A3

Erste Spur und z weite Spur läuft auch. Perfekt. Gut, dann würde ich gerne anfangen und einfach mal fragen, hey, stell dich doch mal vor, was ist dein Background?

I7

Ja, also wurde schon gerade ein bisschen angeteasert von XXXXXXX XXX XXXXX XXXXXXXX XXX XXXXXX XXXXXXXX XXXX XX XXX XX XXXXXXXX XXX XXXX XXX XXXX XXXX XXXXXX XXX XXXXXXXXXXXXX XXXX bin ich dann ins Ausland gegangen für eine Forschungs- und Entwicklungsstelle im Bereich Künstliche Intelligenz, das war die Türkei und dort haben wir halt ich glaube insgesamt 4-5 Projekte oder so in den 1,5 – 2 Jahren die ich da war abgeschlossen im Bereich KI es waren verschiedenste Projekte, hauptsächlich im Bereich Computer Vision und ja, war schon ziemlich interessant und also die Sache ist, ich habe auch viele NTS(?) unterschrieben, ich glaube ich darf nicht über alles reden, aber das was so publik ist, ist so ja ich muss überlegen… Ich glaube dazu sage ich besser nichts aber wir haben halt sehr viel Computer Vision gemacht und war super, super spannend aber ich habe in diesem Bereich halt gemerkt, dass das Zusammenarbeiten an einem KI-Projekt, wo man auch die Daten beispielsweise von den Kunden bekommt und die dann irgendwie großflächig etikettieren muss etc., dass das schon sehr, sehr umständlich war und da haben dann meine Kollegen und ich, das ist noch ein Studienkollege aus der Informatik und einer, den habe ich so an der XX XXXXXXXX kennengelernt wir drei haben da so ein Start-up gegründet ich bin da zurück aus der Türkei Ende 2022 und wurde hier erstmal vom XXX XXXXXXXXXXXXXXXXX gestipendiert äh, gesponsert unterstützt und habe dann im Anschluss noch das Exist(?) bekommen mit meinem Team und jetzt, das läuft diesen Monat aus und ja, wir haben schon unsere, unseren ersten Kunden, unsere ersten Aufträge, so es geht in Richtung Software as a Service, interessanterweise heißt das Produkt, das Projekt, XXXXXXXXXX also ja, interessanter Name auf jeden Fall zu dem Interview heute genau, das ist eigentlich so der Stand der Dinge, wir haben, was wir mit KI im Unternehmen jetzt intern zu tun haben ist einerseits klassisch Deep Learning andererseits, und das war auch für mich eine Lektion, wenn heute von KI geredet wird meinen die nicht mehr Deep Learning die meinen ausschließlich und exklusiv LLMs und, das heißt wir machen auch viel LLMs in letzter Zeit. Ich glaube ich habe alles gesagt.

A3

Okay. Ja, alles gut. Was ist genau euer Geschäftsmodell? Möchtest du darauf eingehen? Und wie nutzt ihr Deep Learning und LLMs konkret im Arbeitsalltag oder in Form eures Produktes?

I7

Ja, also wir haben in allererster Linie XXXXXXXXXX, das ist eine Webplattform, Software as a Service, wo man seine Daten einhaken kann, ja, einklinken kann und dann kann man das in Arbeitspakete einteilen, etikettieren. Wir haben einen super coolen, modernen Labeling-Editor, das ist so im Prinzip das, was uns so besonders ausmacht. Allerdings, wenn es dann, ich sag mal, darum geht, das zu vertreiben, haben wir relativ schnell gemerkt, dass viele Unternehmen noch gar nicht auf dem Stand sind, eigene Deep-Learning-Projekte durchzuziehen. Und da haben wir auch angefangen, so ein bisschen unterstützend mitzuwirken, das heißt, ja, Beratung slash Dienstleistung als Vertriebsweg. Weil wenn wir einmal im Unternehmen drin sind, dann kriegen wir auch das Produkt unseres aktuellen Erachtens noch besser vermarktet, verkauft. Genau. Das ist so unser Geschäftsmodell.

A3

Okay, das heißt quasi dieses Labeling, die Labeling-Oberfläche an sich hat keine KI mit eingebaut, aber ihr seid Consultants für zum Beispiel das Anwenden von Deep Learning Modellen und Large Language Models. Kann man das so zusammenfassen?

I7

Ja, es gibt natürlich unterstützende Aspekte im Labeling Editor. Ich weiß nicht, ihr kennt bestimmt solche Sachen wie Segment Anything, die halt das Labeling besonders schnell vorantreiben können. Da sind wir jetzt gerade dabei, verschiedene KI-Werkzeuge zu ergänzen, sodass das so ein, wir nennen das ALMA, automatisiertes Labeling mit menschlicher Aufsicht. Ich finde das ist ein cooler Begriff. Ja, genau. Dass wir so das ALMA so gesehen jetzt Schritt für Schritt einführen. Und da sind halt mehrere KI-Module, die genutzt werden.

