GEFÖRDERT VOM

Schlussbericht

Verbund: 05H2018 - R&D COMPUTING (Pilotmaßnahme ErUM-Data)

Zuwendungsempfänger:	Universität Hamburg
Projektleitung:	Jun.-Prof. Dr. Gregor Kasieczka
E-Mail:	gregor.kasieczka@physik.uni-hamburg.de

Förderkennzeichen:	05H18GURC1
Förderzeitraum:	01.10.2018 - 30.06.2022
Zuwendung:	189.177,57 €
Projektträger:	Projektträger DESY

Zusätzlicher Kontakt:	gregor.kasieczka@uni-hamburg.de
Zusätzlicher Name:	Gregor Kasieczka

Genutzte Großgeräte:	**Labor**	**Gerät**	**Experiment**
	Themenfeld "Teilchen"	R&D Computing	

Diplomarbeiten:	0
Dissertationen:	0
Habilitationen:	0
Referierte Publikationen:	12
Andere Veröffentlichungen:	3
Patente:	0
Bachelorarbeiten:	3
Masterarbeiten:	1
Staatsexamen:	0

Schlussbericht

Zuwendungsempfänger: Universität Hamburg

Projektleitung: Prof. Kasieczka

Verbund: IDT-UM

Thema: Neue Netzwerkarchitekturen zum automatisierten Lernen und Verstehen fundamentaler Prozesse

Zusammenfassung

Methoden des maschinellen Lernens und der künstlichen Intelligenz spielen eine zunehmend große Rolle in der naturwissenschaftlichen Grundlagenforschung. Aufgrund der großen Mengen und der hohen Komplexität der betrachteten Daten kommt hier der experimentellen Teilchenphysik eine Vorreiterrolle in der Anwendung und Entwicklung dieser Methoden zu.

In diesem Projekt wurden primär zwei Themengebiete behandelt: Einerseits wurde eine Untersuchung Entscheidungsverhalten und den statistischen Eigenschaften von neuronalen Netzwerken durchgeführt. Andererseits wurden Methoden zur Verarbeitung und effizienten Simulation von komplexen Datenstrukturen entwickelt.

Die Untersuchung des Entscheidungsverhaltens und der statistischen Eigenschaften von neuronalen Netzwerken konzentrierte sich auf das Studium von generativen Modellen. Solche Modelle wie Generative Adversarial Networks (GANs), Variational Autoencoder (VAE), oder Normalising Flows können auf einem initialen Datensatz trainiert werden, dessen Wahrscheinlichkeitsverteilung lernen, und daraufhin schnell weitere Beispiele durch Sampling erzeugen. Diese Modelle sind für physikalische Anwendungen von größter Bedeutung da in diesem Forschungsbereich oft umfassende Simulationen erzeugt werden müssen, finden jedoch auch breiten Einsatz in anderen Bereichen. Das aktuell große gesellschaftliche Interesse an die Möglichkeit zur Erzeugung von Kunstwerken mit KI sei hier als Beispiel genannt.

Auf der Seite der Verarbeitung und effizienten Simulation von komplexen Datenstrukturen konnten große Erfolge in der Simulation von Teilchenschauern in hochauflösenden Kalorimetern erzielt werden.

Ein Hindernis in der Verarbeitung von komplexen physikalischen Daten liegt in der Datenstruktur. Hier ist der Übergang von fixen Strukturen wie Gittern zur Darstellung der Daten als Graph oder Punktwolke und die entsprechende Behandlung mit neuronalen Netzwerken eine wichtige Herausforderung. Gemeinsam mit anderen Gruppen im IDT-UM Verbund wurde ein übergreifender öffentlicher Datensatz geschaffen und gezeigt wie eine einheitliche Grapharchitektur zu deren Analyse verwendet werden kann.

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung gefördert. Die Verantwortung für den Inhalt der Veröffentlichung liegt beim Autor.

Bericht

1. Aufgabenstellung und Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Entwicklung von Ansätzen des maschinellen Lernens und des Deep Learnings auf angewandt auf Probleme in der Teilchenphysik. Unter anderem getrieben durch massiv zunehmende Datenmengen aus unterschiedlichen Quellen nimmt die Bedeutung dieser statistisch komplexen multivariaten Auswertungstechniken stetig zu. Im Projekt sollten spezifisch das “Entscheidungsverhalten von neuronalen Netzwerken” sowie sowie die "Verarbeitung von Spur- und Kalorimeterdaten mit Zeitinformation” in diesem Kontext untersucht werden.

