

En el **Capítulo 3** se introduce el fundamento teórico de los modelos generativos así como conceptos fundamentales del aprendizaje no supervisado. Se hace especial énfasis en las arquitecturas generativas y la teoría que las respalda.

El **Capítulo 4** presenta la metodología experimental que este trabajo sigue para lograr los objetivos propuestos.

En el **Capítulo 5 y 6** presentan los resultados obtenidos junto con una discusión de ellos y las conclusiones a las que este trabajo llegó.

Capítulo 2

Estado del arte

El presente capítulo proporciona el estado del arte mediante la revisión de conceptos y trabajos referentes a cascadas atmosféricas, detectores de partículas, simuladores y modelos generativos para establecer el fundamento del desarrollo de este trabajo.

2.1. Antecedentes

El problema de reducir el costo computacional que experimentos en la física de altas energías (HEP) dedican a simulaciones ha tenido mucha atención en los últimos años, tanto que estudiar técnicas de aprendizaje máquina aplicadas al campo, se incluyeron como un área estratégica de inversiones iniciales para enfrentar los desafíos que actualizaciones como HL-LHC presentarán en los próximos años.

El aprendizaje máquina siempre ha estado presente en los flujos de trabajo de experimentos en HEP, sin embargo técnicas modernas del aprendizaje profundo han comenzado a introducirse a procesos de análisis. En específico se ha visto que los modelos generativos permiten acelerar los tiempos de generación de simulaciones, debido a lo anterior es de suma importancia estudiar estas técnicas para que así puedan ser agregadas a próximas versiones de simuladores como GEANTV, CORSIKA8, etc.

Algunos investigadores que usan modelos generativos para acelerar simulaciones son Paganini [7, 8], Erdmann [4, 5], Carminati [1, 2], Glombitza [6, 3]. Paganini *et al* desarrollaron una arquitectura llamada CaloGAN basada en redes generativas adversarias para acelerar simulaciones de cascadas de partículas en calorímetros LAr y así lograr generar cascadas electromagnéticas tridimensionales con una reducción de tiempo computacional cinco órdenes de magnitud menor de lo que le toma a GEANT4.

Erdmann et al usan una arquitectura WGAN para mejorar la estabilidad del entrenamiento y así poder reconstruir propiedades de la cascada simulada. El modelo lo condicionan a un parámetro físico y logran generar simulaciones de un arreglo de calorímetros. Mostrando así que las simulaciones se pueden adaptar para ajustar datos antes del entrenamiento de la red.

Carminati et al presentan una red generativa adversaria convolucional tridimensional para generar la deposición energética de partículas en calorímetros de alta granularidad. Este trabajo es parte del proyecto GEANTV.

Glombitza usa redes convolucionales para reconstruir el máximo y la energía de una cascada atmosférica usando simulaciones generadas con CORSIKA de un arreglo de detectores para el Observatorio Pierre Auger.

En ambos casos Paganini y Erdmann, proporcionan un modelo generativo que es capaz de generar simulaciones de cascadas de partículas acordes a los simuladores Monte Carlo

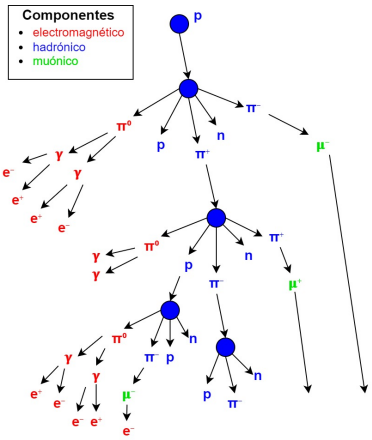

FIGURA 2.1: caption figure 1

existentes. Estos resultados muestran la efectividad de los modelos generativos para reducir los costos computacionales en los actuales flujos de trabajo de los experimentos en HEP.

