GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Schlussbericht

Verbund: 05M2020 - MaGriDo

Zuwendungsempfänger:	Rheinische Friedrich-Wilhelms-Universität Bonn
Projektleitung:	Prof. Dr. Jochen Garcke
E-Mail:	garcke@ins.uni-bonn.de
Förderkennzeichen:	05M20PDB
Förderzeitraum:	01.04.2020 - 28.02.2024
Zuwendung:	295.424,28 €
Projektträger:	Projektträger DESY
Zusätzlicher Kontakt:	
Zusätzlicher Name:	

Genutzte Großgeräte:	Labor	Gerät	Experiment

Diplomarbeiten:	0
Dissertationen:	1
Habilitationen:	0
Referierte Publikationen:	4
Andere Veröffentlichungen:	5
Patente:	0
Bachelorarbeiten:	1
Masterarbeiten:	4
Staatsexamen:	0

Dieser Bericht wurde beim Projektträger über einen individuellen Online-Zugang vom Projektleiter eingereicht und am 30.08.2024 17:36 für eine Veröffentlichung freigegeben.

Schlussbericht

Zuwendungsempfänger:	Rheinische Friedrich-Wilhelms-Universität Bonn Regina-Pacis-Weg 3, 53113 Bonn Ausführende Stelle: Institut für Numerische Simulation
Projektleitung:	Prof. Dr. Jochen Garcke
Verbund:	MaGriDo: Mathematik für maschinelle Lernmethothoden für Graph-basierte Daten mit integriertem Domänenwissen.
Thema:	Teilprojekt 1: Integration von Domänenwissen, Erklärbarkeit und Transferlernen

Zusammenfassung

Methoden des Deep Learning sind in den letzten Jahren erfolgreich für verschiedene Problemstellungen angewendet worden (Bilderkennung etc.). Hier wurden bisher meist sogenannte „end-to-end“ Lernansätze verwendet. Zu deren Umsetzung sind in der Regel sehr große Mengen von strukturierten Daten notwendig sind, welches dazu führt, dass diese in vielen möglichen Anwendungsfällen aus den Naturwissenschaften, Medizin und Industrie nur bedingt einsetzbar sind. Ziel des Verbundvorhabens MaGriDo war es daher, neue Ansätze zu entwickeln, zu analysieren und auf Problemstellungen anzuwenden, die es erlauben existierendes Wissen in die Architektur der Netzwerke einzubauen und somit ermöglichen von den komplementären jeweiligen Stärken von „end-to-end“ Lernansätze und „a priori Modellen/Regeln“ zu profitieren. Solch ein Vorgehen verspricht substantiell effizientere Lösungen für viele der genannten Anwendungsfelder zu ermöglichen.

Da üblicherweise komplexe Systeme sehr gut als Zusammensetzungen von Entitäten und deren Wechselwirkungen repräsentiert werden können, lag der Schwerpunkt der Forschung und Entwicklung in MaGriDo auf sogenannten Graphnetzwerken. Diese enthalten zum Beispiel konventionelle Fully-Connected-NN, Convolution-NN und Recurrent-NN als Spezialfall, können insbesondere auf relationalen Strukturen angewendet werden und ermöglichen eine hierarchische Prozessierung der Eingabedaten.

Die Arbeiten im Teilprojekt 1 haben sich auf die geeignete Auswahl kleiner Trainingsmengen aus großen Pools unmarkierter Datenpunkte konzentriert. Dies ist von wesentlicher Bedeutung, um die Modellleistung zu maximieren und gleichzeitig die Effizienz zu erhalten.

Insbesondere wird Farthest Point Sampling (FPS) untersucht, ein Ansatz zur Datenauswahl, der darauf abzielt, die sogenannte Fülldistanz der ausgewählten Menge zu minimieren. Auf theoretischer Seite wird eine obere Schranke für den maximal erwarteten Vorhersagefehler gezeigt, welche vor allem linear von der Fülldistanz der Trainingsmenge abhängt. Zudem wird empirisch beobachtet, dass die Auswahl eines Trainingssets mit FPS. d.h. mit dem Ziel die Fülldistanz zu minimieren, den maximalen Vorhersagefehler verschiedener Regressionsmodelle signifikant reduziert und alternative Sampling-Ansätze deutlich übertrifft.

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.

Um die durchschnittliche Qualität der Vorhersagen von kontinuierlichen Lipschitz-Regressionsmodellen zu verbessern, wurde mit ”Density Aware Farthest Point Sampling”(DA-FPS) ein neues Datenauswahlverfahren entwickelt. Es zielt darauf, eine neu eingeführte gewichtete Fülldistanz zu minimieren, welche sowohl Punktabstände als auch Verteilungsüberlegungen berücksichtigt. Wir haben eine obere Schranke für den erwarteten durchschnittlichen Vorhersagefehler abgeleitet, die linear von der gewichteten Fülldistanz abhängt. Empirische Untersuchungen bestätigt die theoretischen Analysen.

Bericht

1 Aufgabenstellung und Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Aufgabenstellung Verbundvorhaben:

Ziel des Verbundprojektes war es, tiefe neuronale Netzwerke (TNNs) für Problemstellungen aus der Industrie (weiter) zu entwickeln und zu analysieren, die es erlauben, existierendes Domänenwissen in die Architektur der Netzwerke einzubauen, wobei der Schwerpunkt der Forschung und Entwicklung in auf sogenannte Graphnetzwerke (GNNs) gelegt wurde. Die mathematischen Untersuchungen im Vorhaben sollen zum besseren Verständnis der Architektur von tiefen neuronalen Netzen beisteuern. Die mathematischen Beiträge des Vorhabens umfassen:

- die Untersuchung und Erarbeitung einer Expressivitätsanalyse für Graphnetzwerken,
- die Erweiterung von Interpretierbarkeitsalgorithmen und deren (neue) Theorie auf Graphnetzwerke,
- die Untersuchung effizienter Lernverfahren für das Training von Graphnetzwerken,
- die Betrachtung von Graphnetzwerkarchitektur, Regularisierung und Optimierungsverfahren,
- die strukturierte Integration von Domänenwissen durch geeignete mathematisch Formalismen,
- die Erarbeitung von mathematischen Konzepten zum Transferlernen und aktiven Lernen,
- die Erarbeitung und Untersuchung von systematisch verbesserbaren Wechselwirkungspotentialen,
- die Untersuchung generativer Netzwerke.

Aufgabenstellung Teilprojekt

Voraussetzungen:

Das Verbundkonsortium besteht aus den Gruppen:

- (Verbundkoordinator): Arbeitsgruppe Garcke (Universität Bonn),
- Arbeitsgruppe Kutyniok (LMU München),
- Arbeitsgruppe Lorenz (TU Braunschweig),
- Arbeitsgruppe Hamaekers (Fraunhofer SCAI).