Dabei konnte auf Vorerfahrung in der Entwicklung und Anwendung von Methoden des maschinellen Lernens sowie umfassende Kompetenz in der Gewinnung und Auswertung von teilchenphysikalischen Daten zurückgegriffen werden.

2. Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Zu Projektbeginn gab es zwar bereits umfassende Ergebnisse zur Anwendung von KI auf teilchenphysikalische Daten jedoch limitiert in Tiefe und Umfang. Generative Modelle waren bereits vorgeschlagen, hatten aber nur auf relativ einfachen Datensätzen realisiert werden können. Ebenso war das Problem der Suche nach passenden Symmetrien zwar bekannt, Architekturen wie graphbasierte Modelle aber noch nicht verbreitet.

3. Planung und Ablauf des Vorhabens sowie Kooperation mit Dritten

Es sollten zwei PhD-Positionen besetzt werden - jeweils eine zu den beiden vertretenen Themen. Aufgrund des überraschenden Ausscheidens einer Person musste eine der beiden Stellen neu besetzt. Dabei fand eine Justierung des Schwerpunkts von der Berücksichtigung spezifisch von Zeitinformation zur allgemeinen Entwicklung von zum Umgang mit komplexen Datenstrukturen geeigneten Architekturen satt. Weiters wurde der Schwerpunkt vermehrt auf generative Modelle gelegt mit denen bereits erste Erfolge erzielt worden waren.

Enge Zusammenarbeit mit der DESY Future Colliders Gruppe (Dr. Gaede, Dr. Krueger); der Gruppe von Prof. Plehn in Heidelberg sowie mit anderen Gruppen im IDT-UM Verbund.

4. Verwendung der Zuwendung (wichtigste Positionen des zahlenmäßigen Nachweises, z. B. Investitionen, Personalmittel)

Buhmann: Primär: Entscheidungsverhalten von neuronalen Netzwerken
Schnake: Primär: Verarbeitung von Spur- und Kalorimeterdaten mit Zeitinformation
Korcari: Primär: Verarbeitung von Spur- und Kalorimeterdaten mit Zeitinformation

5. Erzielte Ergebnisse mit Gegenüberstellung der vereinbarten Ziele

Es konnten wegweisende Ergebnisse in der Simulation von hochauflösenden Detektoren mit generativen Modellen erzielt werden. Ebenso wurden grundlegende Erkenntnisse zu den statistischen Eigenschaften von generativen Modellen erzielt werden die über den eigentlich betrachteten wissenschaftlichen Kontext relevant sind. Weiters konnte gezeigt werden wie physikalische Eigenschaften der Daten in latenten Räumen abgebildet werden, und wie diese Beobachtung zur Verbesserung von generativen Modellen verwendet werden kann.

Es wurde ein gemeinsamer Datensatz von wissenschaftlichen Daten aus unterschiedlichen Disziplinen zusammengestellt und gezeigt wie eine einheitliche graphbasierte Netzwerkarchitektur zur Analyse dieser Daten eingesetzt werden kann.

Wie bereits erwähnt wurde bedingt durch einerseits die veränderte Personalsituation und andererseits durch die großen Erfolge auf generativer Seite eine Verstärkung dieser Themen auf Kosten der expliziten Berücksichtigung von Zeitinformation durchgeführt.

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung gefördert. Die Verantwortung für den Inhalt der Veröffentlichung liegt beim Autor.

6. Notwendigkeit und Angemessenheit der geleisteten Arbeit

Maschinelles Lernen und künstliche Intelligenz bieten großes Transformationspotential in der wissenschaftlichen Grundlagenforschung. Die erreichten Ergebnisse zeigen das Potential von generativen Modellen und graphbasierten Methoden konnten damit zum wissenschaftlichen Erkenntnisgewinn und der Effizienzsteigerung im wissenschaftlichen Prozess beitragen. Die Arbeit war damit notwendig und konnte das durchgeführte Projekt konnte aus keinen anderen Quellen finanziert werden.

Der Beitrag von PhD-Studierenden ist von großer Bedeutung in der Forschung und im Vergleich zu Personen mit höherem Qualifikationsgrad kosteneffektiv. Die geleistete Arbeit war damit sicherlich angemessen.

