AI Chart Summarizer

p5 - HS 2023

Luca Mazzotta, Florin Barbisch

Fachhochschule Nordwestschweiz FHNW
Institut für Interaktive Technologien
Bahnhofstrasse 6
5210 Windisch

Contents

1 Einleitung

In einer Zeit, die von Daten dominiert wird, sind Diagramme unverzichtbare Werkzeuge geworden, um komplexe Informationen visuell darzustellen. Sie ermöglichen es, leicht Trends zu identifizieren, Beziehungen zu analysieren und Schlussfolgerungen zu ziehen. Diese visuellen Darstellungen stellen jedoch insbesondere für Laien ohne statistische Vorbildung eine Herausforderung dar. Vor diesem Hintergrund verfolgt unser Projekt das Ziel, ein Tool zu entwickeln, das fähig ist, aus Diagrammen automatisierte Zusammenfassungen zu generieren.

Die Herausforderung bei der Interpretation von Diagrammen inspirierte uns zu unserem Forschungsvorhaben. Wir erkennen ein Defizit an Lösungen. Unser Forschungsziel ist es daher, ein Tool zu entwickeln, das Diagramme analysiert und ihre Kerninhalte in Textform wiedergibt. Ein weiteres Ziel ist die Untersuchung des Einflusses verschiedener Trainingsparameter auf die Leistungsfähigkeit unseres Modells.

Zur Realisierung dieses Vorhabens haben wir einen speziellen Datensatz zusammengestellt, der verschiedene Arten von Diagrammen umfasst. Dieser Datensatz dient als Grundlage für das Training unseres AI Chart Summarizer-Tools. Die ersten Ergebnisse unseres Projekts deuten darauf hin, dass unser Tool in der Lage ist, Zusammenfassungen zu liefern, was die Zugänglichkeit und das Verständnis komplexer Daten für Laien verbessern könnte.

Der Bericht gliedert sich in mehrere Schlüsselsektionen. Nach einer Einführung in die Grundlagen und die Auswahl der Large Language Models (LLM's), beschreiben wir die Herausforderungen bei der Zusammenfassung von Grafiken durch LLM's. Es folgt eine detaillierte Beschreibung unseres Ansatzes für das Fine-Tuning des Modells und die Zusammenstellung unseres Datensatzes. In den Abschnitten Ergebnisse und Diskussion evaluieren wir die Leistung des entwickelten Modells und diskutieren Stärken, Herausforderungen und Grenzen unserer Lösung. Abschliessend bietet das Fazit einen Überblick über unsere Erkenntnisse und Empfehlungen.

Mit dem Ziel, eine benutzerfreundliche Schnittstelle zu schaffen, die es ermöglicht, die in Daten verborgenen Informationen für Laien zu enthüllen, tragen wir dazu bei, dass jeder, unabhängig vom Fachwissen, datengestützte Entscheidungen treffen kann. Dieser Bericht dokumentiert unseren Forschungs- und Entwicklungsprozess und bietet Einblicke in die Potenziale und Herausforderungen, die sich bei der Realisierung eines solchen Tools ergeben.

2 Grundlagen

2.1 Modelle

Die sorgfältige Auswahl und Analyse von LLM's bildet das Fundament dieses Projektes. Im Folgenden werden verschiedene fortschrittliche LLM's aufgezählt und deren Eigenschaften erwähnt.

MatCha-ChartQA: Dieses Modell wurde speziell auf dem ChartQA-Datensatz trainiert, um Fragen basierend auf Diagrammen zu beantworten. Es basiert auf Pix2Struct und konzentriert sich auf Aufgaben, die die Zerlegung von Diagrammen und numerisches Verständnis beinhalten [1]. MatCha-ChartQA wurde auf Fragen und Antworten trainiert, aber nicht auf die Generierung von Beschreibungen [2].

ChatGPT: Als ein geschlossenes Modell gilt es als eines der besten verfügbaren LLMs. Ursprünglich in seiner vierten Generation (GPT-4) ohne visuelle Fähigkeiten gestartet, zeigte es dennoch eine beeindruckende Fähigkeit, Bilder zu beschreiben [3]. Es ist bemerkenswert, dass ChatGPT in deutscher Sprache schlechtere Ergebnisse lieferte als in Englisch [4].

DePlot: Dieses Modell erstellt zunächst eine linarisierte Tabelle in Textform der Daten aus einem Diagramm und verwendet anschliessend ein LLM (z.B. GPT3), um eine Beschreibung des Diagramms zu generieren. Das verwendete LLM ist ebenfalls geschlossen. [5]

BLIP-2: BLIP-2 nutzt einen lightweight Querying Transformer, der in zwei Stufen vor-trainiert wird. In der ersten Stufe wird die visuelle Sprachdarstellung von einem eingefrorenen Image-Encoder bootstrapped. In der zweiten Stufe wird das generative Lernen von vision-to-language anhand eines eingefrorenen Sprachmodells bootstrapped. [6]

MiniGPT4: MiniGPT-4 besteht aus einem Vision-Encoder mit einem vortrainierten ViT und Q-Former, sowie einem fortgeschrittenen Vicuna-Sprachmodell. Die beiden Modell werden durch nun einen linearen Projektionslayer miteinander verbunden. [7]

Fuyu8B: Fuyu8B ist ein Transformer-basiertes Modell, dass das Bild in Patches aufteilt und diese über eine lineare Projektionsschicht direkt in die erste Transformer-Schicht projiziert. [8]

LLaVA: LLaVA ist ähnlichen aufgebaut wie Fuyu8B, besitzt aber mit 13 Milliarden mehr Parameter als Fuyu8B, welches 8 Milliarden Parameter hat. [9]

MECDG: Chen und Zaoh [10] schlagen "Manufacturing Enterprise Chart Description Generation" (MECDG) vor, dass 1) den Text mit OCR und Keypoints extrahiert und 2) mit einem Language Modell Antworten auf User-Input generieren. Allerdings hat das Modell gewisse Redundanzen, da die einzelnen Komponenten unabhängig voneineander sind.

Chart-Text: Abhijit et. al haben Chart-Text entwickelt, um Alt-Texte für Charts zu generieren. Sie verwenden Bilder-Klassifizierungsmodelle und Objekterkennungsmodelle, welche sie mit einem selbst erstellten Datensatz trainiert haben. Sie haben eine Chart-Klassifikationsgenauigkeit von 99.72% erreicht und eine Text-Generationsgenauigkeit von 78.9%. Dieses System setzt

den Fokus allerdings auf blinde Personen. [11]

Jedes dieser Modelle bietet einzigartige Ansätze zur Diagrammanalyse und -interpretation, wobei ihre Leistungsfähigkeit und Genauigkeit je nach Anwendungsbereich variieren. Im weiteren Verlauf werden wir uns genauer mit dem Modell Fuyu8b auseinandersetzen, weil 1) die Transformer-Architektur mit einem Projektionslayer in drei der hier vorgestellten Modelle vorkommt (LLaVA und MiniGPT4 benutzen auch diese Architektur) und 2) wir das Gefühl hatten, dass das finetuning von Fuyu8B am einfachsten sein wird aufgrund der Anzahl Parameter und der Architektur.

2.2 Herausforderungen aktueller LLMs

Die Herausforderungen bei der Zusammenfassung und Beschreibung von Grafiken durch LLM's sind vielfältig und komplex. Zu den Hauptproblemen gehören:

1. **Entwicklung robuster Modelle:** Aufgrund begrenzter Datensätze und Diagrammtypen ist es schwierig, robuste Modelle zu entwickeln. Die Datensatzerstellung ist besonders herausfordernd, da Grafiken und ihre Beschreibungen oft nicht direkt miteinander verknüpft sind, was die Trainingsgrundlage für diese Modelle schwächt. [12]

2. **Begrenzung des Kontextfensters bei LLM's:** Diese Einschränkung beeinträchtigt die Kapazität der Modelle für In-Context-Lernen und stellt somit eine signifikante Hürde dar. [13]

3. **Problematik der Halluzination:** LLM's neigen dazu, falsche oder inkonsistente Aussagen zu generieren, was auf einen Mangel an echtem Wissen hindeutet. Dieses Phänomen tritt nicht bei allen Modellen und Tasks gleich stark auf. Verstärkt tritt es auf, wenn Wissen von visuellen zu sprachlichen Formaten übertragen wird, was die Zuverlässigkeit der generierten Zusammenfassungen und Beschreibungen beeinträchtigt. [14], [15]

4. **Fehlende Fähigkeit zur effektiven Nutzung visueller Informationen:** Die überwiegend textbasierte Trainingsgrundlage der LLM's schränkt ihre Fähigkeit erheblich ein, grafische Informationen zu verarbeiten und zu interpretieren. [16]

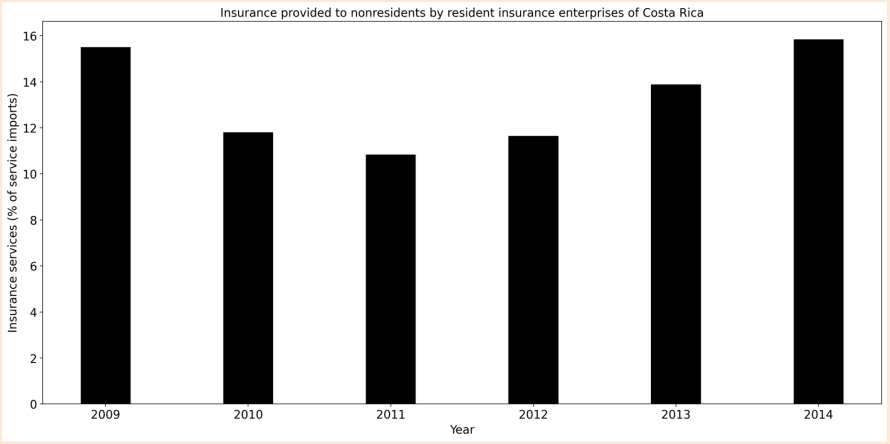
2.3 Datensätze

Um Modelle zu trainieren braucht man Datensätze mit verschiedenen Beispielen, damit das Modell davon lernen kann. Es gibt verschiedene Datensätze, die auf unterschiedliche Arten und Weisen erstellt worden sind.

2.3.1 Autochart

AutoChart ist ein spezialisierter Datensatz für die Chart-to-Text-Generierung, eine Aufgabe, die darauf abzielt, analytische Beschreibungen von visuellen Plots zu generieren. Der Datensatz wurde von Zhu et al. [17] entwickelt und zielt darauf ab, die Forschung im Bereich der analytischen Beschreibung von Charts zu fördern. AutoChart umfasst vier Arten von Diagrammen: Streudiagramme, Liniendiagramme, vertikale und horizontale Balkendiagramme, die mit der Python-Bibliothek Matplotlib erstellt wurden.

Die Beschreibungen in AutoChart wurden durch eine Kombination aus menschlich geschriebenen Vorlagen und paraphrasierten Sätzen erstellt, um eine Vielzahl rhetorischer Bewegungen zu berücksichtigen, die für die Analyse von Diagrammen wesentlich sind. Diese umfassen eine Übersicht über das Diagramm, Beschreibungen der Diagrammkomponenten, Interpretationen der dargestellten Informationen, bewertende Kommentare und Schlussfolgerungen.