2.2. Cascadas Atmosféricas

Problemas como conocer la naturaleza de la masa, la naturaleza de la antimateria, la dimensionalidad del espacio o lograr la unificación de las fuerzas fundamentales; se tienen que abordar tanto de los enfoques teórico como experimental para así intentar responder a preguntas fundamentales como, ¿Cuántas partículas hay?, ¿Cuáles son sus propiedades y cómo interactúan?.

Estudiar la fenomenología de los rayos cósmicos es de vital importancia para poder responder las preguntas anteriores. Laboratorios como DESY, SNOLAB, CERN en conjunto con experimentos como el observatorio Pier Auger o el observatorio HAWC entre otros, representan el actual estado del arte del campo de la física de altas energías, como referencia las partículas generadas en las colisiones del LHC viven por fracciones de segundo, lo cual hace que encontrar indicadores de que una partícula fue creada a pesar de nunca ser detectada, sea una tarea artesanal.

Los **rayos cósmicos** son partículas aceleradas en la galaxia o en objetos astrofísicos extragalácticos que al impactar con la atmósfera terrestre generan **cascadas de partículas** (Figura 2.1). El consenso general es que rayos cósmicos debajo de energías de 3x10e6 GeV son acelerados en los remanentes de supernovas galácticas y para energías mayores no se tiene una clara idea de que es lo que acelera a estas partículas, lo único que se tiene claro es de que dichas partículas tienen un orígen extragaláctico. A modo de comparación estas fuentes aceleran partículas en tres órdenes de magnitud mayor que el equivalente energético del LHC.

Cuando la energía de los rayos cósmicos sobrepasa significativamente 1000 GeV estos tienen que ser estudiados por las cascadas de partículas que generan en la atmósfera. La mayoría de las cascadas son iniciadas por hadrones con energías que van desde 1000 GeV o mayores a 10 TeV, que al entrar isotrópicamente a la atmósfera producen un gran número de productos secundarios en una serie de colisiones sucesivas con los núcleos de los constituyentes atmosféricos ie. N2, O2, Ar. Estos productos secundarios resultantes se comportan de una manera similar, mientras se van propagando a través de la atmósfera.

Las cascadas de partículas descritas anteriormente son conocidas como **cascadas atmosféricas extensas** *EAS*. Estas cascadas atmosféricas se propagan longitudinalmente a lo largo

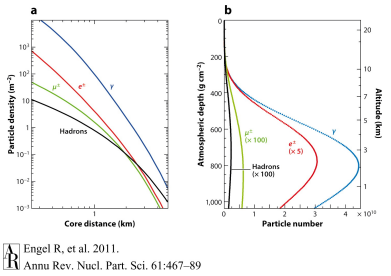

FIGURA 2.2: caption figure 1

de la dirección de incidencia de la partícula primaria, debido al momento transversal de los productos secundarios la cascada también se extiende lateralmente. Como referencia de la complejidad que se tiene, una partícula primaria altamente energética puede crear una cascada gigante de partículas que se propaga esencialmente a la velocidad de la luz a través de la atmósfera y puede alcanzar el nivel del mar si el evento es lo suficientemente energético agregado a lo anterior el número de productos secundarios es del orden 10e10 (Figura 2.2).

2.2.1. Propiedades de las cascadas atmosféricas

Algunas propiedades que caracterizan a las cascadas atmosféricas son:

- **E0** [eV]: energía de la partícula primaria.
- **N** (*shower size*): número total de partículas producidas en un nivel en particular de la atmósfera. Es una función que depende de la energía E0, ángulo de incidencia zenith angle y a altura de la primera interacción del evento primario en la atmósfera h1.
- **Xmax** [$gcm^{-}2$]: profundidad de máximo desarrollo medida desde la parte más alta de la atmósfera. Se desfasa a profundidades mayores conforme la energía del primario se incrementa.
- **Shower Axis**: extensión del vector de momento del primario incidente en la dirección de propagación de la cascada.
- **Dirección de arribo**: dirección de incidencia de la partícula primaria determinada por sus ángulos azimutales y el ángulo zenith.