7. Voraussichtlicher Nutzen, insbesondere Verwertbarkeit der Ergebnisse

Der Schwerpunkt des Forschungsprojekts lag in der Entwicklung von neuen Methoden zur wissenschaftlichen Datenanalyse mittels maschinellem Lernen sowie der Untersuchung der Eigenschaften dieser Methoden. In beiden Fällen liegt ein Hauptziel in der wissenschaftlichen Kommunikation der erzielten Ergebnisse. Dies wurde durch Publikation der Arbeiten und die Vorstellung auf internationalen Workshops und Konferenzen erzielt. Soweit angebracht wurden die verwendeten Software-Codes und relevanten Daten ebenfalls öffentlich zugänglich gemacht um die einfache Nachnutzung zu gewährleisten.
Die Resultate erlauben aufgrund der höheren Leistungsfähigkeit von Methoden des maschinellen Lernens im Vergleich zu klassischen statistischen Techniken eine Effizienzsteigerung in der Auswertung von physikalischen Daten. Gleichzeitig sind die Erkenntnisse zu statistischen Eigenschaften grundlegend genug um auch in anderen Wissenschaftsfeldern Anwendung zu finden.
Über die Forschung hinaus wurde auch umfassende Nachwuchsarbeit in Form der Ausbildung von BSc-, MSc- und PhD-Studierenden in gesellschaftlich und wirtschaftlich relevanten Fähigkeiten in der Entwicklung von modernen digitalen Methoden geleistet.

8. Während der Durchführung des Vorhabens dem Zuwendungsempfänger bekannt gewordenen Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Die Entwicklung von Methoden des maschinellen Lernens für Anwendungen in der Teilchenphysik ist ein hochaktives Forschungsfeld mit mehr als hundert Publikationen pro Jahr. Es war jedoch keine Anpassung der Forschungsziele aufgrund von Ergebnissen anderer Gruppen notwendig.

9. Erfolgte und geplante Veröffentlichungen der Ergebnisse

9.1. Referierte Publikationen (z. B. in Fachzeitschriften oder -büchern und referierte Konferenzproceedings)

[1] GANplifying Event Samples, A. Butter, S. Diefenbacher, G. Kasieczka, B. Nachman, T. Plehn, SciPost Phys. 10, 139 (2021), 2008.06545
[2] DCTRGAN: Improving the Precision of Generative Models with Reweighting
S. Diefenbacher, E. Eren, G. Kasieczka, A. Korol, B. Nachman, D. Shih, JINST 15 P11004, 2009.03796
[3] Calomplification -- The Power of Generative Calorimeter Models, S. Bieringer, S. Diefenbacher, E. Eren, F. Gaede, D. Hundhausen, G. Kasieczka, B. Nachman, T. Plehn, M. Trabs, JINST 17 P09028, 2202.07352
[4] Symmetries, Safety, and Self-Supervision, B. Dillon, G. Kasieczka, H. Olischlager, T. Plehn, P. Sorrenson, L. Vogel, SciPost Phys. 12, 188 (2022), 2108.04253
[5] Getting High: High Fidelity Simulation of High Granularity Calorimeters with High Speed, E. Buhmann, S. Diefenbacher, E. Eren, F. Gaede, G. Kasieczka, A. Korol, K. Krüger

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung gefördert. Die Verantwortung für den Inhalt der Veröffentlichung liegt beim Autor.

May 2020, published in: Comput Softw Big Sci 5, 13, 2005.05334
[6] Hadrons, Better, Faster, Stronger, E. Buhmann, S. Diefenbacher, E. Eren, F. Gaede, D. Hundhausen, G. Kasieczka, W. Korcari, K. Krüger, P. McKeown, L. Rustige, Mach. Learn.: Sci. Technol. **3** 025014, 2112.09709
[7] Shared Data and Algorithms for Deep Learning in Fundamental PhysicsL. Benato, E. Buhmann, M. Erdmann, P. Fackeldey, J. Glombitza, N. Hartmann, G. Kasieczka, W. Korcari, T. Kuhr, J. Steinheimer, H.Stöcker, T. Plehn, K. Zhou, July 2021, Comput Softw Big Sci 6, 9 (2022), 2107.00656
[8] Deep-Learning Jets with Uncertainties and More, S. Bollweg, M. Haußmann, G. Kasieczka, M. Luchmann, T. Plehn, April 2019, published in: SciPost Phys. 8 (2020) 1, 006, e-print: 1904.10004
[9] Per-Object Systematics using Deep-Learned Calibration, G. Kasieczka, M. Luchmann, F. Otterpohl, T. Plehn, March 2020, published in: SciPost Phys. 9, 089 (2020), e-print: 2003.11099