In dem Vorhaben arbeiteten universitäre Arbeitsgruppen mit Expertise in den Disziplinen numerische Mathematik, Analysis, Optimierung, und Angewandte Funktionalanalysis mit der Universtität Bonn zusammen. Die universitären Arbeitsgruppen haben komplementäre Schwerpunkte und adressierten im Vorhaben unterschiedliche Aspekte der betrachteten Lernverfahren, wobei das verbindende Element das Design der Architektur war, sei es aus Sicht der Integration des Domänenwissens und des Vergleichs von Verfahren des Maschinellen Lernens (Bonn), der Expressivität und Interpretierbarkeit (München), des Trainierens (Braunschweig) und der konkreten Anforderungen aus der Anwendung (SCAI mit den assoziierten Projektpartnern). Die Universität Bonn hat insbesondere Expertise im informierten Maschinellen Lernen, der Approximation in hochdimensionalen Räumen und

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.

hat mit verschiedenen Verfahren des Maschinellen Lernens gearbeitet und diese numerisch untersucht.

2 Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Der Stand der Forschung zeigt, dass das maschinelle Lernen, insbesondere das tiefe Lernen, große Erfolge bei der Mustererkennung in verschiedenen Bereichen erzielt hat, wie z.B. Computer Vision, Spracherkennung, Textverständnis und Spiele-KIs. Diese Erfolge basieren auf dem datenbasierten Ansatz des Lernens aus einer Vielzahl von Beispielen. Allerdings stoßen rein datengetriebene Ansätze in einigen Fällen an ihre Grenzen oder führen zu unbefriedigenden Ergebnissen. Zum Beispiel, wenn nicht genügend Daten vorhanden sind, um komplexe und genaue Modelle zu trainieren, oder wenn das Ziel darin besteht, die Vorhersagequalität zu verbessern oder die Trainings- und Inferenzzeit zu reduzieren. Darüber hinaus besteht ein Bedarf an interpretierbaren Modellen und an Transfer-Lernen-Ansätze, um die Abstraktionsfähigkeiten von maschinellen Lernmodellen zu erhöhen. Diese Aspekte werden weitgehend in den Arbeitspaketen 2, 3 und 4 von den jeweiligen universitären Arbeitsgruppen bearbeitet. Am Fraunhofer SCAI lag der Schwerpunkt auf der Entwicklung und Umsetzung von Graphnetzwerken für Anwendungen in der Materialentwicklung, wobei der Fokus auf entsprechenden Fragestellungen bezüglich Molekülen, Polymeren und Gläsern lag.

Neben den genannten klassischen Bereichen wird das maschinelle Lernen, und insbesondere das tiefe Lernen, immer wichtiger und erfolgreicher in verschiedenen Bereichen der Ingenieur- und Naturwissenschaften. Diese Erfolgsgeschichten basieren auf dem datenbasierten Ansatz des Lernens aus einer Vielzahl von Beispielen, siehe auch [RMB+23]. In vielen Fällen stoßen rein datengetriebene Ansätze jedoch an ihre Grenzen oder führen zu unbefriedigenden Ergebnissen. Das offensichtlichste Szenario ist, dass nicht genügend Daten zur Verfügung stehen, um ausreichend komplexe und genaue Modelle zu trainieren. Darüber hinaus kann es das Ziel sein, die Vorhersagequalität eines Modells zu steigern oder dessen Trainings- oder Inferenzzeit zu reduzieren. Ein weiterer wichtiger Aspekt ist, dass ein rein datengetriebenes Modell möglicherweise nicht den Anforderungen entspricht, die beispielsweise durch Naturgesetze oder regulatorische oder Sicherheitsrichtlinien vorgegeben sind. Mit maschinellen Lernmodellen, die immer komplexer werden, gibt es auch einen steigenden Bedarf für Modelle, die interpretierbar und transparent sind, vgl. für Quantenmechanik für Moleküle auch [SAC+17].

All diese Aspekte haben zu einer verstärkten Forschung geführt, wie man maschinelle Lernmodelle verbessern kann, indem man Domänenwissen explizit in den Lernprozess integriert [RMB+23]. Da üblicherweise komplexe Systeme sehr gut als Zusammensetzungen von Entitäten und deren Wechselwirkungen repräsentiert werden können, eignen sich zur Behandlung vieler Problemstellungen Graphnetzwerke [WPC+19], die den Einbau von Domänenwissen in vielen Fällen ermöglichen [BHB+18].

Darüber hinaus gibt es ein Interesse daran, die Abstraktionsfähigkeiten von Modellen des maschinellen Lernens zu erhöhen, um Ansätze des Transferlernens zu erleichtern. Der Ansatz des Transferlernens ist es, auf Basis eines (großen) Datensatzes gewonnenes Wissen zu nutzen, insbesondere in Form des dazugehörigen gelernten ML-Models, um es auf ein anderes, aber verwandtes Problem mit nur wenig gelabelten Daten anzuwenden [PY10]. Die Aufgabe des Transferlernens wurde in der Vergangenheit durch verschiedene Ansätze angegangen. Einer davon ist der so genannte instanzbasierte Transfer, bei dem jede Instanz ein gewisses Gewicht erhält, um anzugeben, wie viel Einfluss sie auf die Vorhersage der eigentlichen Zieldaten haben wird. In kernbasierten Ansätzen wurde eine spezielle Kernmatrix gelernt, die die Ähnlichkeiten zwischen den zu transferierenden und den Zieldaten widerspiegelt. Zudem gibt es Methoden, einen informativen Prior zu lernen, gelernte Modellparameter zu übertragen, oder die Daten in gemeinsame

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.

Merkmalsräume zu projizieren. Entsprechende Referenzen sind in [GV14] oder [PY10] zu finden. In [GV14] wird zudem ein kernbasierter Ansatz vorgestellt, bei der mittels Importance Sampling ein gewichteter Kern zum Transferlernen bestimmt wird.

Bisher konzentrierten sich die meisten Algorithmen des Transferlernens auf die Verbesserung der Verallgemeinerung über verschiedene Verteilungen zwischen Quell- und Zielbereichen oder -aufgaben hinweg. In vielen Anwendungen soll jedoch Wissen über Domänen oder Aufgaben mit unterschiedlichen Charakteristiken hinweg übertragen werden. Insbesondere für TNNs ist die Übertragung der Merkmalsdarstellung in den letzten Jahren erfolgreich angewandt worden. Hierbei werden vortrainierte TNNs genutzt, um diese auf neuen, kleineren Daten nur noch nachzulernen. Vereinfacht gesagt, wird häufig die obere, oder die obersten Schichten, neu gelernt, während die vorher gelernten tieferen Schichten im Wesentlichen beibehalten werden. Bei Graphnetzwerken im adressierten Anwendungsbereich ist Transferlernen beispielsweise in [CZZ+18] erfolgreich mittels erlernter Darstellungen, welche periodische chemische Trends repräsentieren.