2.2.2. Métodos de detección

Los principales detectores de cascadas atmosféricas son arreglos de detectores espaciados uno del otro en distancias que dependen de la energía de los rayos cósmicos que se quieren observar. Para energías de 10e6 GeV la distancia entre los detectores deben de ser del orden de decenas de metros y para energías que sobrepasan 10e9 GeV la distancia es del orden de miles de metros, por ejemplo en el observatorio Pierre Auger la distancia entre detectores es de 1500m. Diferentes métodos observacionales como detectores de aire cherenkov o detectores de fluorescencia son combinados con estos arreglos como en el caso del observatorio Pierre Auger (Figura 2.3).

Sin importar el tipo de sistema de detección que se use, los datos adquiridos representan a la cascada en una etapa en particular de su desarrollo, así como una foto instantánea de la cascada, en el plano de observación. Con estos datos se puede conocer información básica que caracteriza a la cascada atmosférica así como los tiempos de arribo de las partículas cargadas, fotones asociados no ópticos, las distribuciones laterales de partículas y fotones en el

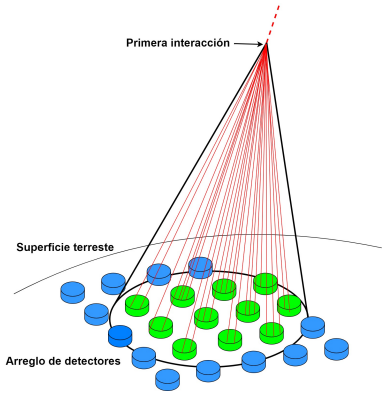

FIGURA 2.3: caption figure 1

plano de observación a una profundidad atmosférica específica. Las propiedades anteriores son inmediatamente accesibles con un arreglo de detectores simple.

La reconstrucción de la partícula primaria depende del modelo hadrónico de interacciones que use el simulador Monte Carlo.

2.3. Análisis de las cascadas atmosféricas

Las cascadas atmosféricas están caracterizadas por un delgado disco de partículas radialmente extenso que se propaga a la velocidad de la luz a través del eje de la cascada. El patrón de las cascadas es circular para cascadas verticalmente incidentes, mientras que la extensión longitudinal y lateral dependen principalmente de la energía de la partícula primaria. El esparcimiento lateral de las partículas en regiones bajas de la atmósfera, llega a cubrir áreas de varios kilómetros cuadrados agregado a lo anterior la mayoría de las partículas arriban en intervalos estrechos de tiempo que van desde unos cuantos nanosegundos en la vecindad del eje de la cascada hasta unos 10 ns a distancias mayores del núcleo de la cascada.

Eventos de baja energía alcanzan su máximo desarrollo en zonas altas de la atmósfera y se mitigan lentamente a mayor profundidad; los componentes que alcanzan a llegar a la superficie son los muones y neutrinos. Para eventos extremadamente energéticos las cascadas logran alcanzar su máximo desarrollo a nivel del mar mientras que sus componentes hadrónicos y electromagnéticos sobrevivientes, son absorbidos en la superficie terrestre y los muones resultantes altamente energéticos continúan propagándose bajo tierra.

Como regla de dedo se puede decir que en promedio una cascada atmosférica está constituida al 1 % por hadrones, alrededor del 10 % son muones y el 90 % o más son electrones o positrones. También para las primeras estimaciones energéticas de la primaria, se tiene que cascadas verticales a una altitud de 5km tienen una energía 1 GeV, 3 GeV para alturas entre 2.5 km y 3 km 10 GeV a nivel del mar.

Para poder visualizar la ocurrencia de estos eventos en distintos regímenes energéticos se puede estudiar el número de partículas que cruzan un área en un tiempo dado, conocido como flujo cósmico. Este flujo sigue una ley de potencias con la forma $\frac{1}{E^3}$, en se puede ver energías alrededor de 10^12 eV el flujo es de 10 partículas primarias por minuto y m^2, para energías entre 10^18 eV y 10^19 eV se tiene 1 partícula primaria por año y km^2 en estos regímenes energéticos las estadísticas son pobres y las incertidumbres altas.

