Workshop Data Science

Drehbuch FS 2025

Simon Felix, Michael Graber, Martin Melchior

Funktion im Rahmen der Gesamtausbildung - Leitidee

Der Workshop Data Science bietet Studierenden zum Ende des Bachelorstudiums die Gelegenheit, in einem Themengebiet der aktuellen wissenschaftlichen Forschung im Forschungsfeld Data Science selbständig erste Schritte zu unternehmen. Dabei werden die im Studium in Modulen und Projekten erworbenen Kompetenzen im Prozess des wissenschaftlichen Erkenntnisgewinns eingesetzt und ausgebaut. Der Fokus liegt dabei besonders auf dem Austausch gesammelter Einsichten und Erkenntnisse mit Fachexperten und Mitstudierenden.

Unterrichtsorganisation

Lern- und Arbeitsformen

Studierende widmen sich im Workshop individuell und selbständig einem Thema der wissenschaftlichen Forschung aus dem Feld Data Science, welches von den Dozierenden zur Auswahl vorgegeben wird. Sie bauen sich Wissen zu Kontext, Umsetzung und Evaluierung von Methoden und Algorithmen auf und stehen dabei im regelmässigen Austausch mit einem Dozierenden, welcher die Arbeit begleitet. Die wissenschaftliche Arbeitsweise soll diskursiv geübt werden. Instrumente, die im Zuge des Semesters eingesetzt werden sollen, sind: 1. Sichtung und Aufarbeitung der relevanten wissenschaftlichen Literatur, 2. Aufbau und Formulierung einer abstrakten Darstellung des Themenbereichs, sowie Identifikation von relevanten Fragestellungen, 3. Formulieren von Hypothesen zu Wirkungszusammenhängen 4. Konzeption und Umsetzung von Experimenten, welche es ermöglichen, Wirkungsweise und Hypothesen zu untersuchen. 5. Sinnvolle, präzise und kompakte graphische Darstellung von Resultaten.

Im Semesterverlauf finden ***4 Veranstaltungen im Plenum*** statt:

Kick-Off Meeting (in der ersten Semesterwoche): Hier werden die Themen vorgestellt und auf die teilnehmenden Studierenden verteilt, sowie die Rahmenbedingungen des Workshops erklärt.

Journal Club-Session (in der 5. Semesterwoche): Studierende stellen in 5-10 Minuten kurz und frei ihr Thema und den wissenschaftlichen Kontext dazu vor. Schemata und Grafiken (keine Slides) können verwendet werden, wo dies dem einfacheren Verständnis dient. Im Anschluss besprechen wir den Inhalt gemeinsam und stellen Fragen.

3-Slides-Session (in der 9. Semesterwoche): Studierende erklären in 5-10 Minuten die für sie interessanten Fragestellungen in ihrem Themenbereich, formulieren Hypothesen und Experimente, die sie untersuchen und umsetzen möchten, sowie den Stand ihrer Arbeiten dazu. Schemata und Grafiken können gezeigt werden, wo dies ein einfacheres Verständnis ermöglicht (maximal 3 Slides). Im Anschluss besprechen wir den Inhalt gemeinsam, reflektieren die vorgeschlagenen Experimente und stellen Fragen.

Abschlusspräsentationen (in der zweitletzten Semesterwoche, in der letzten ist dieses Jahr Pfingstmontag): Die Studierenden präsentieren in 10-15 Minuten ihr Thema, die Zusammenhänge und Erkenntnisse, die sich für sie durch die Bearbeitung ihres Themas ergeben haben, sowie Hypothesen, Experimente und Resultate, die sie formuliert, umgesetzt und aufbereitet haben. Präsentationen können in Slides vorbereitet werden, sollen aber frei vorgetragen werden.

Zeitaufwand für Studierende

Plenarveranstaltungen: 12 h
Regelmässige Treffen mit den begleitenden Dozierenden: 8 h
Selbststudium: 70 h

Leistungsbeurteilung

Zur Beurteilung der Leistung der Studierenden werden folgende Kompetenzen und Kriterien ein bewertet:

Lernerfolg/Durchdringung des Themas

- Der/die Studierende kann das Thema/die Methode konzise erklären
- Er/sie versteht es, Sachverhalte, Wirkungsweisen und Zusammenhänge im Themenbereich präzise zu schildern
- Er/sie verbindet verständig Theorie und Praxis

Methodik und Vorgehen

Zum Erarbeiten eines erweiterten Verständnisses setzt der/die Studierende gezielt folgende Mittel ein:

- Literaturrecherche im Kontext des gewählten Themas
- Formulieren von Hypothesen
- Konzeption und Umsetzung von Experimenten
- Graphische Darstellung von Methoden und Resultaten