3 Planung und Ablauf des Vorhabens sowie Kooperation mit Dritten

Die Planung und Durchführung des Projektes gliederte sich in folgende Arbeitspakete, wobei der jeweilige Teilprojektkoordinator in Klammern vermerkt ist:

- Arbeitspaket AP0 (UB): Koordination:
  Das Projektmanagement hat die Aufgabe die verfügbaren Ressourcen in kontrollierter und strukturierter Weise zu verwalten. Nicht nur der reibungslose Ablauf des Projekts sollte sichergestellt werden, auch die dafür notwendigen Grundlagen sollten geschaffen werden, wie die Kommunikation innerhalb des Konsortiums und mit dem Projekträger.
- Arbeitspaket AP1 (alle, je 2 PM): Demonstratoranwendung:
  Spezifikation einer gemeinsamen Demonstratoranwendung und notwendiger Datenformate, sowie Realisierung von Schnittstellen, gemeinsamer Datenrepositorien und Coderepositorien.
- Arbeitspaket AP2 (LMU): Expressivität und Interpretierbarkeit
- Arbeitspaket AP3 (TUBS): Training und Architektur
- Arbeitspaket AP4 (UB): Integration von Domänenwissen, Erklärbarkeit und Transferlernen
  In diesem Arbeitspaket sollen einerseits die verschiedenen Ansätze zur Integration von Domänenwissen in Verfahren des Maschinellen Lernens untersucht und weiter aus Sicht der Mathematik strukturiert werden. Als zweiter großer Aspekt wird hier Transferlernen genauer untersucht.
  - AP4.1: Nutzung und Generierung von Domänenwissen
  - AP4.2: Bezüge zwischen tiefen neuronalen Netzen und kernbasierten Verfahren
- Arbeitspaket AP5 (SCAI): Anwendung in der Materialentwicklung
- Arbeitspaket AP6 (SCAI): Praxisrelevante Anwendung

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

Es sind im Wesentlichen nur Personalkosten und Reisekosten zu berücksichtigen.

5 Erzielte Ergebnisse mit Gegenüberstellung der vereinbarten Ziele

Die folgende Darstellung umfasst im Detail nur die Aufgaben und Ergebnisse des Teilprojektes 1 „Integration von Domänenwissen, Erklärbarkeit und Transferlernen“ des Verbundprojektes MaGriDo, d.h. AP 1 und AP 4. Hinsichtlich der Ziele und Ergebnisse der Arbeitspakete:

- Arbeitspaket AP2 (TUB): Expressivität und Interpretierbarkeit,
- Arbeitspaket AP3 (TUBS): Training und Architektur,

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.

- Arbeitspaket AP5 (SCAI): Anwendung in der Materialentwicklung
- Arbeitspaket AP6 (SCAI): Praxisrelevante Anwendung

wird auf die Berichte der anderen Teilprojekte verwiesen.

5.1 Arbeitspaket AP1: Demonstratoranwendung (gemeinsam)

Um eine gemeinsame Sicht auf die Aufgabenstellungen und eine Synergie der verschiedenen mathematischen Forschungsarbeiten zu ermöglichen, sollte zum Beginn des Projekts eine geeignete Demonstratoranwendung identifiziert werden. Diese sollte zum einen nah an den Anforderungen der Anwender sein, aber andererseits die (auch gemeinsamen) mathematischen Untersuchungen ermöglichen.

Hier wurde die Vorhersage von Moleküleigenschaften auf dem Benchmarkdatensatz QM9 [Rea:12] als Demonstratoranwendung identifiziert und auch die notwendigen Datenformate, Schnittstellen, gemeinsame Datenrepositorien und Coderepositorien realisiert.

5.2 Arbeitspaket AP4: Integration von Domänenwissen, Erklärbarkeit und Transferlernen

Für das AP4.1 "Nutzung und Generierung von Domänenwissen" haben zum Anfang des Projekts Untersuchungen stattgefunden und erste Einordnungen in eine Taxonomie wurden durchgeführt. Durch den unerwarteten Abgang des ursprünglichen Projektmitarbeiters wurde nach dem "Neustart" des Projekts evaluiert, wie weit es angebracht und effizient ist diese Arbeiten wieder aufzugreifen. Da durch den Übersichtsartikel von Noé. et. al. [NTM+20], veröffentlicht in 2020, ein sehr guter Stand der Wissenschaft aufgearbeitet und beschrieben wurde, wurde entschieden, dass der neue Mitarbeiter sich auf die Forschungsaspekte zum Vergleich von Graph Neural Networks und Kernen fokussiert. Es war nicht zu erwarten war, dass die Fortführung der ursprünglich geplanten Übersichtsaktivitäten zur Integration von Domänenwissen im Rahmen der zur Verfügung stehen Laufzeit einen zusätzlichen wissenschaftlichen Beitrag über den Übersichtsartikel [NTM+20] hinaus geliefert hätte.

Zum allgemeinen Kontext des informierten Maschinellen Lernens wurde der zum Start des Projekts existierender Übersichtsartikel nach weitgehender Überarbeitung als „Informed machine learning - A taxonomy and survey of integrating knowledge into learning systems“ veröffentlicht [RMB+23].

Es fanden theoretische Untersuchung der mathematischen Formalisierungen für die verschiedenen auf Fachwissen basierenden neuronalen Netzarchitekturen statt. Der Schwerpunkt lag vor allem auf der Untersuchung der mathematischen Formalisierung von graphbasierten neuronalen Netzen (GNN), die auf molekularen Graphen trainiert werden. Eine wesentliche Strategie zur Einbeziehung von molekülbezogenem Fachwissen in GNN, die auf molekularen Graphen trainiert werden, besteht darin sicherzustellen, dass das Netz zusätzlich zu der Grundvoraussetzung der Knotenpermutation äquivariant in Bezug auf die euklidische Gruppe E(3) starrer Bewegungen (Rotationen, Translationen und Reflexionen) ist [SHW21]. Übersichten über die sehr aktive Forschung zu graphbasierten neuronalen Netzen finden sich in [RNE+22, WLB23].

Zudem wurden Kernel-Methoden für die Vorhersage molekularer Eigenschaften untersucht. Die Konzentration auf Kernel-Methoden wurde durch ihr solides mathematisches Verständnis motiviert, das sie interpretierbar macht. Insbesondere wurde festgestellt, dass die Wirksamkeit von Kernel-Methoden nicht nur von der Wahl des Kernels, sondern auch von der Wahl des molekularen Deskriptors abhängt. Daher wurde eine solche Untersuchung der Kernel-Methoden parallel zur Analyse der verschiedenen verfügbaren molekularen Deskriptoren, z. B. SOAP, FCHL und Mordred, durchgeführt. Diese Untersuchung führte zur Wahl eines bestimmten molekularen Deskriptors, des Mordred-Deskriptors.