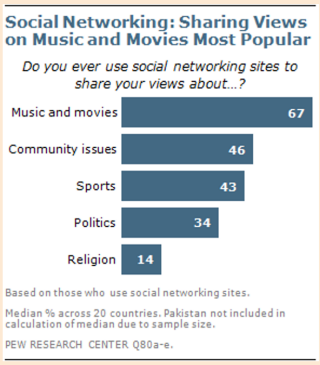
Beschreibung:

The bar chart illustrates the number of insurance provided to nonresidents by resident insurance enterprises of costa rica from 2009 to 2014. Each bar represents the number in the unit of Insurance services (% of service imports). There is a unstable trend of the number. The number in 2014 being the peak, and the lowest number is found in 2011. Number changes may be linked to national policies of the country.

Figure 1: Beispielbild aus dem Autochart-Datensatz.

2.3.2 Chart-to-text (Pew und Statista)

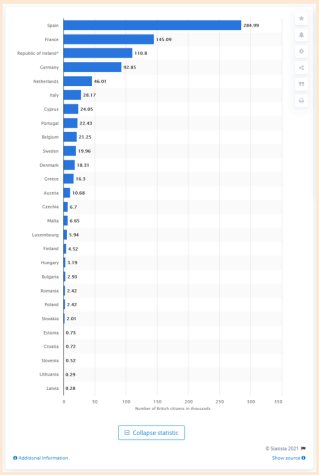
Die Chart-to-text-Datensätze von Pew und Statista bieten eine breite Palette von Diagrammtypen, darunter Balken-, Linien- und Tortendiagramme. Besonders auffällig ist, dass diese Diagramme umfangreiche Textinformationen innerhalb des Diagramms selbst enthalten, was zu detaillierteren und komplexeren Beschreibungen führt. Der Pew-Datensatz neigt zu längeren Beschreibungen mit spezifischen Schlussfolgerungen und detaillierten Berechnungen, während der Statista-Datensatz qualitativ hochwertige Charts jedoch ohne Titel bietet, was den Kontext des Charts etwas schwieriger zu erfassen macht. Die Beschreibungen in beiden Datensätzen bieten jedoch einen umfassenden Kontext darüber, was in den Charts dargestellt wird. [18]



Beschreibung:

A survey of 20 countries conducted last spring by the Pew Research Center's Global Attitudes Project found a median of 67% of social networkers say they use these sites to share opinions about music and movies. Significant numbers also post their views on community issues, sports and politics

Figure 2: Beispielbild aus dem Chart-to-text (Pew) -Datensatz



Beschreibung:

In 2018 , there were approximately 285 thousand British citizens living in Spain , with a further 145 thousand in France and almost 93 thousand in Germany . By comparison , there were only 283 British people living in Latvia , the fewest of any European Union member state . As a member of the European Union , British citizens currently have the right to live and work in any EU member state , although it is unclear what their legal status will be after Brexit is fulfilled and the UK leaves the European Union.

Figure 3: Beispielbild aus dem Chart-to-text (Statista) -Datensatz

2.3.3 ChartSumm

ChartSumm umfasst ebenfalls verschiedene Chart-Typen, wobei Linien-, Balken- und Tortendiagramme dominieren. Die Diagramme in ChartSumm wirken eher schlicht. Die Beschreibungen sind auch eher schlicht gehalten und beschränken sich auf die Beschreibung der Grafik [19].

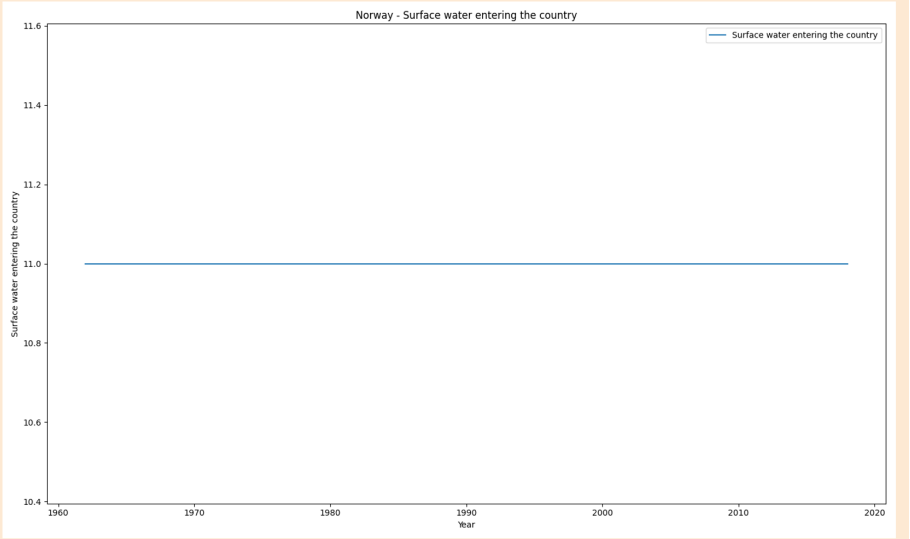


Figure 4: Beispielbild aus dem Chartsumm-Datensatz

2.4 Metriken

2.4.1 BLEU

Der BLEU Score ist eine Methode zur Bewertung von maschinell generiertem Text im Vergleich zu menschlich erstellten Texten. Er wurde entwickelt, um die Qualität von maschinell generierten Übersetzungen zu messen. Der BLEU Score misst die Übereinstimmung zwischen einem generierten Text und dessen Referenztexten, basierend auf der Anzahl übereinstimmender N-Gramme. Ein BLEU Score von 1 (oder 100%) stellt eine perfekte Übereinstimmung dar.

In dieser Arbeit wird der BLEU Score verwendet, um die Qualität der generierten Beschreibungen zu bewerten. Die Verwendung von BLEU ist hier vorteilhaft, da es eine objektive und quantifizierbare Methode zur Beurteilung der Nähe der Zusammenfassungen zu Referenztexten bietet. Trotz seiner Einschränkungen, wie der Konzentration auf N-Gramm-Übereinstimmungen, die den Kontext der Texte weniger berücksichtigen, ist der BLEU Score ein anerkannter Standard in der Bewertung maschinengenerierter Texte.

Ein hoher BLEU Score bedeutet jedoch nicht unbedingt eine hohe Kontextähnlichkeit. Generell gilt: Je höher der BLEU Score, desto grösser ist die

Übereinstimmung des Textes mit den Referenztexten in Bezug auf die verwendeten Wortgruppen. Ein guter BLEU Score deutet darauf hin, dass der generierte Text viele Wortgruppen enthält, die auch in den Referenztexten vorkommen, was oft, aber nicht immer, auf eine höhere Qualität schliessen lässt. [20]

2.4.2 ROUGE-1

ROUGE-1 ist eine spezifische Metrik innerhalb der ROUGE-Familie (Recall-Oriented Understudy for Gisting Evaluation) und wie der BLEU Score auch eine Metrik für den Vergleich von maschinell generierten Texten und von Menschen erstellten Texten. Diese Metrik bewertet jedoch die Qualität von Textzusammenfassungen, indem sie die Anzahl der gemeinsamen Einzelworte (Unigramme) in der generierten Zusammenfassung und den Referenztexten misst.

ROUGE-1 berechnet zwei Hauptkomponenten: Recall und Precision. Recall misst, welcher Anteil der Wörter aus den Referenztexten in der generierten Zusammenfassung vorkommt. Precision hingegen misst, welcher Anteil der Wörter in der generierten Zusammenfassung auch in den Referenztexten enthalten ist.

Diese beiden Werte können dann zu einem F-Measure kombiniert werden, das ein harmonisches Mittel von Recall und Precision darstellt, um eine ausgewogene Bewertung der Textzusammenfassungen zu bieten.

Wie beim BLEU Score variiert der Rouge-1 Score zwischen 0 und 1, wobei 0 die geringste und 1 die höchste Übereinstimmung zwischen der generierten Zusammenfassung und den Referenztexten darstellt.

Während der BLEU Score hauptsächlich sich auf die Übereinstimmung von N-Grammen (Wortgruppen) konzentriert, misst ROUGE-1 spezifisch die Übereinstimmung auf Unigramm-Ebene. Während BLEU eine starke Betonung auf Precision legt und Kontext und Satzstruktur weniger berücksichtigt, ermöglicht ROUGE-1 eine ausgewogenere Bewertung durch die Kombination von Recall und Precision, die besonders für die Bewertung der Inhaltsabdeckung in Zusammenfassungen relevant ist. [21]

2.4.3 WER

WER, kurz für Word Error Rate, ist ebenfalls eine weitere Metrik für einen solchen Vergleich. Sie misst die Qualität, indem sie die Anzahl der Fehler – das heisst der notwendigen Änderungen (Einfügungen, Löschungen und Substitutionen) – berechnet, um von der maschinell generierten Ausgabe zu einem Referenztext zu gelangen. Diese Änderungen werden dann durch die Gesamtanzahl der Wörter der Referenz geteilt, um den WER Score zu erhalten.

Ein niedriger WER weist auf eine höhere Genauigkeit des Systems hin, da weniger Korrekturen erforderlich sind, um Übereinstimmung mit dem Referenztext zu erreichen. Diese Metrik ist besonders nützlich, um die Fehlerquote in den generierten Texten zu quantifizieren.

Im Gegensatz zu BLEU und ROUGE, die auf der Übereinstimmung von

Wortgruppen oder Einzelwörtern basieren, konzentriert sich WER auf die direkte Fehleranalyse.

2.5 LoRA

LoRA, oder Low-Rank Adaptation, ist eine fortschrittliche Technik zur Modifikation grosser vortrainierter Modelle. Der Kerngedanke hinter LoRA ist, die Anpassungsfähigkeit grosser Modelle zu verbessern, ohne ihre umfangreiche Struktur und Parametrisierung vollständig neu trainieren zu müssen und somit die Anzahl der Parameter deutlich zu verringern, was Rechenzeit und Speicher spart. Dies wird erreicht, indem Veränderungen auf eine niedrigrangige Matrix beschränkt werden, die auf die Gewichte bestimmter Schichten des Modells, wie die Aufmerksamkeits- und Feedforward-Netzwerkschichten, angewendet wird. Die niedrigrangige Matrix wirkt dabei als eine Art "Update-Filter", der es ermöglicht, die ursprünglichen Gewichte des Modells fein abzustimmen, ohne sie direkt zu verändern.

Technisch gesehen werden bei der LoRA-Methode zusätzliche Matrixfaktoren zu den Gewichtsmatrizen des Modells hinzugefügt. Diese Faktoren sind von niedrigem Rang, was bedeutet, dass sie weniger Parameter enthalten als die ursprünglichen Gewichtsmatrizen. Durch die Anpassung dieser Faktoren statt der gesamten Gewichtsmatrix kann das Modell effizienter und zielgerichteter für spezifische Aufgaben optimiert werden.