9.2. Andere Veröffentlichungen (z. B. Konferenzbeiträge wie Vorträge und Poster, unreferierte Proceedings, Conference Notes)

[10] Decoding Photons: Physics in the Latent Space of a BIB-AE Generative Networs, E. Buhmann, S. Diefenbacher, E. Eren, F. Gaede, G. Kasieczka, A. Korol, K. Krüger, February 2021, arXiv:2102.12491
[11] Amplifying Statistics using Generative Models, A. Butter, S. Diefenbacher, G. Kasieczka, B. Nachman, T. Plehn, December 2020, published as: ML4PS Workshop at NeurIPS 2020
[12] Generative Models for Particle Shower Simulation in Fundamental Physics
E. Buhmann, S. Diefenbacher, E. Eren, F. Gaede, D. Hundhausen, G. Kasieczka, W. Korcari, A. Korol, K. Krüger, P. McKeown, L. Rustige in: simDL Workshop at ICLR 2021
[13] Amplifying Statistics with Ensembles of Generative Models, A. Butter, S. Diefenbacher, G. Kasieczka, B. Nachman, T. Plehn in: simDL Workshop at ICLR 2021
[14] Machine Learning and LHC Event Generation, A. Butter (ed), T. Plehn (ed), S. Schumann (ed), S. Badger, S. Caron, K. Cranmer, F. A. Di Bello, E. Dreyer, S. Forte, S. Ganguly, D. Gonçalves, E. Gross, T. Heimel, G. Heinrich, L. Heinrich, A. Held, S. Höche, J. N. Howard, P. Ilten, J. Isaacson, T. Janßen, S. Jones, M. Kado, M. Kagan, G. Kasieczka, F. Kling, S. Kraml, C. Krause, F. Krauss, K. Kröninger, R. K. Barman, M. Luchmann, V. Magerya, D. Maitre, B. Malaescu, F. Maltoni, T. Martini, O. Mattelaer, B. Nachman, S. Pitz, J. Rojo, M. Schwartz, D. Shih, F. Siegert, R. Stegeman, B. Stienen, J. Thaler, R. Verheyen, D. Whiteson, R. Winterhalder, J. Zupan, March 2022, e-print: 2203.07460
[15] New directions for surrogate models and differentiable programming for High Energy Physics detector simulation, A. Adelmann, W. Hopkins, E. Kourlitis, M. Kagan, G. Kasieczka, C. Krause, V. Mikuni, B. Nachman, K. Pedro, D. Shih, D. Winklehner
March 2022, e-print: 2203.08806

9.3. Abschlussarbeiten (Bachelor, Master, Diplom, Staatsexamen, Promotion, Habilitation)

[16] Latent Space Representations of Calorimeter Data, Bachelor thesis by Ronja vom Schemm, November 2019
[17] Automating Machine learning in Particle Physics, Bachelor thesis by Nils Gerber, June 2020
[18] Statistics of Generative Models in Physics, Master thesis by Daniel Hundhausen, September 2021
[19] The Blessing of Dimensionality: Event Manifold Dimensionality Estimation for Event Clustering. Bachelor thesis by Malte Jacobsen, September 2021

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung gefördert. Die Verantwortung für den Inhalt der Veröffentlichung liegt beim Autor.

Kurzbericht

- öffentlich -

Zuwendungsempfänger: Universität Hamburg

Projektleitung: Prof. Kasieczka

Verbund: IDT-UM

Thema: Neue Netzwerkarchitekturen zum automatisierten Lernen und Verstehen fundamentaler Prozesse

1. Ziel und Inhalt des Projektes

Methoden des maschinellen Lernens und der künstlichen Intelligenz spielen eine zunehmend große Rolle in der naturwissenschaftlichen Grundlagenforschung. Aufgrund der großen Mengen und der hohen Komplexität der betrachteten Daten kommt hier der experimentellen Teilchenphysik eine Vorreiterrolle in der Anwendung und Entwicklung dieser Methoden zu.

In diesem Projekt wurden primär zwei Themengebiete behandelt: Einerseits wurde eine Untersuchung Entscheidungsverhalten und den statistischen Eigenschaften von neuronalen Netzwerken durchgeführt. Andererseits wurden Methoden zur Verarbeitung und effizienten Simulation von komplexen Datenstrukturen entwickelt.