Im Rahmen der empirischen Untersuchung von verschiedenen Lernverfahren auf großen Datensätzen, wie beispielsweise dem aus dem OpenCatalyst-Projekt, wurde klar dass eine Auswahl

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.

von Teilmengen der Datensätze notwendig ist, damit überhaupt sinnvolle empirische Studien von verschiedenen Modellen mit vertretbarem Rechenaufwand durchführbar sind. Dabei stellt sich raus, dass ein randomisierter Ansatz zwar einfach durchzuführen ist, aber nicht die besten und nicht robuste Ergebnisse liefert. Vielfach verwendete Methoden des aktiven Lernens basieren auf den Vorhersagen eines konkreten ML-Verfahren, die so ausgewählten Daten sind nicht zwingend nützlich für andere ML-Verfahren. Basierend darauf, fand ein Vergleich der Vor- und Nachteile verschiedener Strategien zur Auswahl von Trainingsdatensätzen für Regressionsprobleme, insbesondere für die Vorhersage von Moleküleigenschaften und Kraftfeldnäherungsaufgaben statt. Es wurden Strategien für die Auswahl eines effizienten Trainingssatzes untersucht und verglichen, um Ansätze zu ermitteln, von denen mehrere Lernmodelle profitieren können. Damit soll erreicht werden, dass der Aufwand für das Labeln nicht auf Teilmengen verschwendet wird, die nur für ein Modell oder eine Modellklasse nützlich sind, beziehungsweise nur für eine Analyseaufgabe nutzbar sind.

Beim aktiven Lernen werden die Parameter des Lernalgorithmus iterativ aktualisiert und die Unsicherheiten für nicht gekennzeichnete Proben vorhergesagt. Aktives Lernen kommt jedoch in der Regel nur einem bestimmten Modell, einer bestimmten Modellklasse oder einer konkreten Analyseaufgabe zugute. Andererseits werden beim passiven Lernen Stichproben ausschließlich auf der Grundlage ihrer Position im Merkmalsraum ausgewählt. Passive Lernstrategien können zudem in zwei Klassen unterteilt werden: modellabhängig und modellagnostisch. Es wurde festgestellt, dass die Erforschung modellunabhängiger passiver Lerntechniken eine genauere Untersuchung verdient. Solche Strategien haben das Potenzial, die Leistung mehrerer Lernmodelle zu verbessern und zu vermeiden, dass der Aufwand für die Kennzeichnung von Teilmengen verschwendet wird, die nur für ein bestimmtes Modell oder eine bestimmte Klasse von Modellen von Nutzen sind.

Eine passive Auswahlstrategie kann zudem beim aktiven Lernen selbst als Baustein im Prozess sinnvoll sein. Insbesondere im Anwendungsszenario, wo das Labeln eine aufwändige Simulation umfasst, kann es sinnvoll sein, mittels Strategien das aktiven Lernens eine Kandidatenmenge zu identifizieren, aus denen dann mittels weiteren Labelunabhängigen passiven Strategien die konkret zu labelnden Daten ausgewählt werden. Nur mittels aktiven Lernens mehrere zu labelnde Daten auszuwählen, ist oft ineffizient, da aus dem aktiven Lernen häufig nah beieinander liegende Kandidaten ausgewählt werden. Es stellt sich dann häufig heraus, dass ein Datenpunkt von mehreren in der Nachbarschaft schon gereicht hätte, um die Vorhersagequalität zu verbessern, und somit teure Simulation unnötig durchgeführt wurden. Hier kann eine passive Strategie helfen. Somit adressieren die durchgeführten Untersuchungen zum passiven Lernen letztendlich auch das aktive Lernen im Anwendungskontext nach der ursprünglichen Projektplanung.

Wobei wir uns im Projekt, auch aus Zeitgründen, auf die Untersuchung von passiven, modelagnostischen Auswahlstrategien konzentrieren um diese eingehend zu analysieren. Die Kombination von aktivem und passivem Lernen ist vorgesehene Forschung nach Projektablauf.

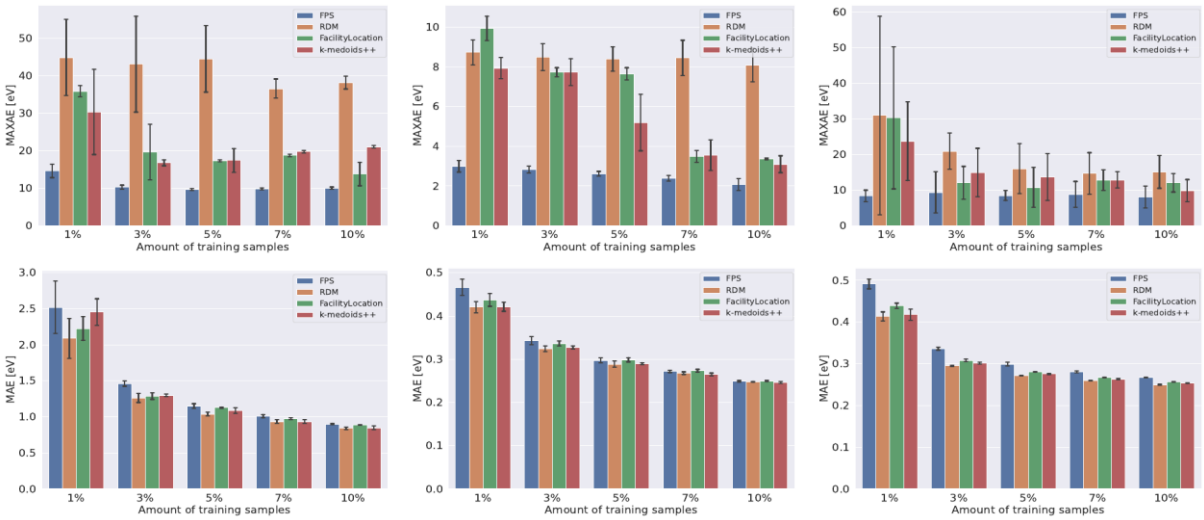
Eine erste empirische Untersuchung der Auswirkungen verschiedener modellunabhängiger Strategien zur Auswahl von Trainingsdaten auf die Vorhersagequalität von maschinellen Regressionslernmodellen für die Vorhersage molekularer Eigenschaften wurde durchgeführt. Die Untersuchung konzentrierte sich in erster Linie auf modellunabhängige und auf Domänenwissen basierende Auswahlstrategien und ihre Auswirkungen auf die Vorhersagequalität von Kernel-Ridge-Regressionsmodellen (KRR) und Graph Neural Networks. Diese empirische Studie in Zusammenarbeit mit den Projektpartnern führte zu einem gemeinsamen Artikel ‘On the Interplay of Subset Selection and Informed Graph Neural Networks’ [BCG+24], welcher im Laufe des Jahres 2023 akzeptiert wurde, die Veröffentlichung als Buchbeitrag steht für 2024 an. Es wurde hierbei die Schlussfolgerung gezogen, dass die Ausnutzung des Domänenwissens zur Maximierung der molekularen Vielfalt im Auswahlprozess der Trainingsmenge die Genauigkeit und Robustheit dieser

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.