In dieser Arbeitet hilft LoRA die LLM's so anzupassen, um sie effizient trainieren zu können. Diese Anpassung erfolgt mit einem Bruchteil der Rechenressourcen, die normalerweise für das vollständige Training oder die Feinabstimmung eines solchen Modells erforderlich wären. Dadurch wird eine hohe Anpassungsfähigkeit erreicht, ohne die umfassenden Kenntnisse zu beeinträchtigen, die das Modell während seiner ursprünglichen Trainingsphase bereits gelernt hat. [22]

2.6 Fuyu8B

Fuyu-8B stellt eine abgespeckte Version eines multimodalen Modells dar, das von der Adept-Plattform entwickelt wurde, und bietet einige bemerkenswerte Merkmale.

Die Architektur des Fuyu-8B-Modells unterscheidet sich von anderen multimodalen Modellen, da es keinen speziellen Bildencoder besitzt. Stattdessen werden Bildausschnitte direkt in die erste Schicht des Transformer-Modells projiziert, wodurch die Notwendigkeit eines separaten Bildencoders entfällt. Dies ermöglicht dem Modell die Verarbeitung von Bildern in beliebigen Auflösungen und grössen, was besonders wichtig ist, um verschiedenste Bildarten zu verstehen.

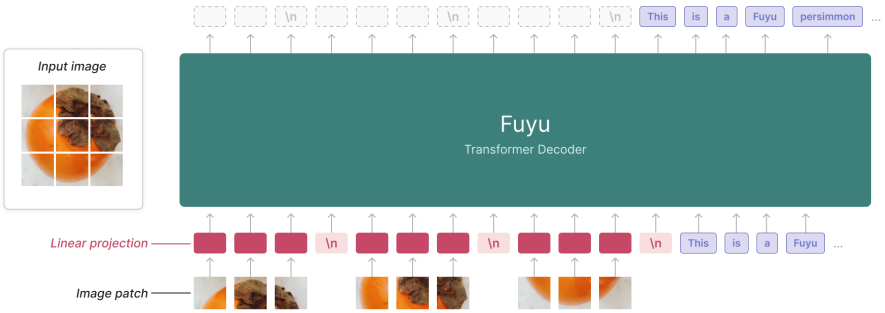


Figure 5: Architektur vom Fuyu-8B-Modell

Die Leistungsüberprüfung des Fuyu-8B-Modells ergab, dass es sich gut auf verschiedenen Bildverstehensdatensätzen schlägt, darunter VQAv2, OKVQA, COCO Captions und AI2D. Obwohl das Modell in erster Linie für den Einsatz in digitalen Agenten entwickelt wurde, erzielte es gute Ergebnisse in diesen Benchmarks, auch ohne spezifische Optimierungen für diese Aufgaben.

Die Autoren weisen jedoch darauf hin, dass die Benchmark-Datensätze einige Herausforderungen aufweisen, insbesondere in Bezug auf die Bewertung von Antworten. Die Frage-Antwort-Datensätze verwenden oft komplizierte Bewertungsmethoden und verlangen spezifische Antwortformate, was zu falschen Bewertungen für korrekte Antworten führen kann.

Die Autoren erwähnen auch, dass ihre internen Modelle, die auf Fuyu basieren, zusätzliche Fähigkeiten für OCR, detaillierte Lokalisierung von Text und UI-Elementen sowie das Beantworten von Fragen zu UI-Bildern bieten. Diese Entwicklungen könnten in zukünftigen Produktversionen integriert werden.

Insgesamt ist das Fuyu-8B-Modell eine vielversprechende Open-Source-Ressource, die Bild- und Textdaten verarbeiten kann. Es bietet eine einfache Architektur und hat sich in verschiedenen Tests bewährt, während es weiterhin Raum für zukünftige Entwicklungen und Verbesserungen bietet. [8]

3 Methodik

Für das Trainieren des Fuyu8b Modells haben wurde LoRA verwendet. Wie im Kapitel 2.5 bereits erwähnt, bietet LoRA durch die Einführung von Low-Rank-Matrizen in bestimmte Layers einen effizienten Ansatz für das Trainieren solcher Modelle. Diese Methode ermöglicht es, das Modell signifikant zu modifizieren, ohne ein umfangreiches Training durchzuführen.

Das Finetuning wurde mit einer Nvidia A100 GPU mit einer Speicherkapazität von 40 GB durchgeführt. Die A100 ist ideal für diesen Zweck aufgrund des hohen Speichers, der für solche grossen Modelle und Bilder Datenmengen, unabdingbar ist. Die Auswertung wurde dann auf einer NVIDIA RTX 4090 mit 24GB Speicher durchgeführt, da bei der Inference deutlich weniger Speicher gebraucht wird.

Um PyTorch-bezogene Out of Memory (OOM)-Errors zu vermeiden, wurde eine spezifische Einschränkung in der Bildverarbeitung implementiert. Jedes Bild wird in Patches von 30x30 Pixeln aufgeteilt (so wie das vom Modell Fuyu8b auch getan wird [23]) und jedes Bild, was über 784 Patches enthält wurde beim Trainieren ignoriert. Ein quadratisches Bild könnte also maximal aus 28x28 Patches oder 840x840 Pixel bestehen. Dieser Schritt gewährleistet fehlerfreies Training, da bei Bildern mit weniger als 784 Patches keine Out of Memory - Errors entstanden sind.

Darüber hinaus wurde eine 4-Bit-Quantisierung beim Laden des Modells verwendet. Diese Technik macht das Training ein bisschen langsamer, aber verringert den Speicherverbrauch erheblich und ermöglichte es erst, so grosse Bilder zu verarbeiten [24].

LoRA wurde nur auf die Query and Key Matrizen angewendet, da es auch für dies vorgesehen ist [22]. Den Rang der LoRA-Layers wurde im Laufe der Experimente zwischen 64 und 512 variiert, was ungefähr 0,389% bis 3,11% der Modellparameter trainierbar machte.

Für die Trainingsstrategie wurde Gradientenakkumulation eingesetzt, wobei ein Accumulation Step von 32 verwendet wurde. Damit konnte trotz des begrenzten Speichers von 40GB mit einer grösseren Batchsize trainiert werden, da die Gradienten akkumuliert werden und nur Zwischenresultate der Weights von einem Bild-Text-Paar zwischengespeichert wurden [25].

Als Optimizer wurde der 8bit quantisierten Lion-Optimizer verwendet. Dieser Optimierer braucht weniger Speicher und Compute-Time als AdamW [26], was es für dieses Projekt attraktiver machte.

Die CrossEntropyLoss-Funktion wurde als Verlust-Funktion ausgewählt. Dies ist eine Standardwahl für Klassifizierungsaufgaben und misst effektiv die Leistung von Modellen, die Wahrscheinlichkeitswerte ausgeben.

Um das Modell zur Generierung kontextuell angemessener Beschriftungen zu steuern, wurde eine spezifische Prompt verwendet: "Generate a coco-style caption." so wie sie im Beispiel in der Model-Card auf Huggingface steht [27]. Diese Prompt richtet die Ausgaben des Modells nach dem Stil des COCO-Datensatzes aus, einem Schlüsselbenchmark im Bereich der Bildbeschriftung.

4 Beschreibung des Datensatzes

4.1 Einleitung

Im Rahmen dieses Projekts wurde ein umfangreicher Datensatz zusammengestellt. Dieser Datensatz setzt sich aus mehreren bestehenden Datenkollektionen zusammen. Zu den integrierten Datensätzen gehören Autochart, Chart-to-text (Pew und Statista) sowie ChartSumm.

4.2 Analyse-Erkenntnisse

Kategorie	n-**Beschreibungen**	∅ **Länge (Worte)**	**Median Länge (Worte)**	**Min. / Max. Länge (Worte)**
Autochart	46,850	101.48	83	52 / 189
Chart-to-text pew	1,490	93.97	79	18 / 545
Chart-to-text statista	27,868	53.69	47	10 / 424
ChartSumm	84,363	45.44	44	0 / 304
Gesamt	160,571	63.67	52	0 / 545

Table 1: Zusammenfassung der Beschreibungen der Datensätze

Rank	**Autochart**	**Chart-to-text pew**	**Chart-to-text statista**	**Chart-Summ**	**Gesamt**
1	number	say	statistic	million	number
2	found	among	percent	years	found
3	lowest	americans	shows	dollars	lowest
4	may	news	united	statistic	may
5	related	pew	year	per	related
6	national	million	million	percent	shows
7	graph	research	dollars	rate	national
8	changes	according	number	us	graph
9	group	adults	billion	year	million
10	change	half	states	shows	changes

Table 2: Top 10 häufigste Worte in verschiedenen Datensätzen

- **Zusammensetzung:** Gesamthaft hat der kombinierte Datensatz 160'571 Beschreibungen. Wovon knapp mehr als die Hälfte aus dem Chart-Summ Datensatz, rund ein Viertel aus dem Autochart Datensatz und der Rest aus den Chart-to-text Datensätzen kommen. Auffällig ist, dass

Dataset	Flesch-Kincaid-Lesbarkeitsindix
Autochart	70.33
Chart-to-text pew	55.65
Chart-to-text statista	62.43
ChartSumm	59.73
Gesamt	63.25

Table 3: Durchschnittliche Flesch-Kincaid-Lesbarkeitsindizes verschiedener Datensätze

Dataset	∅ Breite	σ Breite	∅ Höhe	σ Höhe
Autochart	3178.04	5.06	1579.00	0.0
Chart-to-text pew	380.23	159.44	465.92	205.45
Chart-to-text statista	800.00	0.0	630.07	228.63
ChartSumm	1500.00	0.0	900.00	0.0
Gesamt	1407.30	512.38	865.82	232.25

Table 4: Durchschnittliche Höhe/Breite und Standardabweichungen der Bilder in den Datensätze

der Chart-to-text pew Datensatz deutlich unterrepräsentiert ist mir nur 1'490 Beschreibungen wie in Tabell 1 zu sehen ist.

- **Länge der Beschreibungen:** Wie der Tabelle 1 zeigt, variiert die Länge der Beschreibungen signifikant über die Datensätze. Der Autochart-Datensatz hat durchschnittlich die längsten Beschreibungen (101.48 Worte), während ChartSumm die kürzesten aufweist (45.44 Worte). Der Gesamtdurchschnitt über alle Datensätze beträgt 63.67 Worte, mit einer grossen Spannbreite von 0 bis 545 Worten, was auf eine vielfältige Beschreibungslänge in den verschiedenen Datensätzen hinweist und dass einige Beschreibungen im ChartSumm Datensatz leer sind.

- **Häufigste Wörter:** Über alle Datensätze hinweg ist *number* das am häufigsten vorkommende Wort. In AutoChart sind Begriffe wie *national*, *changes* und *group* prominent, während in Chart-to-text pew und statista wirtschaftliche Begriffe wie *dollars*, *million* und *billion* vorherrschen, wie in der Tabelle 2 zu sehen ist.

- **Flesch-Kincaid-Lesbarkeitsindex:** Gemäss Tablle 3 hat AutoChart den höchsten Indexwert, was auf eine einfachere Lesbarkeit hinweist, während Chart-to-text pew den niedrigsten Wert hat, was auf komplexere Textstrukturen hindeutet.