Los observables importantes que se deben de obtener para poder reconstruir una cascada son, el tiempo de arribo t_i al detector i respecto al tiempo de referencia t_0, la densidad de partículas ρ_i y la posición del detector con respecto al sistema de referencia (x_i, y_i). Con la

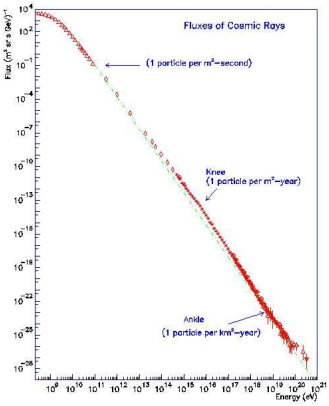

FIGURA 2.4: caption figure 1

distribución lateral de las partículas su puede adquirir la localización del eje de la cascada y el tamaño de la cascada para así obtener un estimado de la energía. Los parámetros anteriores son clasificados como accesibles directamente ya que no necesitan un análisis complejo para su adquisición.

Por último, parámetros indirectamente accesibles que se relacionan con la naturaleza de la partícula primaria, como el tipo de partícula, masa y carga, no se pueden extraer inmediatamente y requieren métodos sofisticados de análisis.

La mayoría de las partículas arriban en intervalos estrechos de tiempo que van desde unos cuantos nanosegundos en la vecindad del eje de la cascada hasta unos 10 ns a distancias mayores del núcleo de la cascada.

2.4. Simulaciones

Simuladores de cascadas atmosféricas son de vital importancia para la evaluación e interpretación de datos experimentales. Las técnicas se reducen a crear e insertar un modelo de cascadas que corresponda a nuestro mejor entendimiento de la realidad, simular cascadas, comparar resultados con los datos experimentales, modificar el modelo o sus parámetros y intentar de nuevo; hacer ajustes pequeños al modelo y repetir hasta que se obtenga un consenso entre la predicción y el experimento.

Las cascadas iniciadas por primarios hadrónicos consisten en la superposición de dos tipos de cascadas, una hadrónica y otra electromagnética. La cascada electromagnética se entiende bien y solo posee problemas prácticos asociados al gran número de partículas participantes en el orden de 10^10. Los programas computacionales que simulan cascadas hadrónicas o electromagnéticas de alta energía o cascadas atmosféricas completas son altamente complejos.

Para tomar en cuenta la complejidad computacional que se tiene, una simulación completa de una cascada debe incluir los componentes electromagnéticos y hadrónicos. Además se debe de tomar en cuenta todos los procesos relevantes, donde la mayoría son de naturaleza estocástica y muchos están en competencia entre sí. También se debe incluir los parámetros que especifican el estado de cada partícula como su masa, carga, energía, momento, ubicación de su creación en el espacio tiempo x,y,z,t, la orientación angular respecto al marco de referencia y parámetros genéticos que revelan la altura de la interacción donde cada partícula fue creada. Estos observables son cruciales para análisis subsecuentes y para la comparación con datos experimentales.

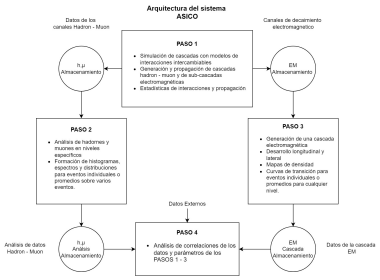

FIGURA 2.5: caption figure 1

2.4.1. Estrategia para simulaciones de EAS

La arquitectura del sistema ASICO sirve como base para entender el proceso de un simulador completo. ASICO fue el primer simulador que generaba cascadas detalladas al usar 12 parámetros que definen a cada partícula y este fue la base para el desarrollo de CORSIKA. Para simular una cascada completa se comienza con la simulación de la cascada hadrónica y se le llama PASO 1, este paso da lugar a datos como la elasticidad, distribuciones de interacciones hadrónicas en diferentes rangos de energías, entre otros; para la parte electromagnética de la cascada se le llama PASO 2 y la combinación de las dos simulaciones para formar la cascada completa se le conoce como el PASO 3.