Arbeitseinsatz und -haltung

- Selbständigkeit
- (Selbst-) kritische und transparent Auseinandersetzung mit der eigenen Arbeit
- Umfang der geleisteten Arbeit

Präsentationen

- Verständlichkeit
- Präzision des Ausdrucks
- Schemata und Grafiken werden sinnvoll eingesetzt, sowie sind vollständig beschriftet
- Einhalten der Zeitvorgabe
- Aktives Mitdenken und Teilnehmen an Diskussionen zu Präsentationen von Mitstudierenden

Berechnung der Note

Die Note wird von den beteiligten Dozierenden gemeinsam auf Basis der oben genannten Kriterien bestimmt, wobei der begleitende Dozierende zu 2/3 Einfluss hat. Der/die Studierende bekommt nach der Schlusspräsentation eine Rückmeldung, welche die Schlussnote in einem Gespräch nach obigen Kriterien begründet.

Arbeitsmittel

- Literaturrecherche
- Schematische Darstellung von Methoden und Prozessen
- Umsetzung von Methoden in Code
- Grafische Darstellung von Resultaten
- Diskursive Auseinandersetzung mit dem Themenbereich zusammen mit dem Dozierenden
- Austausch mit Mitstudierenden und Fachexperten im Rahmen der Plenarveranstaltungen

Drehbuch

Agenda

Kick-Off Veranstaltung - Montag, 17. Februar 2025: 09:15 - 11:00 Uhr
Journal Club-Session - Montag, 17. März 2025: 08:15 - 11:00 Uhr
3-Slides-Session - Montag, 14. April 2025: 08:15 - 11:00 Uhr
Abschlusspräsentationen - Montag, 2. Juni 2025: 08:15 - 11:00 Uhr

Themen FS 2025

Simon Felix

Thema 1: *The History Heuristic and Alpha-Beta Search Enhancements in Practice/A New Paradigm for Minimax Search*

In einem Brettspiel kann der ideale Spielzug identifiziert werden, indem man den Spielbaum systematisch durchsucht. In den vergangenen Jahrzehnten wurden viele Varianten entwickelt, wie das möglichst effizient geschieht [1,2]. Sie sollen diese Verfahren anhand des Brettspiels “Bridget” untersuchen.

[1] Schaeffer, J., 1989. The history heuristic and alpha-beta search enhancements in practice. IEEE transactions on pattern analysis and machine intelligence, 11(11), pp.1203-1212.
[2] Plaat, A., Schaeffer, J., Pijls, W. and de Bruin, A., 1994. A New Paradigm for Minimax Search.

Thema 2: *Vertex Block Descent*

In Spielen und Filmen arbeitet hinter den Kulissen vielfach eine Physik-Simulation, damit sich Stoffe, Haare und Gegenstände realistisch bewegen. Häufig sind solche Simulationen nur in gewissen Situationen stabil und schnell. In [1] wird ein Ansatz vorgestellt, der einfach, schnell und stabil sein soll.

[1] Chen, A.H., Liu, Z., Yang, Y. and Yuksel, C., 2024. Vertex Block Descent. ACM Transactions on Graphics (TOG), 43(4), pp.1-16.

Thema 3: *Constrained Optimization*

In etlichen Anwendungsgebieten von Machine-Learning besteht die Möglichkeit, Vorwissen aus der Domäne zu nutzen. Möchte man zum Beispiel die Herstellkosten eines Artikels schätzen, liegt es auf der Hand, dass die Kosten eines lackierten Artikels höher geschätzt werden sollten, als wenn dieser Artikel nicht lackiert werden müsste. Dazu bestimmt man zusätzliche Contraints, die beim Training eines Modells eingehalten werden müssen. Bibliotheken wie TensorFlow Lattice [1], JAXopt/Optax [2] oder Cooper [3] bieten solche Möglichkeiten. Sie sollen untersuchen, welche Ansätze es gibt und wie zuverlässig diese in der Praxis funktionieren.