- **Grösse und Breite der Bilder:** Nach der Tabelle 4 variieren die Bilder in Grösse und Dimensionen über die verschiedenen Datensätze hinweg unterschiedlich. AutoChart hat beispielsweise die grössten Bilder, während

die Bilder in Chart-to-text pew tendenziell schmaler, aber höher sind.

4.3 Fazit

Die Integration dieser vielfältigen Datensätze bietet eine umfassende Grundlage für die Entwicklung des Tools. Nachdem entfernen einiger einzelnen Datenpaare (aufgrund zu kurzer Beschreibungen, wie in der Tabelle 1 zu erkennen ist) besteht der endgültige Datensatz aus 119'704 Datenpaare. Dieser Datensatz wurde im Anschluss auch gleichmässig in Trainings- (60%) Validierungs- (20%) und Testdatensatz (20%) aufgeteilt, wo darauf geachtet wurde, dass alle Anfangs-Datensätze in jedem Teil gleichmässig verteilt wurden.

5 Experimente

Globales Ziel: Das globale Ziel der Experimente ist es, die optimale Konfiguration des Fuyu-Modells zu identifizieren, indem die Auswirkungen verschiedener Trainingsparameter auf die Modellleistung und -effizienz systematisch untersucht werden.

Trainingsparameter: Um die Auswirkung verschiedener Trainingsparameter zu untersuchen, wurden in jedem Experiment Trainingsparameter angepasst (siehe Tabelle 5). Es wurden aber nicht alle Parameter verändert. So sind folgende Parameter immer gleichgeblieben:

- Quantisierung: 4-bit
- LoRA alpha: 128
- LoRA dropout: 0.05

Experiment	Datenratio	Lernrate	LoRA r	Epochen
1	0.2	0.0001	64	3
2	0.2	0.0001	64	1
3	0.05	0.00005	64	1
4	0.2	0.00005	64	1
5	0.2	0.00005	256	1
6	0.2	0.00005	512	1

Table 5: Übersicht der Parameter für jedes Experiment

5.1 Experiment 1

Hypothese: Die LoRA-Parameter ermöglichen eine angemessene Performanz des Modells auch bei reduzierten Datensätzen. Die Parameter wurden so gewählt, um einen Kompromiss zwischen Modellkomplexität und Rechenzeit zu finden.

Ergebnisse: Der Trainingsverlust zeigte eine schnell abfallende Tendenz, die darauf hindeutet, dass das Modell effektiv lernte und sich schnell an die gegebenen Daten anpasste. Die konsequente Abnahme des Verlustes bis zu einem Plateau erweckt den Eindruck, dass die gewählte Lernrate und die anderen Hyperparameter gut abgestimmt waren, um eine stabile Konvergenz zu erreichen. Jedoch zeigt ein solcher Verlauf nicht automatisch die Qualität des Modells an. Eine ähnliche Kurve kann auch auftreten, wenn die Lernrate und/oder die Anzahl der Epochen zu gross gewählt wurden, was dazu führen kann, dass das Modell sein Wissen vergisst/überschreibt. Um dies zu überprüfen, sind verschiedene Metriken und Beispiele von entscheidender Bedeutung (welche im Kapitel 2.4 beschrieben wurden).

An der Validierungsverlust-Kurve erkennt man wiederum sehr gut, dass das Modell leicht overfittet, da sich diese Kurve zunehmend anstatt abnehmend verhält.

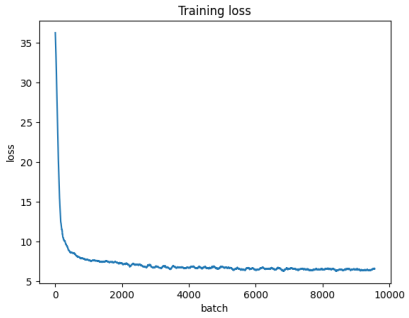


Figure 6: Trainingsloss

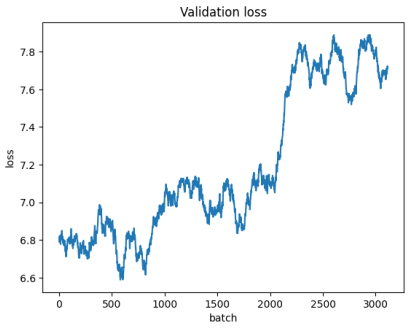


Figure 7: Validierungsloss Experiment 1

5.2 Experiment 2

Hypothese: Eine kürzere Trainingszeit hat einen positiven Einfluss auf die Modellleistung. Dazu wurde die Anzahl Epochen auf 1 gesetzt.

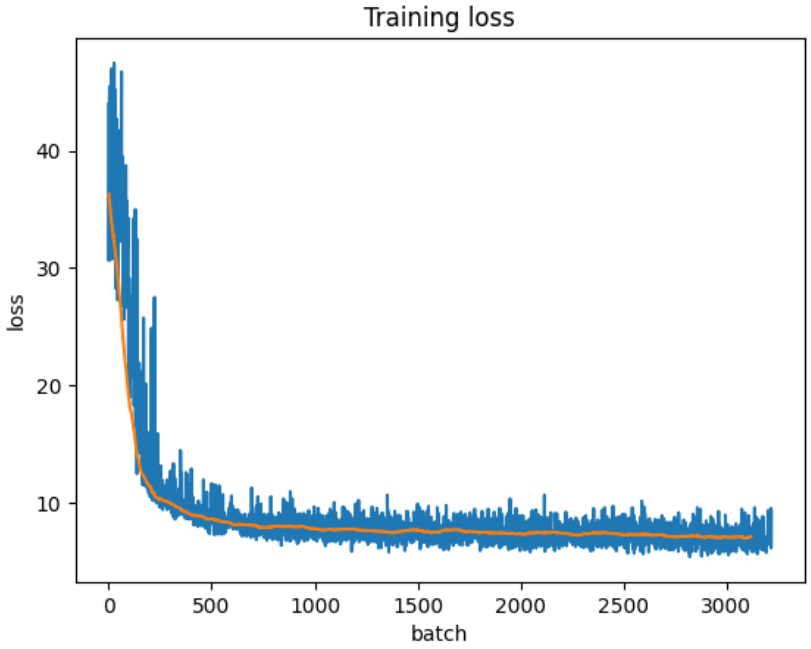


Figure 8: Trainingsloss

Ergebnisse: Auch hier zeigt der Trainingsverlust eine schnell abfallende Tendenz bis zum Erreichen des Plateaus. Dies zeigt hier auch wieder, dass das

Modell aus den Trainingsdaten etwas lernen konnte. Dass das Modell vergisst, lässt sich auch hier noch nicht ausschliessen.

Da in diesem Experiment das Modell lediglich für eine Epoche trainiert wurde, und jeweils am Ende der Epoche validiert wird, ist eine Bewertung der Validierungskurve nichtssagend.

5.3 Experiment 3

Hypothese: Weniger Daten und eine tiefere Lernrate haben einen positiven Einfluss auf die Modellleistung.

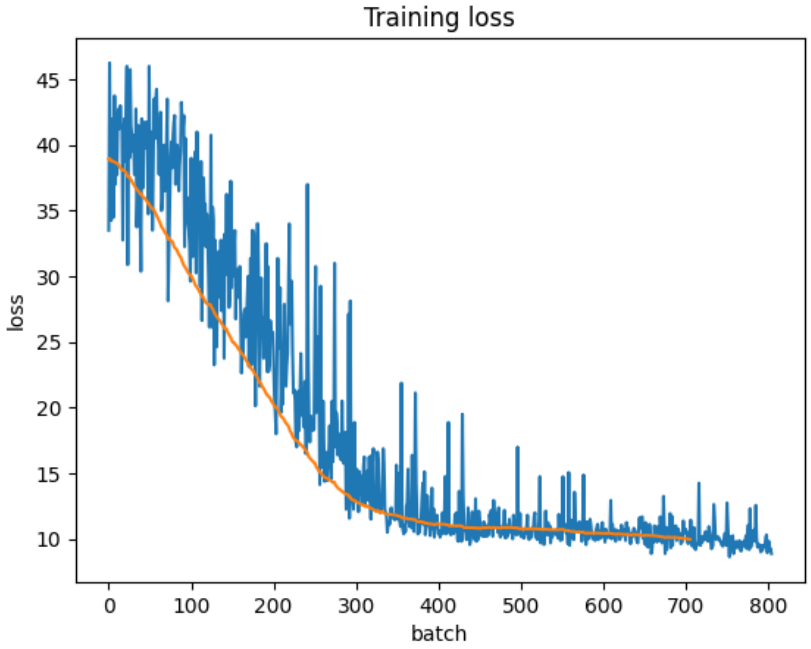


Figure 9: Trainingsloss

Ergebnisse: Der Trainingsverlust zeigt eine markante Abnahme mit einem allmählichen Übergang in ein Plateau. Die Verringerung der Lernrate scheint zu einer langsameren Konvergenz geführt zu haben. Trotz der stark reduzierten Datenmenge konnte das Modell den Trainingsverlust effektiv senken, was darauf hinweist, dass selbst unter stark limitierten Bedingungen eine gewisse Lernfähigkeit erhalten bleibt. Allerdings könnte die geringe Datenmenge auch das Risiko erhöhen, dass das Modell nicht genügend Vielfalt sieht, um generalisierbare Muster zu lernen.

5.4 Experiment 4

Hypothese: Eine tiefe Lernrate aber mehr Daten wirken sich positiv auf die Modellleistung aus.

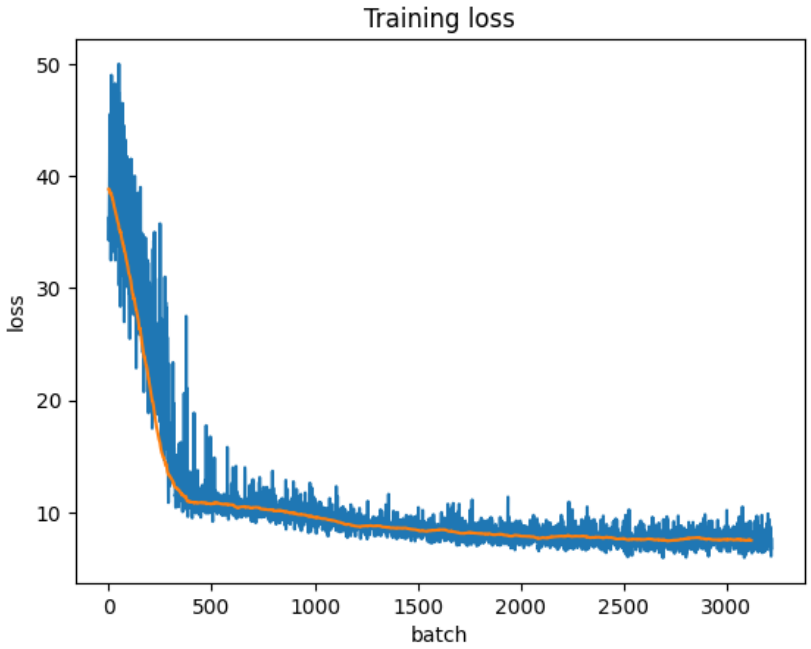


Figure 10: Trainingsloss

Ergebnisse: Der Trainingsverlust zeigt einen ähnlichen Verlauf wie im dritten Experiment, jedoch mit einem anfänglich steileren Abfall und der anschliessenden Stabilisierung auf einem Plateau. Die Erhöhung der Datenmenge scheint eine signifikante Änderung in der Geschwindigkeit der Konvergenz im Vergleich zur verringerten Datenmenge des vorherigen Experiments verursacht zu haben. Jedoch kann man hier auch nicht ausschliessen ob das Phänomen vom "Vergessen" eintritt.