2.4.2. Problemática de las simulaciones

Los parámetros que determinan a cada partícula se deben de asignar en su punto de creación y requieren actualizarse después de cada proceso al cual está sujeta. Esto es al final de cada trayectoria particular, después de propagarse al siguiente punto de interacción o decaimiento y cuando pasan a un nuevo nivel de observación. En cada actualización la partícula y sus parámetros son guardados para su subsecuente análisis y evaluación de los datos de la cascada simulada.

Consecuentemente, la ejecución de programas que simulan cascadas requieren mucho tiempo computacional particularmente para cascadas energéticas donde el número de partículas involucradas se vuelve muy grande. La gran cantidad de datos producidos por estas simulaciones requieren también una gran capacidad de almacenamiento. La complejidad aumenta si los componentes atmosféricos cherenkov o de fluorescencia se incluyen, en este caso los datos corren el riesgo de divergir y métodos computacionales más sofisticados deben de ser usados para su análisis.

En general el desarrollo y la propagación de una cascada a partir del punto de iniciación (la primera interacción) hasta el nivel de observación consume más tiempo que el análisis subsecuente de los datos producidos.

- **Memoria:** Almacenamiento confiable de los datos crudos.
- **Rastreabilidad:** Rastreo de los parámetros que determinan a cada partícula.
- **Fenomenología:** Propagación de las partículas en la atmósfera tomando en cuenta todos los procesos a los que están sujetas.
- **Configuración inicial:** Correlación confiable entre los parámetros iniciales (*i.e.* modelos de interacción, propiedades del detector, propiedades de la primaria) con la cascada final
- **Tiempo:** Existe una relación lineal entre la energía de la partícula y el tiempo que lleva simular la cascada que genera.

2.5. Modelos generativos

Los modelos generativos son un tipo de aprendizaje no supervisado, que describen cómo se genera un conjunto de datos en términos de un modelo probabilístico. Al muestrear de dicho modelo se es capaz de generar datos no observados previamente. Típicamente el marco de trabajo de los modelos generativos involucra las siguientes partes.

- **Los datos:** Conjunto de observaciones, que se asumen ser generadas de acuerdo a una distribución de probabilidad desconocida.
- **El modelo:** Un modelo generativo que intenta imitar lo mejor posible, a la distribución que genera las observaciones. Este modelo es capaz de generar datos no observados que parecen haber sido generados con la distribución desconocida y no debe de generar datos conocidos.

2.5.1. Aplicación de modelos generativos

Algunas de las tareas modernas de los modelos generativos son:

- **Generación de datos novedosos:** Se generan datos nunca antes vistos que pueden ser utilizados para imitar fenómenos o para ayudar a flujos en modelos discriminativos con un pre entrenamiento autosupervisado.
- **Compresión de datos:** El modelo es capaz de aprender las características más importantes que determinan a la observación y así logra reducir la dimensionalidad del espacio de características donde vive la observación original.
- **Tecnologías de síntesis condicional:** Proporciona un método capaz de generar información novedosa condicionada a un dominio específico. Lo anterior permite una suerte de transformación de datos de un dominio a otro.

Las tareas anteriores han logrado avances en:

- Generacion de rostros humanos
- Transformación de imágenes
- Transferencia de estilos
- Texto a imagen
- Texto a voz
- Edicion de imagenes
- Super resolución
- Generación de objetos 3D
- Predicción de fotogramas en videos

Capítulo 3

Modelos generativos

Este capítulo proporciona una breve y concreta introducción a los modelos generativos, su fundamento teórico y algunas de las principales áreas de investigación. Además de mostrar las principales aplicaciones en HEP.