[1] Gupta, M., Cotter, A., Pfeifer, J., Voevodski, K., Canini, K., Mangylov, A., Moczydlowski, W. and Van Esbroeck, A., 2016. Monotonic calibrated interpolated look-up tables. Journal of Machine Learning Research, 17(109), pp.1-47.
[2] Blondel, M., Berthet, Q., Cuturi, M., Frostig, R., Hoyer, S., Llinares-López, F., Pedregosa, F. and Vert, J.P., 2022. Efficient and modular implicit differentiation. Advances in neural information processing systems, 35, pp.5230-5242.
[3] Gallego-Posada, J. and Ramirez, J., 2022. Cooper: a toolkit for Lagrangian-based constrained optimization [online]

Thema 4: *The East German encryption machine T-310 and the algorithm it used*

Die Deutsche Demokratische Republik (DDR) hat Geheimdienst-Nachrichten in den 1980er Jahren mit der eigenen Verschlüsselungsmaschine T-310 chiffriert [1]. Die eingesetzten Algorithmen sind wenig bekannt und kaum analysiert. Um die Stärke von Verschlüsselungsverfahren zu beurteilen, können SAT-Solver eingesetzt werden [2]. Sie sollen untersuchen, wie sicher das DDR-Verschlüsselungsverfahren ist.

[1] Schmeh, K., 2006. The East German encryption machine T-310 and the algorithm it used. Cryptologia, 30(3), pp.251-257.
[2] Soos, M., 2010, June. Grain of salt—an automated way to test stream ciphers through SAT solvers. In Tools (Vol. 10, pp. 131-144).

Martin Melchior

Thema 5: *Ein-Schritt Diffusion für Bildgenerierung*
In den letzten Jahren haben sich Diffusionsmodelle als State of the Art für die KI-basierte Bildgenerierung etabliert, so wie beispielsweise in DALL-E, ChatGPT, Stable Diffusion, ImageGen. Einer der Nachteile von Diffusionsmodellen (z.B. gegenüber GANs) ist, dass die Generierung langsam ist: Für ein Bild müssen iterativ rund 1000 Schritte durchlaufen werden, wobei in jedem Schritt ein komplexes neuronales Netz zur Anwendung gelangt.
Im Paper “One Step Diffusion via Shortcut Models” [2] wird eine Methode vorgestellt, welche erlauben soll, diesen Vorgang auf wenige oder gar nur einen Schritt zu reduzieren. Hier werden Sie mehr zu Diffusionsmodellen im Allgemeinen kennenlernen, und im Speziellen die im Paper vorgestellte Methode verstehen und ausprobieren.

[1] J Ho, A Jain, P Abbeel, *Denoising DIffusion Probabilistic Models*, NIPS 2020.
[2] K Frans, et al, *One Step Diffusion via Shortcut Models*, ICLR 2025.

Thema 6: *Zusätzliches Prompting für Text-to-Image Diffusionsmodelle*
Eine Herausforderung bei generativen Modellen für Bilder ist es, wie man ihnen beibringen kann, was sie generieren sollen. In modernen Chat-Programmen wie Chat-GPT wird das mit geeigneten Prompts erreicht. Bei Bildern könnte zusätzliche Flexibilität gewonnen werden, indem z.B. Skizzen vom zu generierenden Objekt übergeben werden.
Im Paper “Adding Conditional Control to Text-to-Image Diffusion Models” [1] wird das erreicht, indem ein bestehendes vortrainiertes Diffusionsmodell mittels eines “ControlNet” erweitert wird, welches die zusätzlichen Bedingungen einflechten soll. Hier werden Sie mehr zu Diffusionsmodellen im Allgemeinen kennenlernen [2], und im Speziellen die im Paper vorgestellte Methode verstehen und ausprobieren.

[1] L Zhang, A Rao, M Agrawala, *Adding Conditional Control to Text-to-Image Diffusion Models*, ICC 2023.
[2] J Ho, A Jain, P Abbeel, *Denoising DIffusion Probabilistic Models*, NIPS 2020.

Thema 7: *Convolutions und Transformer kombiniert*
Convolutional Neural Nets haben in der vorhergehenden Dekade (~2010-2010) neben den Sprachmodellen zum grossen von KI-getriebenen Modellen beigetragen. Im 2017 ist mit dem Paper “Attention is all you need” [1] die Transformer-Architektur eingeführt worden, zuerst in der Anwendung auf Sprache, dann in [2] auch auf Bilder. Damit wurde nahegelegt, dass es keine weiteren Architektur-Bausteine bedarf als Attention und Fully-Connected Layers. Eigentlich hat man damit aber auch gewisse Vorzüge von Convolutions aufgegeben.
Im Paper “CvT: Introducing Convolutions to Vision Transformers” wird eine Kombination von Transformer-Bausteine mit Convolutions vorgestellt. Hier werden Sie die vorgestellte Architektur untersuchen und mit CNN-Architekturen aber auch Transformer Architekturen vergleichen.