Modelle verbessert. Insbesondere zeigen topologiebasierte ML-Modelle, die auf vielfältigen Molekülsätzen trainiert wurden, eine signifikante Verringerung des maximalen absoluten Fehlers im Test. Genauer gesagt, die Datenauswahl erfolgte anhand molekularer Deskriptoren auf der Grundlage von Mordred, einer Bibliothek, die Domänenwissen nutzt, um topologische Darstellungen molekularer Graphen zu erstellen. Die Auswahl der Trainingsdaten erfolgt dann mit dem Farthest-Point-Sampling-Algorithmus (FPS).

Zum theoretischen Verständnis sind sogenannte Fülldistanzen aus der numerischen Analysis von Approximationsverfahren verwendet worden. Vereinfacht gesagt, gibt die Fülldistanz eine Schranke an, wie weit bereits ausgewählte Trainingsdaten von den noch nicht ausgewählten Daten entfernt sind. Es wurden theoretische und empirische Untersuchungen über die Auswirkungen der Minimierung der Fülldistanz bei der Vorhersagequalität von kontinuierlichen Lipschitz-Regressionsmodellen durchgeführt. Diese Eigenschaft gilt sowohl für tiefe neuronale Netze und kernbasierte Verfahren und greift damit auch den untersuchten Aspekten der Bezüge zwischen diesen Lernverfahren auf. Insbesondere beweisen wir eine obere Schranke für den maximal zu erwartenden Vorhersagefehler von Lipschitz-kontinuierlichen Lernalgorithmen ein, die von der Kenntnis der Datenmerkmale abhängt und die linear von der Fülldistanz der Trainingsmenge abhängig ist. Wir zeigen, dass die Auswahl des Trainingssatzes mit dem Ziel, die vorgeschlagene Schranke zu minimieren, d.h. durch Minimierung der Fülldistanz des Trainingssatzes, den maximalen Approximationsfehler verschiedener Lernalgorithmen, wie z.B. Kernel-Regressionsmodelle, Feed-Forward-Neuronale Netze und Graph-Neuronale Netze, verringert. Die Auswahl von Trainingssätzen mit FPS minimiert dabei die Fülldistanz bis auf einen Faktor zwei, ein besserer Faktor wäre ein NP Problem. Wir haben den Ansatz zur Minimierung der Fülldistanz zudem mit anderen modernen modellunabhängigen Sampling-Techniken verglichen. Die Ergebnisse bei der Nutzung von kernbasierten Regressionsverfahren sind Abbildung 1 zu sehen, es zeigen sich beim maximalen Fehler konsistent die Vorteile des FPS-basierten Verfahrens.

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.



(a) KRR on QM7 (b) KRR on QM8 (c) KRR on QM9

Abbildung 1: Ergebnisse für Regressionsaufgaben auf QM7, QM8 und QM9 unter Verwendung von kernbasierten Regressionsverfahren, die auf Mengen unterschiedlicher Größe trainiert wurden, ausgedrückt als Prozentsatz der verfügbaren Datenpunkte und ausgewählt mit unterschiedlichen Stichprobenstrategien. Maximaler Fehler (MAXAE) (obere Zeile) und mittlerer absoluter Fehler (MAE) (untere Zeile) sind für jede Trainingsmenge und jeden Stichprobenansatz dargestellt.

Zusätzlich haben wir eine umfangreiche numerische Analyse durchgeführt, um besser zu verstehen, wie der FPS-Auswahlprozess funktioniert. Insbesondere ging es darum, zu verstehen, welche Punkte durch den FPS-Auswahlprozess priorisiert werden. Diese Analyse hat gezeigt, dass FPS Punkte an den Enden der Datenverteilung bevorzugt und Punkten in dünn besiedelten Regionen des Merkmalsraums zu Beginn des Sampling-Prozesses eine höhere Relevanz zuweist. Darüber hinaus hat unsere experimentelle Analyse in Bezug auf die Möglichkeiten der FPS den empirischen Beweis erbracht, dass die FPS besonders bei geringen Trainingsdatenbudgets von Vorteil ist. Wir haben zudem argumentiert, dass es möglicherweise geeignetere Sampling-Strategien als FPS gibt, um größere Datenmengen auszuwählen.

Desweiteren haben wir theoretisch und empirisch gezeigt, dass der FPS-basierte Auswahlprozess bei Kernel-Regressionsmodellen zu einer erhöhten Modellstabilität führt. Die Stabilität der Regressionsmodelle wurde anhand der Zustandszahl der berechneten Kernelmatrix gemessen. Wir konnten auf Basis von Aussagen der numerischen Approximationstheorie zeigen, dass die Auswahl eines Trainingssatzes mit dem FPS zu einer Kernelmatrix mit einer im Vergleich kleinen Kondition führt und sich somit die Modellstabilität im Vergleich zu den anderen von uns betrachteten Benchmark-Auswahlstrategien erhöht.

Diese Ergebnisse wurden in einem Artikel mit dem Titel ”On minimizing the training set fill distance in machine learning regression” 2023 auf arXiv veröffentlicht und sind 2024 im Journal of Data-centric Machine Learning Research erschienen [CG24a]. Der Code zu diesem Artikel ist veröffentlicht und frei zur Verfügung gestellt worden [CG24d].

Neben diesen Untersuchungen bezüglich des maximal zu erwartenden Vorhersagefehlers, haben wir zudem untersucht, wie Moleküle auf der Grundlage ihrer topologischen Deskriptoren ausgewählt werden können, um die durchschnittliche Qualität der Vorhersagen von kontinuierlichen Liptschitz-Regressionsmodellen zu verbessern und gleichzeitig den maximalen Fehler gering zu halten. Diese Studie führte zur Entwicklung von ”Density Aware Farthest Point Sampling”(DA-FPS), einem neuen Datenauswahlverfahren, das darauf abzielt, eine gewichtete Fülldistanz der ausgewählten Menge zu

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.

minimieren. Die gewichtete Fülldistanz ist eine skalare Größe, die wir definiert haben, um zu quantifizieren, wie gut die Trainingsmenge die interessierende Merkmalsdomäne repräsentiert, wobei sowohl Punktabstände als auch Verteilungsüberlegungen berücksichtigt werden. Wir haben eine obere Schranke für den erwarteten Vorhersagefehler von Lipschitz-kontinuierlichen Regressionsmodellen abgeleitet, die linear von der gewichteten Fülldistanz der Trainingsmenge abhängt. Wir haben bewiesen, dass DA-FPS suboptimale Minimierer für das von uns definierte gewichtete Fülldistanz-Minimierungsproblem liefert und somit suboptimale Minimierer für die vorgeschlagene Schranke bereitstellt. Wir haben den DA-FPS-Ansatz zur Minimierung der gewichteten Fülldistanz zudem mit anderen modernen modellunabhängigen Sampling-Techniken verglichen. Es zeigen sich konsistent die Vorteile des DA-FPS-Verfahrens. Um zudem einen Vergleich mit einem Verfahren des aktiven Lernens zu haben, nutzen wir einen Orakelansatz. Es werden jeweils auf Basis der Fehler auf den Testdaten die Daten ausgewählt, welche den größten Fehler zwischen KRR-Vorhersage und den exakten Labels haben. Das ist in der Praxis natürlich nicht möglich, aber in den Experimenten. Es zeigt sich, dass eine Auswahl mit den Orakel-Aktiven Lernens zwar bei KRR auf Dauer die besten Ergebnisse liefert, diese Daten aber bei Feedforward Neural Networks die schlechtesten Ergebnisse liefert. Somit sind die mit aktivem Lernen ausgewählten Trainingsdaten nicht gut für manche andere Modellklassen geeignet. Es müsste noch eingehender untersucht werden, wie sich das bei anderen Modellklassen verhält. Diese Ergebnisse sind Abbildung 2 zu sehen.