3.1. Introducción

Así como los modelos discriminativos han sido el centro del progreso en metodologías del aprendizaje máquina en los últimos años ya que es más sencillo monitorear su desempeño y así poder elegir la mejor metodología que se ajuste a la tarea. Los modelos generativos suelen ser más difíciles de evaluar, lo cual hace que encontrar aplicaciones industriales sea más complicado.

Los modelos generativos han probado su efectividad para generar muestras que son capaces de imitar a observaciones reales así como rostros humanos con StyleGAN de NVIDIA o GPT3 de openAI para generar texto. Los modelos anteriores han impulsado el interés para expandir el campo del aprendizaje de máquina a través de modelos que aprenden a generar muestras indistinguibles de observaciones reales. Los avances en el campo podrían ser fundamentales para el desarrollo de una máquina que haya adquirido una inteligencia comparable a la de los humanos.

El campo de los modelos generativos es diverso y la definición de los problemas toman una gran variedad de formas. Sin embargo, para cada tarea, los desafíos que se tienen son los mismos. Entender cómo es que el modelo maneja un alto grado de dependencias condicionales entre las características de la observación y como es que logra encontrar una observación viable, en un espacio de alta dimensionalidad, que concuerda con los datos observados a partir de un conjunto pequeño de observaciones, es de vital importancia para desarrollar metodologías más robustas.

3.1.1. Fundamento teórico

Como punto de partida se debe de reconocer la diferencia clave entre los modelos discriminativos y los modelos generativos.

Def.

Los modelos discriminativos estiman p(y | x) - la probabilidad del observable y dada la observación x.

Los modelos generativos estiman p(x) - la probabilidad de observar x.

En otras palabras, los modelos discriminativos intentan estimar la probabilidad de que una observación x pertenezca a la categoría y, y los modelos generativos intentan estimar la probabilidad de ver la observación x. Por lo tanto el modelo debe de ser probabilístico en vez de ser determinista y debe incluir un elemento estocástico que influencia las observaciones individuales generadas por el modelo.

3.2. Líneas de investigación

Detrás del creciente interés en la academia por los modelos generativos se encuentran dos razones con una gran importancia teórica, que se describen a continuación:

Se debe de buscar un entendimiento completo de cómo se generan las observaciones para así poder formar inteligencias artificiales más sofisticadas que van más allá de lo que pueden lograr los modelos discriminativos.

Es altamente probable que los modelos generativos sean centrales para futuros desarrollos en otros campos del aprendizaje máquina.

3.3. Modelos generativos

Un modelo generativo describe cómo se genera un conjunto de observaciones en términos de un modelo probabilístico. Al generar muestras de este modelo, se es capaz de generar observaciones nunca antes vistas.

Al tener un conjunto de observaciones que representen la entidad que se quiere generar. El objetivo del modelo generativo es generar nuevas observaciones que sigan las mismas reglas con las cuales las observaciones originales fueron generadas. Lo anterior es posible debido a que se asume que existe alguna distribución de probabilidad que explica, porqué ciertas observaciones son más probables de encontrarse en un conjunto y otras no.

El trabajo del modelo es imitar una distribución desconocida lo más cercano posible, para luego muestrear de ella y generar nuevas observaciones, distintas de las conocidas, que además parezca que son parte del conjunto de entrenamiento.

El marco de trabajo de los modelos generativos es el siguiente:

1 Se tiene un conjunto de datos de observaciones X.

2 Se asume que las observaciones X se generaron de acuerdo a una distribución de probabilidad desconocida p_{datos}.

3 Se diseña un modelo generativo p_{modelo} que intenta imitar a p_{datos}.

4 Se muestrea de p_{modelo} para generar observaciones que parecen ser obtenidas de p_{datos}.

Consideramos que p_{modelo} hace un buen trabajo si:

Puede generar observaciones que parecen ser obtenidas de p_{datos}.