5.5 Experiment 5

Hypothese: Ein höherer LoRA-Rank wirkt sich positiv auf die Modellleistung aus.

Ergebnisse: Die Trainingsverlust-Kurve sieht ähnlich aus wie bei Experiment 3, was darauf schliessen lässt, dass die Erhöhung des LoRA-Rank Parameters keinen Signifikanten Einfluss auf den Verlauf der Trainingsverlust-Kurve hat.

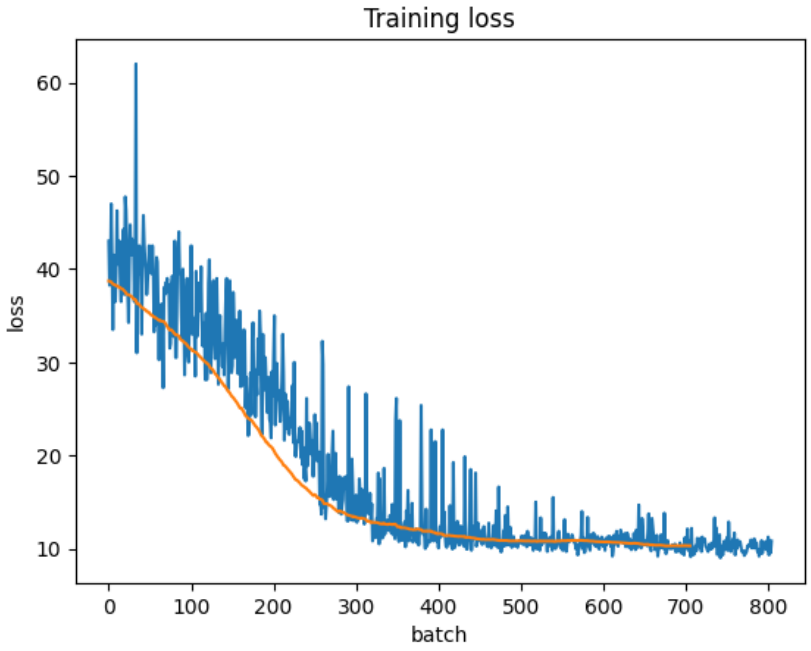


Figure 11: Trainingsloss

5.6 Experiment 6

Hypothese: Ein noch höherer LoRA-Rank wirkt sich positiv auf die Modellleistung aus.

Ergebnisse: Die Analyse der Trainingsverlustkurve könnte aufzeigen, ob die erhebliche Erhöhung des LoRA-Ranks von 256 auf 512 keinen merklichen Unterschied in der Lernkurve des Modells erzeugt. Erwartet wurde, dass eine höhere Kapazität in der Repräsentation durch LoRA zu einer schnelleren Konvergenz führen könnte.

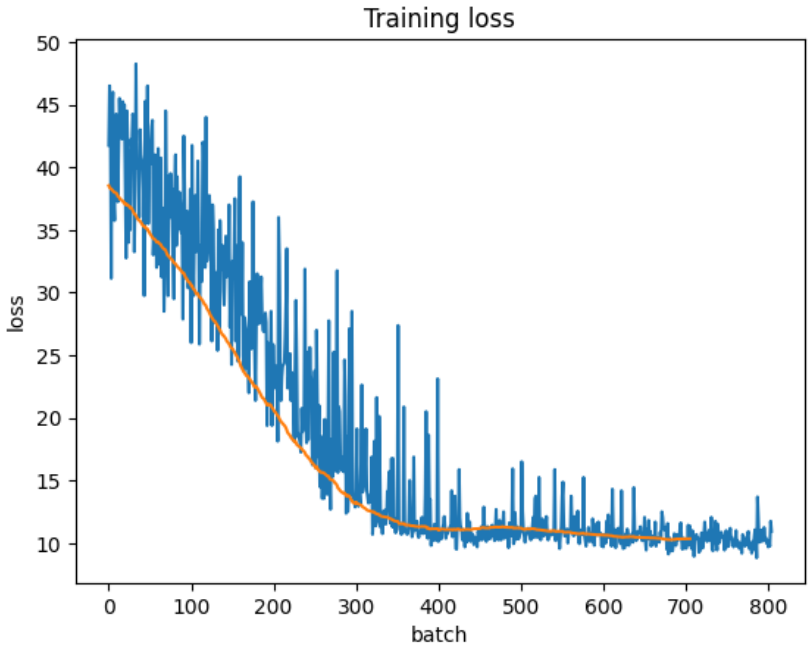


Figure 12: Trainingsloss

6 Ergebnisse

In diesem Kapitel werden die Resultate der leistungsstärksten Modelle für jedes Experiment detailliert untersucht. Um die optimalen Parameter für jedes Modell zu ermitteln, wurde ein Gridsearch über die Parameter Temperatur (`temperature`), Sampling-Methode (`do_sample`) und Wiederholungsstrafe (`repetition_penalty`) durchgeführt. Die Ergebnisse sind in der Tabelle 6 dargestellt.

Der Parameter `temperature` beeinflusst die Vorhersagegenauigkeit des Modells: Ein niedriger Wert wie 0.1 macht das Modell vorhersehbarer und konservativer, während ein höherer Wert wie 1 zu kreativeren und überraschenderen Ergebnissen führt.

Die Sampling-Methode bestimmt, ob bei der Vorhersage Zufälligkeit einbezogen wird (`True`) oder nicht (`False`), wobei `True` eine grössere Diversität in den Antworten ermöglicht.

Der `repetition_penalty`-Parameter beeinflusst, wie stark das Modell Wiederholungen vermeidet: Ein höherer Wert wie 1.5 reduziert die Wahrscheinlichkeit von wiederholten Inhalten stärker als ein niedrigerer Wert wie 1.0.

Die Auswahl der besten Modelle basierte auf den Metriken BLEU, ROUGE und WER die im Kapitel 2.4 erklärt wurden. Die Tabelle 6 illustriert die besten Modelle in absteigender Reihenfolge gemäss ihrer Leistung auf diesen Metriken,

wobei das oberste Modell als das effektivste nach dem BLEU-Wert gilt.

Model_name	**Temperature**	**Do-sample**	**Repetition-penalty**	**BLEU**	**ROUGE**	**WER**	**Experiment**
4bit_1_epoch lr_0.00005 tr_0.05_lora_64	0.7	True	1.5	2.77e-03	0.150	0.952	3
4bit_1_epoch lr_0.00005 tr_0.05_lora_512	0.7	False	1.0	1.65e-03	0.133	0.958	6
4bit_1_epoch lr_0.00005 tr_0.05_lora_256	0.1	True	1.0	1.19e-03	0.130	0.960	5
4bit_3_epoch lr_0.0001 tr_0.2_lora_64	0.4	True	1.5	5.59e-80	0.185	3.015	1
4bit_1_epoch lr_0.00005 tr_0.2_lora_64	1.0	True	1.5	4.66e-80	0.149	1.549	4
4bit_1_epoch lr_0.0001 tr_0.2_lora_64	0.7	True	1.5	1.74e-155	0.17.	2.301	2

Table 6: Resultate der Experimente sortiert nach BLEU-Score

In den Abbildungen 13, 14, 15, 16 und 17 werden die Auswirkungen dieser Modelle auf fünf verschiedene Testbilder dargestellt. Dabei werden die Beschreibung der Modelle in der Reihenfolge ihres BLEU-Scores sortiert.

6.1 Vergleich der Experimente

6.1.1 Experiment 1: 4bit_3_epoch_lr_0_0001_tr_0_2_LoRA_64

Im ersten Experiment wurde deutlich, dass trotz einer vielversprechenden Konvergenz des Trainingsverlustes der Verlauf des Validierungsverlustes über die Epochen hinweg stetig anstieg, was auf ein mögliches Overfitting hindeuten könnte. Die optimierten Hyperparameter des Modells, ermittelt durch Gridsearch, waren eine Temperatur von 0.4 und eine Repetitionsstrafe von 1.5. Diese Einstellungen sollten theoretisch zu einer Balance zwischen Diversität und Genauigkeit der generierten Texte führen und Wiederholungen einschränken.

Trotz der sorgfältigen Einstellung der Parameter erbrachte das Modell in der Praxis jedoch keine kohärenten oder zusammenhängenden Ergebnisse, wie die Analyse im Kapitel "Ergebnisse" aufzeigt. Die generierten Texte des Modells erschienen wirr und zusammenhangslos, was die effektive Nutzung des Fuyu-Modells zur Diagrammzusammenfassung in Frage stellt. Es scheint, als ob das

Modell durch den Fine-Tuning-Prozess wesentliche Teile seines vorherigen Wissens "überschrieben" hat, was zu unbrauchen Beschreibungen geführt hatte. Ebenfalls sieht man in Tabelle 6 auch, dass es nur Platz 4 erreicht hat von allen Experimenten.

6.1.2 Experiment 2: 4bit_3_epoch_lr_0_0001_tr_0_2_LoRA_64

Das zweite Experiment zeigte eine ähnlich positive Entwicklung des Trainingsverlustes wie Experiment 1, diesmal jedoch mit dem Training über lediglich eine Epoche. Der Gridsearch ergab, dass die besten Ergebnisse bei einer Temperatur von 0.7 und einer Wiederholungsstrafe von 1.5 erzielt werden.

Dennoch offenbarten die Metriken in Tabelle 6, dass das Modell insgesamt nicht erfolgreich war. Mit einem BLEU-Wert von 1.742641e-155, einem ROUGE-Wert von 0.172967 und einem WER von 2.301069 landete es auf dem letzten Platz unter allen Experimenten.

Die generierten Beschreibungen untermauern diese schlechten Metriken ebenfalls, wie in den Bildern 13, 14, 15, 16 und 17 zu sehen ist.

6.1.3 Experiment 3: 4bit_1_epoch_lr_0_00005_tr_0_05_LoRA_64

Im dritten Experiment zeigt sich erneut eine vielversprechende Entwicklung des Trainingsverlustes. Das Modell wurde für eine Epoche trainiert und die optimierten Hyperparameter (eine Temperatur von 0.7 und eine Repetitionsstrafe von 1.5) wurden mittels Gridsearch ausgewählt.

Die Ergebnisse in Tabelle 6 zeigen, dass dieses Experiment das erfolgreichste Modell aller Experimente ist. Mit einem BLEU-Wert von 0.00277, einem ROUGE-Wert von 0.149854 und einem WER von 0.952, bestätigen diese Metriken die Beurteilung, deuten aber gleichzeitig auf eine immernoch sehr schlechte Performance hin. Diese Werte markieren eine deutliche Verbesserung im Vergleich zu den vorherigen Experimenten.

