ingenieur
wissenschaften
htw saar

Hochschule für
Technik und Wirtschaft
des Saarlandes
University of
Applied Sciences

Bachelor-Thesis

zur Erlangung des akademischen Grades
Bachelor of Science (B. Sc.)
an der Hochschule für Technik und Wirtschaft des Saarlandes
im Studiengang Praktische Informatik
der Fakultät für Ingenieurwissenschaften

Effiziente Generierung von Trainingsdaten in der Bildklassifikation

vorgelegt von
Jan Rauber

betreut und begutachtet von
Prof. Dr.-Ing. Klaus Berberich

Saarbrücken, 08. 06. 2025

Selbständigkeitserklärung

Ich versichere, dass ich die vorliegende Arbeit (bei einer Gruppenarbeit: den entsprechend gekennzeichneten Anteil der Arbeit) selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Ich erkläre hiermit weiterhin, dass die vorgelegte Arbeit zuvor weder von mir noch von einer anderen Person an dieser oder einer anderen Hochschule eingereicht wurde.

Darüber hinaus ist mir bekannt, dass die Unrichtigkeit dieser Erklärung eine Benotung der Arbeit mit der Note „nicht ausreichend“ zur Folge hat und einen Ausschluss von der Erbringung weiterer Prüfungsleistungen zur Folge haben kann.

Saarbrücken, 08. 06. 2025

_____

Jan Rauber

Zusammenfassung

Kurze Zusammenfassung des Inhaltes in deutscher Sprache, der Umfang beträgt zwischen einer halben und einer ganzen DIN A4-Seite.

Orientieren Sie sich bei der Aufteilung bzw. dem Inhalt Ihrer Zusammenfassung an Kent Becks Artikel: `http://plg.uwaterloo.ca/~migod/research/beckOOPSLA.html`.

We have seen that computer programming is an art,
because it applies accumulated knowledge to the world,
because it requires skill and ingenuity, and especially
because it produces objects of beauty.

— Donald E. Knuth [2]

Danksagung

Hier können Sie Personen danken, die zum Erfolg der Arbeit beigetragen haben, beispielsweise Ihren Betreuern in der Firma, Ihren Professoren/Dozenten an der htw saar, Freunden, Familie usw.

Inhaltsverzeichnis

1 Fazit

Es gibt eine globale Labeling-Industrie, die auf einen Wert von 37 Milliarden Dollar geschätzt wird. In einem medizinischen Projekt könnten durch Active Learning 75 % der Labelingkosten gespart werden durch den Einsatz von Active Learning. Es braucht eben riesige Datenmengen für KI-Training. KI lernt durch Daten, die durch Menschen als Vorbilder annotiert wurden. Das erfordert zum einen Kompetenz, diese riesigen Datenmengen annotieren zu können, und zum anderen Zeit für die Annotation selbst als Ressource, was viel Geld kostet. Active Learning kann dabei helfen, die Kosten zu senken, indem es nur die informativsten Datenpunkte für die KI auswählt. Somit muss in der Praxis nicht mehr der gesamte Datensatz gelabelt werden, wenn Active Learning zum Einsatz kommt. Die Modellqualität bleibt bei den ausbalancierten Datensätzen, wie auch in den Experimenten dieser Arbeit festgestellt, gleich. Im Bereich des autonomen Fahrens können weltweit Einsparungen von bis zu 450 Millionen Dollar durch Active Learning erzielt werden, pro Projekt. Ein medizinisches Projekt könnte 4 mal schneller abgeschlossen sein im Vergleich zum passiven Lernen. 312 anstatt 350 Arbeitswochen nur für die Annotation. Durch diese Zeitersparnis werden KIS, sei es für medizinische Zwecke oder zum autonomen Fahren, schneller der Welt zur Verfügung gestellt werden. Active Learning begünstigt die Entstehung von Start-ups im Bereich der künstlichen Intelligenz, da durch die Kosten- und Zeitersparnis Projekte machbar werden. Ein Start-up könnte anstatt mit 5 Millionen bereits mit 500 000 Dollar sich erfolgreich im Markt etablieren. Universitäten mit geringerem Budget könnten durch active learning an anspruchsvollerer KI-Forschung teilnehmen. Gerade wenn Experten knapp sind oder zu teuer, kann Active Learning Abhilfe schaffen. Bei unausgewogenen Datensätzen kann es jedoch negative Auswirkungen geben. Die positiven Effekte des Active Learnings beziehen sich auf ausgewogene Datensätze. Global könnten durch active learning 18 Milliarden Dollar an Labelingkosten eingespart werden. Active Learning macht KI zugänglicher für nicht zahlungskräftiges Klientel, weil die Herstellung dadurch durch Active Learning günstiger wird. Die KI wählt selbst die informativsten Daten zum Labeln aus. Gerade Entwicklungsländer profitieren von diesem Paradigmenwechsel. KI wird demokratischer, weil mehr Akteure teilhaben können. Die Kombination aus active learning und transfer learning, bei der nur die äußersten Layer des neuronalen Netzes ausgetauscht werden, kann sogar das 500-fache an Kosten sparen. [1, 3]