[1] A Vaswani et al, Attention Is All You Need, NIPS 2017.
[2] A Dosovitskiy et al, *An Image is worth 16X16 words: Transformers for Image Recognition at scale*, ICLR 2021.
[3] H Wu et al, *CvT: Introducing Convolutions to Vision Transformers*, ICCV 2021

Thema 8: *Knowledge Distillation, Self-Distillation*
Bei Knowledge Distillation wird typischerweise eine kompaktere Version eines grösseren Modells erstellt: Das Wissen wird während einem Trainingsvorgang vom grossen Modell (Teacher) zu einem kleineren Modell (Student) transferiert. Dieser Vorgang wird als Knowledge Distillation bezeichnet und ist nützlich, da die Student-Modelle häufig viel weniger Ressourcen benötigen bei gleichbleibender oder fast gleichbleibender Performance. Bei Self-Distillation, wie im Paper “Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation” [1] vorgestellt, wird schrittweise im Rahmen eines Trainings ein kleineres Modell gelernt. Hier soll diese Methode auf ihre Wirksamkeit untersucht werden - gleichzeitig auch die Prinzipien von Knowledge Distillation (Student-teacher Setting) kennengelernt werden.

[1] L Zhang et al, *Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation*, ICCV 2019.

Michael Graber

Thema 9: *KD-LoRA / PC-LoRA*
Bei der Weiterentwicklung von Deep Learning-Modellen in den vergangenen 15 Jahren waren wiederkehrend expansive Bewegungen hin zu immer grösseren Modellen zu beobachten in Abwechslung mit methodischen Fortschritten, die eine Kompression von Modellen ermöglichen. Zwei relativ kürzliche methodische Innovationen sind Knowledge Distillation und Low-Rank Adaptation (LoRA). Im letzten Jahr haben nun zwei Publikationen [1,2] eine Kombination dieser beiden Ansätze umgesetzt. Wir wollen diese Ansätze verstehen, umsetzen und hinsichtlich ihrer Eigenschaften bei Anwendung auf sinnvollen, kleinen Beispielen vergleichen.

[1] Azimi, Rambod, et al. "KD-LoRA: A Hybrid Approach to Efficient Fine-Tuning with LoRA and Knowledge Distillation." arXiv preprint arXiv:2410.20777 (2024).
[2] Hwang, Injoon, et al. "PC-LoRA: Low-Rank Adaptation for Progressive Model Compression with Knowledge Distillation." arXiv preprint arXiv:2406.09117 (2024).

Thema 11: *LLMs as Optimizers*
Die Optimierung von Zielfunktionen ist eine der zentralen Tätigkeiten der Informatik. Dabei werden die Zielfunktionen meist mathematisch formuliert und mittels verschiedener algorithmischen und numerischen Optmierungsverfahren optimiert. LLMs ermöglichen die Instruktion von Computern mittels Sprache, auch für das Lösen von Optimierungsproblemen, wie kürzlich gezeigt wurde [1]. Wir wollen diesen Ansatz explorieren, um verschiedene Optimierungsprobleme zu lösen!

[1] Yang, Chengrun, et al. "Large Language Models as Optimizers", arXiv: 2309.03409

Thema 10: *TabPFN*
Die Vorhersage mit tabellarischen Daten ist eine der letzten Bastionen klassischer Machine Learning-Ansätze, die noch nicht von Deep Learning überrollt wurden. Ein neues Paper [1] behauptet nun, eine Deep Learning-Methode entwickelt zu haben, die sämtliche bisherigen Ansätze, auch klassiche, mit grossem Abstand zu schlagen imstande sei, sofern der Datensatz nicht allzu gross ist (< 10’000 Punkte). Wir wollen diese neue Methode hinsichtlich ihres Modellierungsansatzes aufarbeiten und verstehen, um sie dann auf einer Reihe von Problemen anzuwenden. Unser Ziel ist es zu verstehen, wo die Grenzen dieses neuen Ansatzes liegen.

[1] Hollmann, Noah, et al. "Accurate predictions on small data with a tabular foundation model." Nature 637.8045 (2025): 319-326.

Thema 12: *Prompting Strategies: Chain-of-Thought, ReACT, ..*
Large Language Models (LLMs) können komplexere Probleme lösen, wenn sie via Prompt systematisch instruiert werden. In den letzten Jahren wurden verschiedene Prompting Strategien entwickelt, ja Prompt Engineering wurde zu einer ernstzunehmenden Tätigkeit. Hier wollen wir uns am Beispiel von Chain-of-thought und React ins Gebiet Prompt Engineering und Prompt Optimization einarbeiten, uns eine Übersicht über SOTA-Ansätze verschaffen und diese auf Beispielproblemen miteinander konkret vergleichen.

[1] Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." Advances in neural information processing systems 35 (2022): 24824-24837.
[2] Yao, Shunyu, et al. "React: Synergizing reasoning and acting in language models." arXiv preprint arXiv:2210.03629 (2022).