Diese wissenschaftlichen Ergebnisse wurden in einem Artikel mit dem Titel ”Density Aware Farthest Point Sampling” 2024 aufgeschrieben, der aktuell unter Begutachtung ist [CG24b].

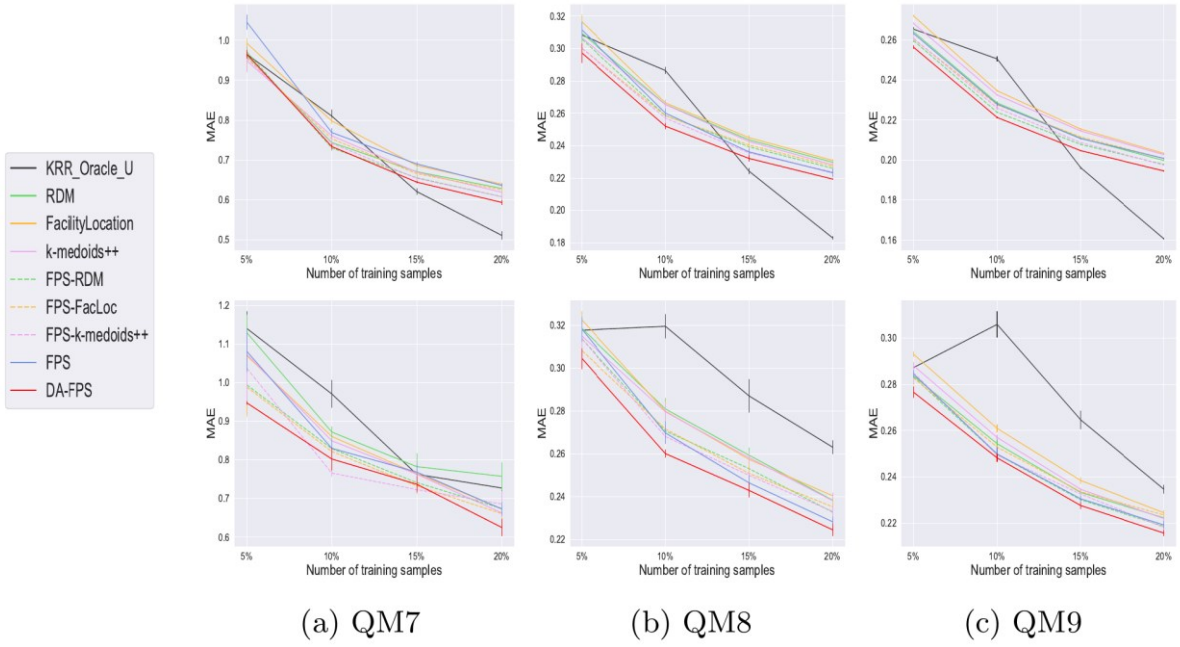


Abbildung 2: MAE für Regressionsaufgaben auf QM-Datensätzen unter Verwendung von KRR mit dem Gauß-Kernel (oben) und Feedforward Neural Networks (FNN) (unten), trainiert auf Datensätzen unterschiedlicher Größe und mit verschiedenen Sampling-Strategien, einschließlich des zusätzlichen aktiven Best Oracle-Aktiven Lernansatzes auf der Grundlage von KRR: "KRR Oracle U". Die Fehlerbalken stellen den Standardfehler des Mittelwerts über fünf Durchläufe dar.

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

Durch die Forschungsarbeiten an praxisrelevanter Methodenentwicklung zur Anwendung und Analyse von sogenannten Graphnetzwerken wurden an den beteiligten Universitäten und dem Fraunhofer-Institut SCAI weitergehende bzw. neue Expertise aufgebaut, die ohne externe Finanzierung nicht geleistet hätte werden können. Es sind neue Erkenntnisse erarbeitet worden, die in verschiedenen, allgemein verfügbaren Publikationen veröffentlicht wurden, auch um weitere Forschungsarbeiten von anderen zu ermöglichen. Die Notwendigkeit und Angemessenheit der geleisteten Arbeit sind in den Arbeitsergebnissen fachlich eingehend dargestellt.

Neben den Forschungsergebnissen sind durch das Projekt auch Impulse für die akademische Lehre entstanden.

7 Voraussichtlicher Nutzen, insbesondere Verwertbarkeit der Ergebnisse

Die neuen wissenschaftlichen Erkenntnisse wurden teilweise bereits mittels wissenschaftlicher Fachzeitschriften und Bücher verbreitet [RMB+23, BCG+24, CG24a] und es ist eine weitere Veröffentlichung in Vorbereitung [CG24b].

Die Minimierung der Fülldistanz des Trainingssatzes kann bei Anwendungen von großem Nutzen sein, bei denen die Datengenerierung mittels numerischer Simulationen oder Laborexperimente rechenintensiv, zeitaufwändig oder kostspielig sind. In solchen Anwendungen werden ML-Regressionsmodelle verwendet, um die unbekannten Eigenschaften schnell vorherzusagen. Ihre Genauigkeit ist jedoch stark von der Qualität der Trainingsdaten abhängig. Daher ist eine sorgfältige Auswahl der Trainingsdaten entscheidend, um genaue und robuste Vorhersagen für neue Datenpunkte zu gewährleisten. Unsere Forschung zur Minimierung der Fülldistanz der Trainingsmenge kann dazu verwendet werden, Trainingsmengen zu identifizieren, die das Potenzial haben, die Robustheit der Vorhersagen für eine breite Palette von Regressionsmodellen und Aufgaben zu verbessern. Dieser Ansatz verhindert die Verschwendung teurer Ressourcen zur Generierung von Daten, welche nur für ein bestimmtes Lernmodell oder eine bestimmte Aufgabe von Nutzen sein können.

Obwohl die Datensatzauswahlstrategie an den Daten und der Anwendung aus dem Projektkontext entwickelt und evaluiert wurde, sind die Ergebnisse allgemeiner Natur. Die praktischen Algorithmen auf Basis von FPS können ohne weiteres in anderen Domänen verwendet werden. Der Code zu [CG24a] wurde veröffentlicht und ist frei verfügbar [CG24d], bei Annahme der Veröffentlichung ist gleiches für den Code zu [CG24b] vorgesehen.