Die im Kapitel 6 generierten Beschreibungen unterstreichen die Qualität dieses Modells. In den meisten Fällen liefert es gute Beschreibungen, im Vergleich zu den vorherigen Experimenten, die auch Informationen zur dargestellten Grafik enthält. Das Modell erkennt oft den richtigen Charttyp und extrahiert relevante Texte aus der Grafik. Obwohl Fehler auftreten, verknüpft es die erkannten Texte grösstenteils korrekt mit der Grafik. Besonders interessant ist, dass das Modell oft den Fokus auf den Titel der Grafik legt, was in den meisten Fällen zu verbesserten Beschreibungen führt.

6.1.4 Experiment 4: 4bit_1_epoch_lr_0_00005_tr_0_2_LoRA_64

Im vierten Experiment zeigt sich erneut eine vielversprechende Entwicklung des Trainingsverlustes. Das Modell wurde für eine Epoche trainiert, und die optimierten Hyperparameter (eine Temperatur von 1.0 und eine Repetitionsstrafe von 1.5) wurden mittels Gridsearch ausgewählt.

Trotz der vielversprechenden Trainingskurve zeigt die Auswertung in Tabelle 6, dass dieses Experiment den fünften Platz belegt, was darauf hinweist, dass es

zu den schlechteren Modellen gehört. Dies ist insbesondere interessant, da mehr Daten für das Training verwendet wurden als bei Experiment 3, aber ansonsten dieselben Parameter benutzt wurden. Dies lässt vermuten, dass das Modell möglicherweise sein Wissen durch das Training mit einer grösseren Datenmenge überschreibt.

Die erzielten Metriken bestätigen diese Einschätzung, mit einem BLEU-Wert von 4.656737e-80, einem ROUGE-Wert von 0.149040 und einem WER von 1.549063. Diese Metriken spiegeln sich auch in den generierten Beschreibungen wider, die keinen klaren Sinn ergeben und wie zufällig ausgewählte Token wirken.

6.1.5 Experiment 5: 4bit_1_epoch_lr_0_00005_tr_0_05_LoRA_256

Die Trainingskurve im fünften Experiment zeigt erneut vielversprechende Ergebnisse. Gemäss Tabelle 6 belegt dieses Modell den dritten Platz. Es verwendet die gleichen Hyperparameter wie das beste Modell (Experiment 3), jedoch mit einem grösserem LoRA Rank.

Die besten Parameter gemäss Gridsearch waren eine Temperatur von 0.1 und eine Repetitionsstrafe von 1.0, was interessanterweise die niedrigste Temperatur für alle Modelle in den Experimenten ist.

Die erzielten Metriken zeigen jedoch eine Verschlechterung im Vergleich zu Experiment 3, mit einem BLEU-Wert von 1.188312e-03, einem ROUGE-Wert von 0.130065 und einem WER von 0.959555. Die generierten Beschreibungen ähneln denen des besten Modells (Experiment 3) und zeigen keine klaren qualitative Unterschiede.

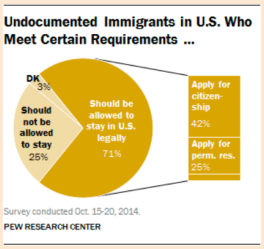
6.1.6 Experiment 6: 4bit_1_epoch_lr_0_00005_tr_0_05_LoRA_512

Die Trainingskurve im sechsten Experiment sieht ebenfalls gut aus. Gemäss Tabelle 6 belegt dieses Modell den zweiten Platz. Es verwendet die gleichen Hyperparameter wie das beste Modell (Experiment 3) und Experiment 5, jedoch mit einem noch grösseren LoRA Rank.

Die besten Parameter gemäss Gridsearch waren eine Temperatur von 0.7 und eine Repetitionsstrafe von 1.0.

Die erzielten Metriken zeigen jedoch eine Verschlechterung im Vergleich zu Experiment 3, mit einem BLEU-Wert von 1.645735e-03, einem ROUGE-Wert von 0.133363 und einem WER von 0.957850. Die generierten Beschreibungen ähneln denen des besten Modells (Experiment 3) und Experiment 5, jedoch sind die Metriken minimal besser als bei Experiment 5 und immer noch schlechter als bei Experiment 3. Dies könnte darauf hinweisen, dass nur das Erhöhen des LoRA Rank-Parameter nicht die gewünschten Verbesserungen bringt.

Example 1



Ground thruth:

When it comes to immigration policy, roughly seven-in-ten (71%) continue to support a way for undocumented immigrants to gain legal status if certain requirements are met. One-in-four (25%) say they should not be allowed to stay.
Most of those who say there should be a pathway to legal status support the ability for undocumented immigrants to apply for citizenship: 42% overall say they should be able to apply for citizenship, while 25% of the public thinks they should be able to apply for permanent residency, but not citizenship.

Generated texts:

Model - Experiment 3 (4bit_1_epoch_lr_0_00005_tr_0_05_lora_64)
A pie chart with the percentage of people's willingness to stay in U, S who meet the certain requirements to meet a certain requirement

Model - Experiment 6 (4bit_1_epoch_lr_0_00005_tr_0_05_lora_512)
A pie chart with the percentage of people's contributions to U.S. who meet certain requirements.

Model - Experiment 5 (4bit_1_epoch_lr_0_00005_tr_0_05_lora_256)
A pie chart showing the percentage of people who are not allowed to stay in U.S.

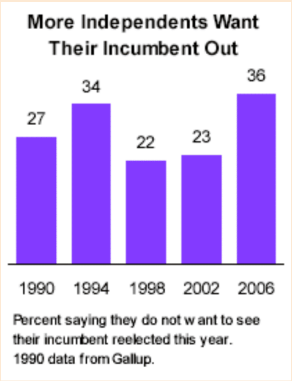
Model - Experiment 1 (4bit_3_epoch_lr_0_0001_tr_0_2_lora_64)
(and) or insS is of people was who in the U20 users/ survey are states41'0 dollars such growth their with thousand The from517 news rate1 national the for that about13 were million%) all on200 in the U population more increasedbased9 compared to In14 Americans6 as state which then30 adults power to less " over15 at [online COVID years25 it opinion the coronavirus United States most around market hads are16 Denmark total approximately
[…]

Model - Experiment 4 (4bit_1_epoch_lr_0_00005_tr_0_2_lora_64)
more than percent dollars Thes there were million% and8 years, approximately less than1320020'211 increase13 almost of) to was Ud15 thousand0 517111 is billion between the percent of business industry719 sales respondents cases* including population16 moren daily9 in3324 people618 were60 in the30),S 4014 over10 year100 millions from the total25 deaths70. reported been around after48 which about34 (). numbers every decade the
[…]

Model - Experiment 2 (4bit_1_epoch_lr_0_0001_tr_0_2_lora_64)
Thes182 percent andS56019 was13 million In' in percent of with year11 of)%16 Ud S thousand0 617 dollars is billion According toe for' about of thes are approximately21 years This to on between from US United States64 (924 people respondentsy were12 in the30), had companies33 as715 at are online less than1425 birth or. reported population around the average484 services increase with a). was a40 GDP the difference57 the63 product there
[…]

Figure 13: Beispielbild 1

Example 2



Ground thruth:

Last month, 36% of independent voters said they don't want to see the incumbent in their district reelected. This is as high as in October 1994 (34%), shortly before the historic 1994 midterm when Democrats lost control of Congress.

Generated texts:

Model - Experiment 3 (4bit_1_epoch_lr_0_00005_tr_0_05_lora_64)
A bar chart with the percentage of people who have not paid their insurance out

Model - Experiment 6 (4bit_1_epoch_lr_0_00005_tr_0_05_lora_512)
The bar chart shows how many % of people felt their independents

Model - Experiment 5 (4bit_1_epoch_lr_0_00005_tr_0_05_lora_256)
A bar chart with the percentage of people who have not paid their insurance.

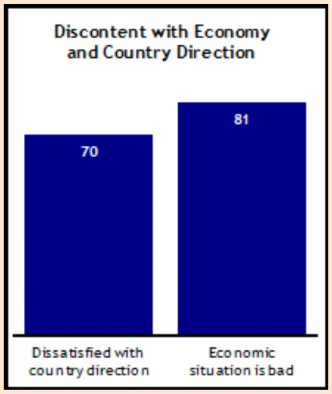
Model - Experiment 1 (4bit_3_epoch_lr_0_0001_tr_0_2_lora_64)
(% and in20) is InSs U the coronavirus' or were over by0 are people United States in the17 survey of about to9 population18 news%) more dollars the growth The41 then307 roughly who on200165 that was million very/ all market19 rate when13'8 since Denmark which from40based compared to with37 increased religious information both25 amounted to respondents most states the average demand has hit national less Americans41 for percent
[...]

Model - Experiment 4 (4bit_1_epoch_lr_0_00005_tr_0_2_lora_64)
million percent7,13020130 and% were2419 thousand dollars respondents billion population years between10s) U almost2011570 approximately The6 in the2003 been more). of259 deaths more than5111in the difference According to less than the total2 the34n60.2911 reported there were),3326* tod or439638 mostly1837 overalls were the average for Number of926410031 yeary4069 persons from6862 average1746 its4136the times (48 minus35 days
[...]

Model - Experiment 2 (4bit_1_epoch_lr_0_0001_tr_0_2_lora_64)
was191820 and in3 In2 United States U1 to around04 with200 from percent of dollars7178 of about924s (thousand isS This percent year605 billion or30 million11 was a had)%610 in the on21 were euros ' between13e, approximately GDP3116/ by64 of the online there were the income the difference25 populationn.201 including respondents, increase57 volumes are: The product1563 companies rate'
[...]

Figure 14: Beispielbild 2

Example 3



Ground thruth:

As of fall 2009, seven-in-ten (70%) Czechs were dissatisfied with the way things were going in their country. Roughly eight-in-ten (81%) described the current economic situation in the Czech Republic as somewhat or very bad, with many (32%) saying very bad.

Generated texts:

Model - Experiment 3 (4bit_1_epoch_lr_0_00005_tr_0_05_lora_64)
A bar chart showing the difference between percent and country direction

Model - Experiment 6 (4bit_1_epoch_lr_0_00005_tr_0_05_lora_512)
A bar chart showing the percentage of a discounts for economy

Model - Experiment 5 (4bit_1_epoch_lr_0_00005_tr_0_05_lora_256)
A bar chart showing the percentage of a dependent with economy and country direction.