Puede generar observaciones que son sustancialmente diferentes a las observadas en X. En otras palabras, el modelo no debería de reproducir cosas que ya conoce.

3.3.1. Conceptos básicos

Def:

Espacio Muestral

El espacio muestral es el conjunto de todos los posibles valores que una observación x puede tomar.

Función de densidad de probabilidad

Una función de densidad de probabilidad, $p(x)$, es una función que mapea un punto x del espacio muestral a un número entre 0 y 1. La suma de la función de densidad sobre todos los puntos del espacio muestral es igual a 1 para que sea una distribución bien definida.

Por definición tenemos que solo existe una p_{datos} pero existen infinitas distribuciones p_{modelo} que pueden estimar p_{datos}. Para encontrar una distribución adecuada se tiene que usar un modelo paramétrico.

Modelo paramétrico

Un modelo paramétrico, $p_\theta(x)$, es una familia de funciones de densidad que pueden ser descritas por medio de un número finito de parámetros, θ.

Capítulo 4

Metodo experimental

En este capítulo se detallan las técnicas utilizadas así como el método experimental para la generación y evaluación de cascadas atmosféricas..

4.1. Recolección y manejo de datos

Dado que el objetivo de este trabajo es implementar un modelo generativo que sea capaz generar simulaciones de cascadas atmosféricas, se necesita un banco de datos de dichas simulaciones. Una de las primeras maneras de obtener este conjunto de datos será tomar el simulador CORSIKA y generar el número de eventos necesarios para poder entrenar a la red neuronal. Como segunda opción de recolección se buscará un banco de simulaciones disponibles al público con características fundamentales como la energía de la primaria, la altura de máximo desarrollo, etc.

Como se ha mencionado en capítulos anteriores, las tareas que cada simulador conlleva para generar una cascada, incluyen subrutinas se encargan de la parte hadrónica, la parte electromagnética, el modelo de interacción, el rastreo de parámetros y procesamientos adicionales para reducir la carga computacional; por lo cual los datos que se utilizaran no deben de concentrarse en alguna de estas subrutinas para poder acotar el dominio de la simulación.

Para generar un resultado significativo sin pérdida de generalidad, se tomarán en cuenta las siguientes consideraciones:

Utilizar un número relativamente “grande” de simulaciones. El número de observaciones debe ser un número “grande”, como referencia se intentará obtener un conjunto de datos con al menos el mismo número de observaciones que el conjunto MNIST.

Preprocesar el conjunto de datos. Para datos como los tiempos de arribo, el dominio de cada simulación diferirá en magnitudes que deben de tomarse en cuenta, por lo tanto técnicas de estandarización podrán ser utilizadas para tener una mejor estabilidad al entrenar el modelo.

Separación en conjuntos de entrenamiento, validación y evaluación. Una vez que se cumpla lo anterior, los datos estarán listos para poder entrenar al modelo.

4.2. Desarrollo del método

A continuación se muestra el método desglosado. Obtener un banco de datos de simulaciones de cascadas atmosféricas acotadas a algún proceso seleccionado.

Hacer el preprocesamiento de datos adecuado para normalizar la entrada de datos y así estabilizar el entrenamiento sin dejar de tomar en cuenta las consideraciones teóricas que dicho procesamiento implica.

Elegir una arquitectura generativa adversaria y entrenarla para intentar generar nuevas cascadas.

Se tomarán las nuevas muestras y se inspeccionarán visualmente comparándolas con las observaciones originales.

Después de haber confirmado la similitud con las observaciones originales, se hará un análisis exploratorio de la dimensionalidad del vector latente y un análisis al proceso determinista de entradas y salidas de dicha red.

Una vez logrado lo anterior, cambiar de arquitectura y repetir el proceso, si se puede.