A3

Okay, alles klar, sehr interessant. Was ist denn genau deine Rolle in dem Unternehmen als Gründer?

I7

Genau. Ich bin der Geschäftsführer bei uns, das heißt, ich mache so gesehen natürlich in Absprache mit den Kollegen, aber irgendwo die finale Entscheidung. Ich bin der KI-Entwickler und auch ich bin nicht der Projektleiter tatsächlich. Das Projekt leitet der andere Kollege. Der hat so ein Big Picture und Software-Architektur ist da ein Ticken besser. Aber ich mache dann halt diese spezialisierten Lösungen, speziell wenn es um KI-Implementierung geht oder bestimmte Module, die fertig geschrieben werden müssen. Genau. Das ist meine Rolle. Und natürlich, interessanterweise ist Programmieren noch gar nicht mal die größte Sache, die ich gerade mache, sondern in letzter Zeit sehr viel Reden, Veranstaltungen, Netzwerken, sehr viele Office-Produkten, PowerPoint, Word, etc., Excel. Also, ja, interessanterweise.

A3

Das glaube ich. Also der Management-Anteil wird quasi immer mehr. Ja. Okay, interessant. Du, aber du machst trotzdem immer noch was, du programmierst immer noch. Du hast Vorerfahrung in der Computer Vision. Hast du schon gesagt, hast du irgendwie aus dem Studium noch irgendwelche besonderen Erfahrungen oder Expertise angesammelt?

I7

Ja, tatsächlich war meine Masterarbeit absolut industrierelevant, habe ich hinterher gemerkt. Ich habe halt während des Studiums sehr viel Computer Graphics, Computer Vision, kann man eigentlich sagen, gemacht, sehr viel am XXXXXXXXX X. Hinterher habe ich auch die Masterarbeit in diesem Bereich gemacht und das war dann halt so eine Schnittmenge zwischen Computer Graphics und AI. Das heißt, was ich dort gemacht habe, ist, ich habe im Prinzip über Kamerabilder die Materialparameter von einer Oberfläche, die man ja zum Rendern eigentlich benötigt für eine Szene, die habe ich mittels einer KI extrahieren können und natürlich, wenn ich die Materialparameter habe, kann ich die ganzen Szenen nochmal komplett neu beleuchten. War eine super, super spannende Arbeit. War das erste Mal, dass ich KI selbst genutzt habe oder beziehungsweise in der Programmierung genutzt habe, also als Implementierung. Ich habe mir immer gedacht, so, nee, das ist so ein Themengebiet, damit möchte ich nichts zu tun haben. Das scheint mir so wie Legos zusammenstecken, Daten draufschmeißen und das schien mir nicht so interessant. Da habe ich das in der Masterarbeit gemacht. Ich muss sagen, eigentlich hatte ich recht. Das war wirklich Legos zusammenstecken, Daten draufschmeißen und hoffen, dass das funktioniert. Natürlich ein bisschen involvierter. Also ganz einfach ist das nicht, aber irgendwo im Ganzen stimmt das schon. Hat mir sehr viel Spaß gemacht und das Tweaken an den Modellen in den Layern war auch sehr, sehr spannend. Und das hat am Ende dazu geführt, dass ich tatsächlich eben so eine Liebe zu der ganzen Thematik entwickelt habe. Und auch genau das, was ich dort gemacht habe. Ich habe beispielsweise ein UNET genutzt, ein selbstgeschriebenes. Das habe ich auch hinterher öfter im Job auch nochmal irgendwo auf eine ähnliche Art und Weise umsetzen können. Und das war sehr erfolgsbringend. Also das war schon sehr interessant.

A3

Ah, sehr schön. Okay, aber dann bist du ja quasi jetzt für dein Startup ein bisschen von Computer Vision weggekommen. Was für Modelle oder was für Technologien, KI-Technologien nutzt du denn so oder nutzt ihr aktuell konkret? Zum Beispiel für dieses ALMA.

I7

Genau, also da werden halt mehrere Sachen aktuell in Betracht gezogen, die werden jetzt ausgetestet, durchgetestet einerseits ist es das Segment Anything, das geht natürlich immer noch in Richtung Computer Vision, da wir halt Bilder und Videos in erster Linie als ich sag mal, First Class Citizen behandeln bei uns ist das halt so, dass wir immer noch sehr viel Computer Vision Modelle haben Objektdetektionsmodelle was wir im Prinzip auch sehr viel nutzen in letzter Zeit ist, jetzt weg von Computer Vision tatsächlich, wie du gesagt hast ist natürlich Open AI wir haben auch angefangen mit Cloud rumzuexperimentieren haben da AWS Zugänge bekommen, das war auch sehr spannend und sehr wichtig weil halt viele Leute Privatsphäre bedenken haben korrekterweise, finde ich Lama 3 und ähnliche Modelle, die Open Source sind, damit machen wir auch sehr viel in letzter Zeit, also sehr viel Experimente, auch einen neuen riesen Server haben wir uns zugelegt damit wir das auch im größeren Stil testen können, genau solche Sachen

A3

Okay. Was sind denn so, du hast gerade schon Privatsphäre genannt, was sind denn so noch weitere Herausforderungen, die du in diesen alltäglichen Prozessen hast, also im Deployment von KI-Modellen oder auch vielleicht im Finden von verschiedenen KI-Modellen, die ihr nutzen könnt?