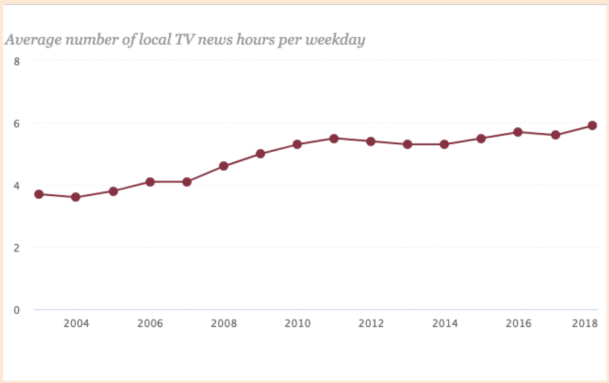
Model - Experiment 1 (4bit_3_epoch_lr_0_0001_tr_0_2_lora_64)
(and%0 people users in was or') fromS is that most of overs then soccer national scarce20 who16 political when dollars in the satisfaction the same survey less for adults5%) on/ four born are200 U were their religious: percent91 nearly In15 " the coronavirus compared to more which' a high about17 before7 roughly anti both million).19 all According to amongin the of the Americans to news13 have been and the growth while51 opinion share as risk), [the *[…]*

Model - Experiment 4 (4bit_1_epoch_lr_0_00005_tr_0_2_lora_64)
respondents20 million72n market0 population of1330 less than111 U ands).15 reported33% percent years,11 thousand8 to604 times), billion92416 more292011200 were10 approximately —)

Model - Experiment 2 (4bit_1_epoch_lr_0_0001_tr_0_2_lora_64)
(and200 million from19 was517 200 around The21 of% had67 In) U316Ss dollars to4 billion United States for percent of percent18 in the sales or30 in S of the income between population'9 about8 by is This there were' the average age25 the difference industry companies at are thousandy15).60 approximately year thisin11 users increase were market over per68 growth health including with a69 as),64 world., when/ the with more overall *[…]*

Figure 15: Beispielbild 3

Example 4



Ground thruth:

The average amount of weekday local TV news programming was increased slightly in 2018, according to the RTDNA/Hofstra University survey. Local TV stations dedicated an average of 5.9 hours to news programming per weekday in 2018, up slightly from 5.6 hours in 2017.

Generated texts:

Model - Experiment 3 (4bit_1_epoch_lr_0_00005_tr_0_05_lora_64)
A bar chart with the number and percentage of news hours per week per weekend day

Model - Experiment 6 (4bit_1_epoch_lr_0_00005_tr_0_05_lora_512)
A graph showing the number of local TV news hours per weekend.

Model - Experiment 5 (4bit_1_epoch_lr_0_00005_tr_0_05_lora_256)
A graph showing the number and percentage of news hours per week.

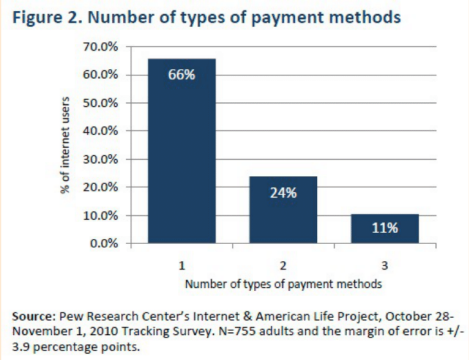
Model - Experiment 1 (4bit_3_epoch_lr_0_0001_tr_0_2_lora_64)
percent, to16820 and0 in theSs in In billion U total or was18 dollars (people of million The19 had about euros512 the%764 is news) approximately3 from when13 between96 figures around"10 amounted to respondents GDP are percent of United States the coronavirus March by15 were rate A4 of the survey with14 Americans COVID industry the average dollars in lived in years increase thousand at17 population roughly until countries in the U
[…]

Model - Experiment 4 (4bit_1_epoch_lr_0_00005_tr_0_2_lora_64)
more, According to) in15 percent7 million and The' to were billion020 approximatelyn in the24191 sales% dollars of the average (respondents since Us is daily2009 reported less than the difference spending by its per revenue for years almoste estimated M16 people the number of525 gross which13 from1118"6 was every17 source18 pounds In Russia3010 over more than annual population Number of60 industry48 S2931 satisfaction been world half
[…]

Model - Experiment 2 (4bit_1_epoch_lr_0_0001_tr_0_2_lora_64)
percent of20020 (0S as18 billion U S to around) percent and at fromd The statistic had about2 In This the% in was of reported dollars approximately or4613 betweene5 with.' in the10 of the7 million179 United States24 by income were. A ares are total population14 online15 US the averages were21 for is2512 thousand68 respondents,201 According to value internet11100 M'n33 the difference $57).2964 business60 overall after rate8 there were
[…]

Figure 16: Beispielbild 4

Example 5



Ground thruth:

As shown in Figure 2, most internet users who have accessed online content have utilized only one method of access and payment (66%).

Generated texts:

Model - Experiment 3 (4bit_1_epoch_lr_0_00005_tr_0_05_lora_64)
A bar chart with the percentage of payment methods available for each type of project

Model - Experiment 6 (4bit_1_epoch_lr_0_00005_tr_0_05_lora_512)
A bar chart with the percentage of payment methods.

Model - Experiment 5 (4bit_1_epoch_lr_0_00005_tr_0_05_lora_256)
A bar chart showing the percentage of payment methods.

Model - Experiment 1 (4bit_3_epoch_lr_0_0001_tr_0_2_lora_64)
(% people7 and is in'9 of survey are) or20 from more19 that was200 AmericansSs U In%) in the13 were1 The as less5 on had million about the1716 all to news if'40 rate0 who15 dollars market8414 has the coronavirus nearly1810 among almost growth COVID41/30 of the percent which the total2 thousand25 five3 figures new around6 when — states11 increased then over it most four the average civilization power roughly according to36 more than38
[...]

Model - Experiment 4 (4bit_1_epoch_lr_0_00005_tr_0_2_lora_64)
percent200150211)18'49,2611117phone and7"rafts3 been).1960s52in632014429 According to33*635%15 respondents14205421216 of40485437d427.131123925 the averagehalf of aals are317032table at1093413795 less than57rack andships in71room is1593558447A5541 might have been92483Units356366944512128button is6212086701140ebras11912851329ify theage of91465024f458821381009323555. the difference150nets600809076rests
[...]

Model - Experiment 2 (4bit_1_epoch_lr_0_0001_tr_0_2_lora_64)
from200182434 and19s percent of151 in dollars17 to The10S In billion of (S was were percent0 as)d with approximately million had the GDP16%9 around in the value for less than 25e30 market year thousand60s are by118n.64014 the average reported33 is This, gross at volume 13 which United States or people U about over3127/ its" increase US64 euros product201. M per share of.63 growth29s were price $ since domestic there werein when
[...]

Figure 17: Beispielbild 5

7 Diskussion

7.1 Zusammenfassung der Erkenntnisse

In der vorliegenden Analyse wurden sechs Experimente miteinander verglichen und analysiert. Die Experimente zeigten eine Reihe von Ergebnissen, die von mangelnder Kohärenz bis hin zu verbesserten, sinnvolleren Beschreibungen reichten.

Experiment 1 und 2 offenbarten Schwächen in der Modellkohärenz, trotz vielversprechender Trainingsverluste, was auf Overfitting hinweisen könnte. Insbesondere Experiment 2 schnitt mit den niedrigsten Metriken aller Experimente ab. Experiment 3 hingegen stellte sich als das erfolgreichste heraus, mit den besten Werten in BLEU, ROUGE und WER. Dies deutet darauf hin, dass die verwendeten Hyperparameter und Trainingsmethoden eine effektivere Balance erzielen konnten. Experiment 4 zeigte trotz ähnlicher Parameter wie Experiment 3 schlechtere Ergebnisse, was die Wichtigkeit der Datenmenge und der Feinabstimmung der Parameter unterstreicht. Experiment 5 und 6 verbesserten sich im Vergleich zu den ersten beiden Experimenten, blieben aber hinter Experiment 3 zurück, trotz des Einsatzes eines höheren LoRA Ranks.

Interessant ist, dass die BLEU-Werte so schlecht sind. BLEU-Werte von 0.60 sind für Bildbeschreibungen nicht unüblich [28]. Dennoch sind die gemessenen BLEU-Werte deutlich schlechter als ein Hunderstel davon. Eine Erklärung könnte darin liegen, dass man Grafiken auf viel unterschiedlichere Arten beschreiben kann als ein simples Bild (z.B. mit/ohne Interpretation, Trends, etc.).

7.2 Diskussion der Modellkonfigurationen

Die Ergebnisse legen nahe, dass die Konfiguration der Modelle eine entscheidende Rolle spielt. Insbesondere die Feinabstimmung von Hyperparametern wie Temperatur und Repetitionsstrafe sowie die Variation von Trainingsparametern wie die LoRA Parameter, Lernraten und Datenmenge sind entscheidend. Es ist offensichtlich, dass ein Gleichgewicht zwischen diesen Parametern gefunden werden muss, insbesondere um mit grösseren Datenmengen effektiv zu arbeiten. Diese Erkenntnis unterstreicht die Notwendigkeit weiterer Forschung und Experimente, um optimale Konfigurationen zu identifizieren, die kohärente und genaue Diagrammzusammenfassungen ermöglichen.

Jedoch muss beachtet werden, dass solche umfassenden Untersuchungen sehr zeitintensiv sind, besonders angesichts der langen Trainingsdauern der Modelle. In diesem Projekt wurden aus diesem Grund nicht alle potenziell erfolgversprechenden Modellkonfigurationen ausführlich untersucht. Zukünftige Projekte könnten jedoch von einem tieferen Eintauchen in diese Aspekte profitieren, um das volle Potenzial der automatisierten Diagrammzusammenfassung auszuschöpfen.

8 Fazit

Unser Projekt zur automatisierten Diagrammzusammenfassung mit KI bietet eine Reihe von Erkenntnissen und Empfehlungen für zukünftige Arbeiten in diesem Bereich. Eine wichtige Erkenntnis ist die potenzielle Verbesserung der Modellleistung durch den Einsatz grösserer Bilder. Diese könnten aufgrund ihres erhöhten Detailgrades und Kontextes sowie einer besseren Auflösung von Beschriftungen zu präziseren Zusammenfassungen führen.

Des Weiteren könnten fortschrittlichere Techniken zur Speicheroptimierung (wie Gradient-Checkpointing [29] und DeepSpeed ZeRO [30]) das Trainieren von noch grösseren Modellen auch auf Consumer-Hardware ermöglichen. Hierdurch liesse sich die Systembelastung reduzieren, während zugleich die Leistungsfähigkeit erhöht wird. Die Generierung diversifizierter Diagramme, möglicherweise durch den Einsatz von KI-Modellen wie ChatGPT und Code-Interpretern, ist ebenfalls vielversprechend [31]. Eine solche Diversifizierung der Trainingsdaten könnte das Modell besser auf reale Anwendungen vorbereiten.

Das Experimentieren mit verschiedenen Prompts im Training könnte auch neue Möglichkeiten eröffnen, da unterschiedliche Prompts die Genauigkeit und Kreativität des Modells beeinflussen können [32].

Eine umfassendere Erprobung verschiedener Parameterkombinationen, ermöglicht durch zusätzliche Rechenleistung, könnte zu einer Optimierung der Modellkonfigurationen führen. Auch die Erforschung unterschiedlicher Modellarchitekturen könnte dazu beitragen, die am besten geeigneten Ansätze für die Diagrammzusammenfassung zu identifizieren.