Über das MaGriDo Projekt hinaus werden die Ergebnisse und Erkenntnisse insbesondere in weiterführenden Forschungsarbeiten an der Universität Bonn verwertet. Die Arbeiten zur Datensatzauswahlstrategie fließen ein in die Arbeiten des Instituts für Numerische Simulation in Datenkompetenzzentrum Rhein-Ruhr (DKZ.2R). Eines der Aufgaben des INS im DKZ.2R ist es Schulungsmaterial zum Thema Datensatzauswahl zu erstellen.

Weitere studentische Abschlussarbeiten im Themenfeld des Projekts sind angedacht.

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

Zur Nutzung von Domänenwissen in ML-Verfahren zur Vorhersage und Design der Eigenschaften von Molekülen bzw, Kraftfelder für atomistische Systeme sind viele Aussagen im Übersichtsartikel [NTM+20] aus dem Jahr 2020 zu finden. Die Forschung an graph-basierten Neuronalen Netzen war und ist sehr aktiv, siehe beispielsweise die Übersichtartikel [RNE+22, WLB23].

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.

9 Erfolgte und geplante Veröffentlichungen der Ergebnisse

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

[RMB+23] von Rueden, L., Mayer, S, Beckh, K., ..., Garcke, J., Bauckhage, C., and Schuecker, J.. Informed machine learning - A taxonomy and survey of integrating knowledge into learning systems. IEEE Transactions on Knowledge and Data Engineering, 35(1):614–633, 2023.

[BCG+24] Breustedt, N., Climaco, P., Garcke, J., Hamaekers, J., Kutyniok, G., Lorenz, D. A., Oerder, R. & Shukla, C. V. (2024). On the Interplay of Subset Selection and Informed Graph Neural Networks. arXiv preprint arXiv:2306.10066. (Zur Veröffentlichung akzeptiert)

[CG24a] Climaco, P. and Garcke, J.. On minimizing the training set fill distance in machine learning regression. Journal of Data-centric Machine Learning Research, 2024, auch arXiv:2307.10988.

[CG24b] Climaco, P. and Garcke, J.. Density Aware Farthest Point Sampling, 2024, eingereicht.

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

[CG23a] Climaco, P. and Garcke, J. Investigating the effects of minimising the training set fill distance in machine learning regression, Workshop on High-Dimensional Approximation, HDA 2023, Canberra, Australien. Vortrag.

[CG23b] Climaco, P. and Garcke, J. Investigating the effects of minimising the training set fill distance in machine learning regression, GAMM Workshop on Computational and Mathematical Methods in Data Science (COMinDS), Köln. Vortrag.

[CG23c] Climaco, P. and Garcke, J. Investigating minimizing the training set fill distance in machine learning regression. ICML Workshop: Data-centric Machine Learning Research, 2023, Hawai, USA. Poster.

[CG23d] Climaco, P. and Garcke, J. On minimizing the training set fill distance in regression, 10th International Congress on Industrial and Applied Mathematics, ICIAM 2023, Tokyo, Japan. Vortrag.

[CG24c] Climaco, P. and Garcke, J. On minimizing the training set fill distance in machine learning regression, GAMM Jahrestagung, 2024, Magdeburg. Vortrag.

[CG24d] Climaco, P. and Garcke, J. Using Fill Distance for Dataset Selection in Machine Learning Regression, Code-Veröffentlichung. https://zenodo.org/doi/10.5281/zenodo.13374151

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

[Erl20] Erlenbach, Lukas. Active Learning for Bayesian Neural Networks with Gaussian Processes. Masterarbeit, Institut für Numerische Simulation, Universität Bonn, 2020.

[Hin20] Hines Chaves, Derrick Armando. Multi-Scale Spectral-Based Deep Graph Convolutional Networks for Molecular Property. Masterarbeit, Institut für Numerische Simulation, Universität Bonn, 2020. (Garcke: Erstbetreuer, Haemaekers: Co-Betreuer).

[Oll21] Ollech, M. Analysis of Neural Tangent Kernels. Masterarbeit, Institut für Numerische Simulation, Universität Bonn, 2021.

[Wal21] Waldorf, Konstantin: Graph-convolutional neural networks for chemical applications, Masterarbeit, Institut für Numerische Simulation, Universität Bonn, 2021. (Garcke: Erstbetreuer, Haemaekers: Co-Betreuer).

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.

[Hau23] F. Hauser. Ähnlichkeit von neuronalen Netzen. Bachelorarbeit, Institut für Numerische Simulation, Universität Bonn, 2023.

[Cli24/25] Climaco, P. On minimizing the training set fill distance in machine learning regression (vorläufiger Titel), Doktorarbeit, Institut für Numerische Simulation, Universität Bonn, in Vorbereitung zur Einreichung Ende 2024/Anfang 2025.

Literaturverzeichnis

[BHB+18] Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., ... & Gulcehre, C. (2018). Relational inductive biases, deep learning, and graph networks. *ArXiv:1806.01261*.

[CZZ+18] Chen, C., Ye, W., Zuo, Y., Zheng, C., & Ong, S. P. (2018). Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *arXiv:1812.05055v2*.

[GV14] Garcke, J. & Vanck, T.. Importance Weighted Inductive Transfer Learning for Regression. In Proceedings of ECMLPKDD 2014, 466–481, 2014.

[NTM+20] Noé, F., Tkatchenko, A., Müller, K. R., & Clementi, C.. Machine learning for molecular simulation. Annual Review of Physical Chemistry, 71, 361–390 (2020).

[PY10] Pan, S.J. and Yang, Q., 2010. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), pp.1345–1359.

[Rea:12] Rupp, Matthias, et al. "Fast and accurate modeling of molecular atomization energies with machine learning." Physical review letters 108.5 (2012): 058301.

[RNE+22] Reiser, P., Neubert, M., Eberhard, A. *et al.* Graph neural networks for materials science and chemistry. *Commun Mater* **3**, 93 (2022). https://doi.org/10.1038/s43246-022-00315-6

[SAC+17] Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R., & Tkatchenko, A. (2017). Quantum-chemical insights from deep tensor neural networks. *Nature communications*, *8*, 13890.

[SHW21] Satorras, V. G., Hoogeboom, E. & Welling, M. E(*n*) equivariant graph neural networks. In International Conference on Machine Learning 9323–9332 (2021).

[WLB23] Wang, Y., Li, Z., Barati Farimani, A. Graph Neural Networks for Molecules. In: Qu, C., Liu, H. (eds) Machine Learning in Molecular Sciences. Challenges and Advances in Computational Chemistry and Physics, vol 36. Springer (2023).

[WPC+19] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2019). A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596*.