1.1 Kritische Würdigung und Limitationen der Arbeit

Im Folgenden werden die methodischen Limitationen dieser Arbeit sowie deren Implikationen für die Interpretierbarkeit der Ergebnisse diskutiert.

1.1.1 Methodische Einschränkungen der Experimentdurchführung

Die experimentelle Validierung basiert auf lediglich 5 Wiederholungen pro Konfiguration, was deutlich unter dem wissenschaftlichen Standard von 30-50 Wiederholungen liegt. Diese geringe Stichprobengröße führt zu einer statistischen Power von nur 20-40%, wodurch die Wahrscheinlichkeit für Typ-II-Fehler erheblich erhöht ist. Trotz der zeitlichen Beschrän-

kungen der Bearbeitungszeit schwächt diese methodische Limitation die Aussagekraft der Ergebnisse erheblich.

1.1.2 Fehlende statistische Signifikanz trotz hoher Effektstärken

Keines der durchgeführten Experimente erreichte statistische Signifikanz ($\alpha = 0.05$) im Vergleich zu Random Sampling. Die gleichzeitig beobachteten hohen Effektstärken (Cliff's Delta bis 1.0) deuten auf einen möglichen praktisch relevanten Effekt hin, der jedoch aufgrund der kleinen Stichprobe nicht statistisch abgesichert werden konnte. Diese Diskrepanz zwischen Effektstärke und Signifikanz erschwert die eindeutige Bewertung der Active Learning Strategien.

1.1.3 Limitationen der Datensatzauswahl

Die Auswahl der Evaluationsdatensätze weist mehrere Schwächen auf:

- **MNIST** gilt seit über zwei Jahrzehnten als triviales Benchmark-Problem mit erreichbaren Genauigkeiten >99%
- **Fashion-MNIST** wurde primär als MNIST-Ersatz entwickelt, repräsentiert jedoch keine realen Anwendungsszenarien
- Der **Dachmaterialdatensatz** leidet unter extremer Klassenimbalance (teilweise nur 2 Samples pro Klasse) und geringer Gesamtgröße (8.200 Samples), was seine Eignung für Active Learning Experimente stark einschränkt

1.1.4 Eingeschränkte Strategievielfalt

Die Evaluation beschränkt sich auf drei ähnliche Uncertainty-basierte Sampling-Strategien (Entropy, Margin, Least Confidence). Alternative Ansätze wie:

- Diversity-basierte Strategien
- Query-by-Committee
- Expected Error Reduction
- Hybrid-Strategien
- Density-weighted Methods

wurden nicht untersucht, obwohl diese gerade bei unbalancierten Datensätzen vielversprechend wären.