2. Ablauf und Ergebnisse des Vorhabens

Die Untersuchung des Entscheidungsverhaltens und der statistischen Eigenschaften von neuronalen Netzwerken konzentrierte sich auf das Studium von generativen Modellen. Solche Modelle wie Generative Adversarial Networks (GANs), Variational Autoencoder (VAE), oder Normalising Flows können auf einem initialen Datensatz trainiert werden, dessen Wahrscheinlichkeitsverteilung lernen, und daraufhin schnell weitere Beispiele durch Sampling erzeugen. Diese Modelle sind für physikalische Anwendungen von größter Bedeutung da in diesem Forschungsbereich oft umfassende Simulationen erzeugt werden müssen, finden jedoch auch breiten Einsatz in anderen Bereichen. Das aktuell große gesellschaftliche Interesse an die Möglichkeit zur Erzeugung von Kunstwerken mit KI sei hier als Beispiel genannt.
Konkret wurde dabei zuerst untersucht wie die physikalischen Eigenschaften eines Datensatzes von Teilchenschauern mit dem latenten Raum von generativen Surrogatmodellen korrelieren. Dies erlaubt einerseits die Interpretation und das bessere Verständnis der Modelle, andererseits konnten so direkt Verbesserungsmöglichkeiten in der Gestaltung der Netzwerkarchitektur identifiziert und realisiert werden.
Eine weitere wichtige Frage beschäftigt sich mit der statistischen Sinnhaftigkeit von generativen Modellen. Dabei konnte an einem allgemeinen Beispiel gezeigt werden dass die Ausgabe von generativen Modellen in der Tat eine bessere Annäherung an die zugrundeliegende Verteilung erzielt als die zum Training verwendeten Daten. In konnte dieses Ergebnis auf teilchenphysikalisch relevante Verteilungen erweitert werden.

Letztlich wurden auch die Bestimmung der inheränten Dimension von Datenräumen, die Nutzung von Symmetrieeigenschaften, sowie Bayessche Methoden zur Quantifizierung von Unsicherheiten betrachtet.
Auf der Seite der Verarbeitung und effizienten Simulation von komplexen Datenstrukturen konnten große Erfolge in der Simulation von Teilchenschauern in hochauflösenden Kalorimetern erzielt werden. Die Resultate zur Simulation von elektromagnetischen und hadronischen Teilchenschauern stellen dabei jeweils die bisher exakteste Simulation des Verhaltens von komplexen Detektoren durch generative Modellen dar und sind richtungsweisend für den weiteren Einsatz dieser Methoden.
Ein Hindernis in der Verarbeitung von komplexen physikalischen Daten liegt in der Datenstruktur. Hier ist der Übergang von fixen Strukturen wie Gittern zur Darstellung der Daten als Graph oder Punktwolke und die entsprechende Behandlung mit neuronalen Netzwerken eine wichtige Herausforderung. Gemeinsam mit anderen Gruppen im IDT-UM Verbund wurde ein übergreifender öffentlicher Datensatz geschaffen und gezeigt wie eine einheitliche Grapharchitektur zu deren Analyse verwendet werden kann.

3. Konkreter Nutzen sowie Anwendungsmöglichkeiten der Ergebnisse

Der Schwerpunkt des Forschungsprojekts lag in der Entwicklung von neuen Methoden zur wissenschaftlichen Datenanalyse mittels maschinellem Lernen sowie der Untersuchung der Eigenschaften dieser Methoden. In beiden Fällen liegt ein Hauptziel in der wissenschaftlichen Kommunikation der erzielten Ergebnisse. Dies wurde durch Publikation der Arbeiten und die Vorstellung auf internationalen Workshops und Konferenzen erzielt. Soweit angebracht wurden die verwendeten Software-Codes und relevanten Daten ebenfalls öffentlich zugänglich gemacht um die einfache Nachnutzung zu gewährleisten.
Die Resultate erlauben aufgrund der höheren Leistungsfähigkeit von Methoden des maschinellen Lernens im Vergleich zu klassischen statistischen Techniken eine Effizienzsteigerung in der Auswertung von physikalischen Daten. Gleichzeitig sind die Erkenntnisse zu statistischen Eigenschaften grundlegend genug um auch in anderen Wissenschaftsfeldern Anwendung zu finden.
Über die Forschung hinaus wurde auch umfassende Nachwuchsarbeit in Form der Ausbildung von BSc-, MSc- und PhD-Studierenden in gesellschaftlich und wirtschaftlich relevanten Fähigkeiten in der Entwicklung von modernen digitalen Methoden geleistet.