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:	Rheinische Friedrich-Wilhelms-Universität Bonn Regina-Pacis-Weg 3, 53113 Bonn Ausführende Stelle: Institut für Numerische Simulation
Projektleitung:	Prof. Dr. Jochen Garcke
Verbund:	MaGriDo: Mathematik für maschinelle Lernmethothoden für Graph-basierte Daten mit integriertem Domänenwissen
Thema:	Teilprojekt 1: Integration von Domänenwissen, Erklärbarkeit und Transferlernen

1. Ziel und Inhalt des Projektes

In den letzten Jahren wurden Deep Learning-Methoden erfolgreich für verschiedene Problemstellungen wie Bilderkennung eingesetzt. Bisher wurden dabei meist sogenannte "end-to-end" Lernansätze verwendet. Diese erfordern in der Regel große Mengen an strukturierten Daten, wodurch sie in vielen Anwendungsfällen aus den Naturwissenschaften, der Medizin und der Industrie nur begrenzt einsetzbar sind.

Das Ziel des Verbundvorhabens MaGriDo war es daher, neue Ansätze zu entwickeln, zu analysieren und auf Problemstellungen anzuwenden, die es ermöglichen, vorhandenes Wissen in die Netzwerkarchitektur einzubauen. Dadurch können die jeweiligen Stärken von "end-to-end" Lernansätzen und "a priori Modellen/Regeln" kombiniert werden. Dieses Vorgehen verspricht effizientere Lösungen für viele Anwendungsfelder zu ermöglichen.

Die Arbeiten im Teilprojekt 1 haben sich auf die geeignete Auswahl kleiner Trainingsmengen aus großen Pools unmarkierter Datenpunkte konzentriert. Dies ist von wesentlicher Bedeutung, um die Modellleistung zu maximieren und gleichzeitig die Effizienz zu erhalten.

2. Ablauf und Ergebnisse des Vorhabens

Das Vorhaben umfasste die Zusammenarbeit von universitären Arbeitsgruppen mit Expertise in numerischer Mathematik, Analysis, Optimierung und angewandter Funktionalanalysis in Kooperation mit dem Fraunhofer SCAI. Die universitären Arbeitsgruppen hatten unterschiedliche Schwerpunkte und behandelten verschiedene Aspekte der betrachteten Lernverfahren. Das Design der Architektur diente als verbindendes Element, sei es durch die Integration von Domänenwissen und Datensatzauswahl (Bonn), die Betrachtung von Expressivität und Interpretierbarkeit (München), das Training der Modelle (Braunschweig) oder die Anforderungen aus der Anwendungsperspektive (SCAI mit den assoziierten Projektpartnern). Darüber hinaus waren die assoziierten Partner Covestro und Schott, Experten für konkrete Fragestellungen in den Anwendungsbereichen Moleküle/Polymere und Gläser, beteiligt.

Die Planung und Durchführung des Projektes gliederte sich in folgende inhaltliche Arbeitspakete, wobei der jeweilige Teilprojektkoordinator in Klammern vermerkt ist:

- Arbeitspaket AP1 (alle, je 2 PM): Demonstratoranwendung:
- Arbeitspaket AP2 (LMU): Expressivität und Interpretierbarkeit
- Arbeitspaket AP3 (TUBS): Training und Architektur
- Arbeitspaket AP4 (UB): Integration von Domänenwissen und Transferlernen
- Arbeitspaket AP5 (SCAI): Anwendung in der Materialentwicklung
- Arbeitspaket AP6 (SCAI): Praxisrelevante Anwendung

Im TP1 der Universität Bonn wurde insbesondere Farthest Point Sampling (FPS) untersucht, ein Ansatz zur Datenauswahl, der darauf abzielt, die sogenannte Fülldistanz der ausgewählten Menge zu minimieren. Auf theoretischer Seite wird eine obere Schranke für den maximal erwarteten Vorhersagefehler gezeigt, welche vor allem linear von der Fülldistanz der Trainingsmenge abhängt. Zudem wird empirisch beobachtet, dass die Auswahl eines Trainingssets mit FPS. d.h. mit dem Ziel die Fülldistanz zu minimieren, den maximalen Vorhersagefehler verschiedener Regressionsmodelle signifikant reduziert und alternative Sampling-Ansätze deutlich übertrifft. Um die durchschnittliche Qualität der Vorhersagen von Regressionsmodellen zu verbessern, wurde mit ”Density Aware Farthest Point Sampling”(DA-FPS) ein neues Datenauswahlverfahren entwickelt, welches in der Herleitung auf der neu eingeführten gewichteten Fülldistanz basiert.

3. Darstellung der wesentlichen Ergebnisse und deren konkreter Nutzen sowie ggf. die Zusammenarbeit mit anderen Forschungseinrichtungen

Die Minimierung der Fülldistanz des Trainingssatzes kann bei Anwendungen von großem Nutzen sein, bei denen die Datengenerierung mittels numerischer Simulationen oder Laborexperimente rechenintensiv, zeitaufwändig oder kostspielig sind. In solchen Anwendungen werden ML-Regressionsmodelle verwendet, um die unbekannten Eigenschaften schnell vorherzusagen. Ihre Genauigkeit ist jedoch stark von der Qualität der Trainingsdaten abhängig. Daher ist eine sorgfältige Auswahl der Trainingsdaten entscheidend, um genaue und robuste Vorhersagen für neue Datenpunkte zu gewährleisten. Unsere Forschung zur Minimierung der Fülldistanz der Trainingsmenge kann dazu verwendet werden, Trainingsmengen zu identifizieren, die das Potenzial haben, die Robustheit der Vorhersagen für eine breite Palette von Regressionsmodellen und Aufgaben zu verbessern. Darüber hinaus haben wir mit der gewichteten Fülldistanz ein weiteres und neues Konzept eingeführt um auch die Minimierung des durchschnittlichen Fehlers modellagnostisch bei der Datenauswahl zu berücksichtigen. Diese Ansätze verhindern die Verschwendung teurer Ressourcen zur Generierung von Daten, welche nur für ein bestimmtes Lernmodell oder eine bestimmte Aufgabe von Nutzen sein können.

Obwohl die Datensatzauswahlstrategie mittels FPS oder DA-FPS an den Daten und der Anwendung aus dem Projektkontext entwickelt und evaluiert wurde, sind die Ergebnisse allgemeiner Natur. Die praktischen Algorithmen auf Basis von FPS und DA-FPS können ohne weiteres in anderen Domänen verwendet werden. Der Code dazu wurde beziehungsweise wird veröffentlicht.

Über das MaGriDo Projekt hinaus werden die Ergebnisse und Erkenntnisse insbesondere in weiterführenden Forschungsarbeiten an der Universität Bonn verwertet. Die Arbeiten zur Datensatzauswahlstrategie fließen ein in die Arbeiten des Instituts für Numerische Simulation in Datenkompetenzzentrum Rhein-Ruhr (DKZ.2R). Eines der Aufgaben des INS im DKZ.2R ist es Schulungsmaterial zum Thema Datensatzauswahl zu erstellen.