4.2.1. Mapa del método experimental

Capítulo 5

Resultados

5.1. Dsicusion de resultados obtenidos

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

Capítulo 6

Conlusiones

6.1. Conclusiones puntuales obtenidas

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

6.2. Aplicacion y extension generadas del trabajo

6.3. Trabajos futuros

Apéndice A

Frequently Asked Questions

A.1. How do I change the colors of links?

The color of links can be changed to your liking using:

`\hypersetup{urlcolor=red}`, or

`\hypersetup{citecolor=green}`, or

`\hypersetup{allcolor=blue}`.

If you want to completely hide the links, you can use:

`\hypersetup{allcolors=.}`, or even better:

`\hypersetup{hidelinks}`.

If you want to have obvious links in the PDF but not the printed text, use:

`\hypersetup{colorlinks=false}`.

Bibliografía

[1] F Carminati y col. «Three dimensional Generative Adversarial Networks for fast simulation». En: *Journal of Physics: Conference Series* 1085.3 (2018), pág. 032016. ISSN: 1742-6596. DOI: 10.1088/1742-6596/1085/3/032016. URL: https://iopscience.iop.org/article/10.1088/1742-6596/1085/3/032016https://iopscience.iop.org/article/10.1088/1742-6596/1085/3/032016/meta.

[2] Federico Carminati y col. «Generative Adversarial Networks for fast simulation». En: *Journal of Physics: Conference Series* 1525.1 (2020), pág. 012064. ISSN: 1742-6596. DOI: 10.1088/1742-6596/1525/1/012064. URL: https://iopscience.iop.org/article/10.1088/1742-6596/1525/1/012064https://iopscience.iop.org/article/10.1088/1742-6596/1525/1/012064/meta.

[3] M. Erdmann, J. Glombitza y D. Walz. «A deep learning-based reconstruction of cosmic ray-induced air showers». En: *Astroparticle Physics* 97 (2018), págs. 46-53. ISSN: 0927-6505. DOI: 10.1016/J.ASTROPARTPHYS.2017.10.006. URL: https://www.sciencedirect.com/science/article/pii/S0927650517302219.

[4] Martin Erdmann, Jonas Glombitza y Thorben Quast. «Precise Simulation of Electromagnetic Calorimeter Showers Using a Wasserstein Generative Adversarial Network». En: *Computing and Software for Big Science 2019 3:1* 3.1 (2019), págs. 1-13. ISSN: 2510-2044. DOI: 10.1007/S41781-018-0019-7. URL: https://link.springer.com/article/10.1007/s41781-018-0019-7.

[5] Martin Erdmann y col. «Generating and refining particle detector simulations using the Wasserstein distance in adversarial networks». En: *Computing and Software for Big Science* 2.1 (2018). arXiv: 1802.03325. URL: https://arxiv.org/abs/1802.03325v1.

[6] Jonas Glombitza. «Air-Shower Reconstruction at the Pierre Auger Observatory based on Deep Learning». En: *PoS* ICRC2019 (2020), pág. 270. DOI: 10.22323/1.358.0270. URL: http://www.auger.org/archive/authors_icrc_2019.htmlhttps://inspirehep.net/literature/1819455.

[7] Michela Paganini, Luke de Oliveira y Benjamin Nachman. «Accelerating Science with Generative Adversarial Networks: An Application to 3D Particle Showers in Multi-Layer Calorimeters». En: *Physical Review Letters* 120.4 (2017). DOI: 10.1103/PhysRevLett.120.042003. arXiv: 1705.02355v2. URL: http://arxiv.org/abs/1705.02355http://dx.doi.org/10.1103/PhysRevLett.120.042003.

[8] Michela Paganini, Luke de Oliveira y Benjamin Nachman. «CaloGAN: Simulating 3D High Energy Particle Showers in Multi-Layer Electromagnetic Calorimeters with Generative Adversarial Networks». En: *Physical Review D* 97.1 (2017). DOI: 10.1103/physrevd.97.014021. arXiv: 1712.10321. URL: https://arxiv.org/abs/1712.10321v1.