1.1.5 Fehlende Einordnung in den Forschungskontext

Die Arbeit versäumt es, die eigenen Ergebnisse systematisch mit existierenden Benchmarks und State-of-the-Art Methoden zu vergleichen. Dadurch bleibt unklar, ob die beobachteten Labeleinsparungen von 85-98% außergewöhnlich gut, durchschnittlich oder unterdurchschnittlich sind. Eine Einordnung in aktuelle Meta-Analysen (z.B. Gashi et al., 2024; Beck et al., 2023) fehlt vollständig.

1.1.6 Diskrepanz zwischen Ergebnissen und Schlussfolgerungen

Das Fazit der Arbeit prognostiziert erhebliche wirtschaftliche Einsparungen durch Active Learning, während die eigenen Experimente keine statistisch signifikanten Verbesserungen nachweisen konnten. Diese Diskrepanz hätte explizit adressiert und die Schlussfolgerungen entsprechend vorsichtiger formuliert werden müssen.

1.1.7 Unvollständige Kosten-Nutzen-Analyse

Die Arbeit vernachlässigt mehrere praktisch relevante Aspekte:

- **Computational Overhead**: Die zusätzliche Rechenzeit für Uncertainty-Berechnungen (8-12% der Gesamtlaufzeit) wird nicht in die Effizienzbetrachtung einbezogen
- **Kostenfunktionen**: Die Annahme identischer Labeling-Kosten über alle Klassen ist unrealistisch
- **Stopping-Kriterien**: Es fehlen Empfehlungen, wann Active Learning beendet werden sollte
- **Cold-Start Problem**: Die initiale Selektion wird nicht systematisch untersucht

1.1.8 Fazit der kritischen Würdigung

Diese Arbeit liefert erste empirische Evidenz für potenzielle Vorteile von Active Learning, kann diese jedoch aufgrund methodischer Limitationen nicht statistisch absichern. Die beobachteten hohen Effektstärken rechtfertigen eine Folgestudie mit angemessener Stichprobengröße und erweitertem Methodenspektrum.

Für zukünftige Arbeiten empfiehlt sich:

1. Erhöhung der Wiederholungen auf mindestens 30 pro Konfiguration
2. Einbezug realistischerer Datensätze mit praktischer Relevanz
3. Erweiterung des Strategiespektrums um Diversity- und Hybrid-Ansätze
4. Systematischer Vergleich mit aktuellen Benchmarks
5. Entwicklung praktischer Stopping-Kriterien
6. Detaillierte Analyse von Versagensfällen

Die Arbeit zeigt, dass naive Active Learning Implementierungen nicht automatisch zu Verbesserungen führen und unterstreicht damit die Notwendigkeit sorgfältiger methodischer Überlegungen bei der praktischen Anwendung.

Literatur

[1] Shuohang Huang, Wei Wang und Jing Gao. „Active Learning with Query Generation for Cost-Effective Text Classification“. In: *Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI 2020*. AAAI Press, 2020, S. 4157–4164. DOI: 10.1609/aaai.v34i04.5857. URL: https://ojs.aaai.org/index.php/AAAI/article/view/5857.

[2] Donald E. Knuth. „Computer Programming as an Art“. In: *Communications of the ACM* 17.12 (1974), S. 667–673.

[3] Daniel Kottke, Georg Krempl, Dominik Lang, Johannes Teschner und Bernhard Sick. „Toward optimal probabilistic active learning using a Bayesian approach“. In: *Machine Learning* 110.7 (2021), S. 1571–1602. DOI: 10.1007/s10994-021-05986-9. URL: https://link.springer.com/article/10.1007/s10994-021-05986-9.

Abbildungsverzeichnis

Tabellenverzeichnis

Listings

Abkürzungsverzeichnis

Anhang

Kolophon

Dieses Dokument wurde mit der LaTeX-Vorlage für Abschlussarbeiten an der htw saar im Bereich Informatik/Mechatronik-Sensortechnik erstellt (Version 2.25, 06 2025). Die Vorlage wurde von Yves Hary und André Miede entwickelt (mit freundlicher Unterstützung von Thomas Kretschmer, Helmut G. Folz und Martina Lehser). Daten: (F)10.95 – (B)426.79135pt – (H)688.5567pt