Eine weitere Empfehlung ist die Einbeziehung textlicher Informationen (wie OCR) in den Input des Modells. Trotz der Fähigkeit unseres aktuellen Modells, Texte in Grafiken effektiv zu erkennen, könnte dieser Schritt die Genauigkeit der Zusammenfassungen verbessern. Sollten in Zukunft leistungsfähigere GPU's verfügbar sein, könnte auch die Überprüfung des Trainings ohne Quantisierungstechniken erfolgen, um zu beurteilen, ob dies die Modellleistung weiter steigern kann, auch wenn eine Quantisierung nicht zwingend einen negativen Einfluss hat [24].

Zusammengefasst zeigt unser Projekt auf, dass das Gebiet der automatisierten Diagrammzusammenfassung vielversprechend ist und durch weitere Forschung und Innovation weiterentwickelt werden kann. Durch die Umsetzung dieser Erkenntnisse und Empfehlungen könnte die Entwicklung effektiverer Tools für die Diagrammzusammenfassung weiter vorangetrieben werden.

9 Einsatz von KI-Tools im Projekt

Im Rahmen des Projekts kamen verschiedene KI-Tools zur Anwendung (ChatGPT / Copilot), die hauptsächlich beim Programmieren eingesetzt wurden. Die Ergebnisse dieser Tools entsprachen jedoch nicht den Erwartungen. ChatGPT wurde zudem für das Verständnis komplexer Themen genutzt. Weiterhin fand es Anwendung bei der Korrektur und Verbesserung von Texten.

References

[1] K. Lee, M. Joshi, I. Turc, *et al.*, *Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding*, arXiv:2210.03347 [cs], Jun. 2023. DOI: 10.48550/arXiv.2210.03347. [Online]. Available: http://arxiv.org/abs/2210.03347 (visited on 01/16/2024).

[2] F. Liu, F. Piccinno, S. Krichene, *et al.*, *MatCha: Enhancing Visual Language Pretraining with Math Reasoning and Chart Derendering*, arXiv:2212.09662 [cs], May 2023. DOI: 10.48550/arXiv.2212.09662. [Online]. Available: http://arxiv.org/abs/2212.09662 (visited on 12/12/2023).

[3] Z. Yang, L. Li, K. Lin, *et al.*, *The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)*, arXiv:2309.17421 [cs], Oct. 2023. DOI: 10.48550/arXiv.2309.17421. [Online]. Available: http://arxiv.org/abs/2309.17421 (visited on 01/16/2024).

[4] V. D. Lai, N. T. Ngo, A. P. B. Veyseh, *et al.*, *ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning*, arXiv:2304.05613 [cs], Apr. 2023. DOI: 10.48550/arXiv.2304.05613. [Online]. Available: http://arxiv.org/abs/2304.05613 (visited on 01/16/2024).

[5] F. Liu, J. M. Eisenschlos, F. Piccinno, *et al.*, *DePlot: One-shot visual language reasoning by plot-to-table translation*, arXiv:2212.10505 [cs], May 2023. DOI: 10.48550/arXiv.2212.10505. [Online]. Available: http://arxiv.org/abs/2212.10505 (visited on 12/12/2023).

[6] J. Li, D. Li, S. Savarese, and S. Hoi, *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models*, arXiv:2301.12597 [cs], Jun. 2023. DOI: 10.48550/arXiv.2301.12597. [Online]. Available: http://arxiv.org/abs/2301.12597 (visited on 12/12/2023).

[7] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, *MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models*, arXiv:2304.10592 [cs], Oct. 2023. DOI: 10.48550/arXiv.2304.10592. [Online]. Available: http://arxiv.org/abs/2304.10592 (visited on 12/12/2023).

[8] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, *et al.*, *Fuyu-8B: A Multimodal Architecture for AI Agents*, en, Oct. 2023. [Online]. Available: https://www.adept.ai/blog/fuyu-8b/ (visited on 12/12/2023).

[9] H. Liu, C. Li, Q. Wu, and Y. J. Lee, *Visual Instruction Tuning*, arXiv:2304.08485 [cs], Apr. 2023. DOI: 10.48550/arXiv.2304.08485. [Online]. Available: http://arxiv.org/abs/2304.08485 (visited on 12/12/2023).

[10] L. Chen and K. Zhao, "An Approach for Chart Description Generation in Cyber–Physical–Social System," en, *Symmetry*, vol. 13, no. 9, p. 1552, Sep. 2021, Number: 9 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2073-8994. DOI: 10.3390/sym13091552. [Online]. Available: https://www.mdpi.com/2073-8994/13/9/1552 (visited on 09/27/2023).

[11] A. Balaji, T. Ramanathan, and V. Sonathi, "Chart-Text: A Fully Automated Chart Image Descriptor," en, May 2023.

[12] J. Thiyam, S. R. Singh, and P. K. Bora, "Chart classification: An empirical comparative study of different learning models," in *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, ser. ICVGIP '21, New York, NY, USA: Association for Computing Machinery, Dec. 2021, pp. 1–9, ISBN: 978-1-4503-7596-2. DOI: 10.1145/3490035.3490291. [Online]. Available: https://doi.org/10.1145/3490035.3490291 (visited on 12/19/2023).

[13] N. Ratner, Y. Levine, Y. Belinkov, *et al.*, *Parallel Context Windows for Large Language Models*, arXiv:2212.10947 [cs], Aug. 2023. DOI: 10.48550/arXiv.2212.10947. [Online]. Available: http://arxiv.org/abs/2212.10947 (visited on 12/19/2023).

[14] T. Norlund, L. Hagström, and R. Johansson, "Transferring Knowledge from Vision to Language: How to Achieve it and how to Measure it?" In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, J. Bastings, Y. Belinkov, E. Dupoux, *et al.*, Eds., Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 149–162. DOI: 10.18653/v1/2021.blackboxnlp-1.10. [Online]. Available: https://aclanthology.org/2021.blackboxnlp-1.10 (visited on 12/19/2023).

[15] W. Dai, Z. Liu, Z. Ji, D. Su, and P. Fung, *Plausible May Not Be Faithful: Probing Object Hallucination in Vision-Language Pre-training*, arXiv:2210.07688 [cs], Feb. 2023. DOI: 10.48550/arXiv.2210.07688. [Online]. Available: http://arxiv.org/abs/2210.07688 (visited on 01/16/2024).

[16] W. Wang, L. Dong, H. Cheng, *et al.*, *Visually-Augmented Language Modeling*, arXiv:2205.10178 [cs], Feb. 2023. DOI: 10.48550/arXiv.2205.10178. [Online]. Available: http://arxiv.org/abs/2205.10178 (visited on 12/19/2023).

[17] J. Zhu, J. Ran, R. K.-W. Lee, Z. Li, and K. Choo, "AutoChart: A Dataset for Chart-to-Text Generation Task," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, Held Online: INCOMA Ltd., Sep. 2021, pp. 1636–1644. [Online]. Available: https://aclanthology.org/2021.ranlp-1.183 (visited on 09/27/2023).

[18] S. Kantharaj, R. T. Leong, X. Lin, *et al.*, "Chart-to-Text: A Large-Scale Benchmark for Chart Summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 4005–4023. DOI: 10.18653/v1/2022.acl-long.277. [Online]. Available: https://aclanthology.org/2022.acl-long.277 (visited on 01/18/2024).

[19] R. Rahman, R. Hasan, A. A. Farhad, M. T. R. Laskar, M. H. Ashmafee, and A. R. M. Kamal, "ChartSumm: A Comprehensive Benchmark for Automatic Chart Summarization of Long and Short Summaries," *Proceedings of the Canadian Conference on Artificial Intelligence*, Jun. 2023, arXiv:2304.13620 [cs]. DOI: 10.21428/594757db.0b1f96f6. [Online]. Available: http://arxiv.org/abs/2304.13620 (visited on 01/18/2024).

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds., Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. [Online]. Available: https://aclanthology.org/P02-1040 (visited on 01/18/2024).

[21] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013 (visited on 01/18/2024).

[22] E. J. Hu, Y. Shen, P. Wallis, *et al.*, *LoRA: Low-Rank Adaptation of Large Language Models*, arXiv:2106.09685 [cs], Oct. 2021. DOI: 10.48550/arXiv.2106.09685. [Online]. Available: http://arxiv.org/abs/2106.09685 (visited on 01/18/2024).

[23] B. Li, P. Zhang, J. Yang, Y. Zhang, F. Pu, and Z. Liu, *OtterHD: A High-Resolution Multi-modality Model*, arXiv:2311.04219 [cs], Nov. 2023. DOI: 10.48550/arXiv.2311.04219. [Online]. Available: http://arxiv.org/abs/2311.04219 (visited on 01/18/2024).

[24] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, *QLoRA: Efficient Finetuning of Quantized LLMs*, arXiv:2305.14314 [cs], May 2023. DOI: 10.48550/arXiv.2305.14314. [Online]. Available: http://arxiv.org/abs/2305.14314 (visited on 01/18/2024).

[25] J. Lamy-Poirier, *Layered gradient accumulation and modular pipeline parallelism: Fast and efficient training of large language models*, arXiv:2106.02679 [cs], Jun. 2021. DOI: 10.48550/arXiv.2106.02679. [Online]. Available: http://arxiv.org/abs/2106.02679 (visited on 01/18/2024).

[26] X. Chen, C. Liang, D. Huang, *et al.*, *Symbolic Discovery of Optimization Algorithms*, arXiv:2302.06675 [cs], May 2023. DOI: 10.48550/arXiv.2302.06675. [Online]. Available: http://arxiv.org/abs/2302.06675 (visited on 01/18/2024).

[27] Adept AI Labs, *Adept/fuyu-8b · Hugging Face*, Nov. 2023. [Online]. Available: https://huggingface.co/adept/fuyu-8b (visited on 01/19/2024).

[28] K. Xu, J. Ba, R. Kiros, *et al.*, *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*, arXiv:1502.03044 [cs], Apr. 2016. [Online]. Available: http://arxiv.org/abs/1502.03044 (visited on 01/19/2024).

[29] T. Chen, B. Xu, C. Zhang, and C. Guestrin, *Training Deep Nets with Sublinear Memory Cost*, arXiv:1604.06174 [cs], Apr. 2016. DOI: 10.48550/arXiv.1604.06174. [Online]. Available: http://arxiv.org/abs/1604.06174 (visited on 01/19/2024).

[30] S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He, *ZeRO: Memory Optimizations Toward Training Trillion Parameter Models*, arXiv:1910.02054 [cs, stat], May 2020. DOI: 10.48550/arXiv.1910.02054. [Online]. Available: http://arxiv.org/abs/1910.02054 (visited on 01/19/2024).

[31] Y. Han, C. Zhang, X. Chen, *et al.*, *ChartLlama: A Multimodal LLM for Chart Understanding and Generation*, arXiv:2311.16483 [cs], Nov. 2023. DOI: 10.48550/arXiv.2311.16483. [Online]. Available: http://arxiv.org/abs/2311.16483 (visited on 01/19/2024).

[32] X. L. Li and P. Liang, *Prefix-Tuning: Optimizing Continuous Prompts for Generation*, arXiv:2101.00190 [cs], Jan. 2021. DOI: 10.48550/arXiv.2101.00190. [Online]. Available: http://arxiv.org/abs/2101.00190 (visited on 01/19/2024).