I7

Also einerseits, was sehr gut wäre, wir haben immer so eine Übersichtstabelle, der kriegt diese Information sehr gut gelöst, der kriegt das sehr gut gelöst. Aber was ich ziemlich gut finden würde, wäre eigentlich, dass man eher so etwas hat wie, das kann man für diese Use Case besonders gut nutzen, für diesen Use Case besonders gut nutzen, weil so eine Zahl bei irgendeinem Test bringt mir nichts. Dass er irgendwie das Jurastudium, die Tests in den Vereinigten Staaten besser hinbekommt als ein Mensch, bringt mir auch nichts. Und dass er Programmieraufgaben, die ja oftmals sehr ähnlich strukturiert aufgebaut sind, sehr gut lösen kann, bringt mir auch nichts. Ich muss wirklich sehen, wie das in der Umsetzung in, ich sag mal, tatsächlichen Anwendungsfällen genutzt werden kann. Und das ist schon die erste große Challenge. Das heißt, ich weiß nicht, wenn ein LLM, egal wie das präsentiert wird, welche Kapazitäten das hat. Ich muss das wirklich sehr tief selber testen und bis an die Limits bringen, um herauszufinden, können wir das benutzen und wenn ja, wie können wir das benutzen. Genau, das ist so einer der größten Challenges, würde ich sagen. Und ansonsten natürlich, und damit lebt man aktuell sowieso, wie kann man den dazu bringen, wirklich die Aufgaben, die man möchte, auch nur die Aufgaben zu lösen, Halluzinationen einzudämmen, auf Fakten basiert arbeiten zu lassen. Da gibt es auch viele Techniken, die wir jetzt auch nutzen, aber ja, das sind so die Challenges. Wir versuchen dagegen zu wirken, sag ich mal. Ich weiß nicht, ob das die Frage ganz beantwortet.

A3

Ja, auf jeden Fall. Wie ist es denn mit den Nicht-LLM? Jetzt bei zum Beispiel deinen Computer Vision Modellen, die ihr nutzt. Habt ihr da auch irgendwie Herausforderungen?

I7

Ja, auch aus dem Job fällt mir da direkt eine Sache ein, die sehr wichtig ist. Das ist wirklich eine Sache, Computer Vision in der Praxis, finde ich, ist überhaupt nicht so wie, ach, wir haben einfach einen Objektdetektor, das war's. Wir haben sehr hochspezifische Daten. Die Daten müssen nicht unbedingt Bilder und Videos sein, wie wir sie kennen. Es können mikroskopische Aufnahmen sein, es können Wärmebilder sein, es kann alles Mögliche an Kamerabildern sein und dafür funktionieren halt durchschnittliche Modelle meistens gar nicht. Und gerade wenn dann auch noch die Qualität oder bestimmte Aspekte in einem Video oder einem Foto nicht so vorliegen, wie das die KI eigentlich kennt, dann sind die meisten Sachen für die meisten Anwendungsfälle in meiner Erfahrung unbrauchbar. Also kann sie gar nicht benutzen. Und deswegen muss man sehr, sehr oft Custom-Lösungen im Bereich Computer Vision anlegen, um etwas Bestimmtes lösen zu können. Da reicht noch nicht mal ein Transfer-Learning. Also das muss schon oft sehr, sehr, sehr spezifisch sein, um eine auch sehr spezifische Aufgabe lösen zu können. Damit würde ich sagen, insgesamt ist Computer Vision bzw. Deep Learning braucht wesentlich mehr Expertise und auch Experten als die ganze LLM-Geschichte. Ich glaube, da muss man nicht so involviert drin sein, dass man das hinbekommt.

A3

Okay, verstehe. Und siehst du irgendwie was, das ist jetzt vielleicht ein bisschen in die Glaskugel schauen, aber was wir auch in der KI-Forschung zum Beispiel besser machen können, um diese Probleme irgendwie zu lösen? Was siehst du da, was wird benötigt?

I7

Also wenn du diese Probleme sagst, meinst du jetzt gerade, was ich gesagt habe, dass da sehr spezielle KI-Modelle angelegt werden müssen, um bestimmte Aufgaben zu lösen?

A3

Ja, gut, ich meine, ihr löst das Problem durch Customizability, aber siehst du noch irgendwie was anderes, was euch helfen könnte?

I7

Also die Sache ist, das ist halt Forschung und Entwicklung und ich weiß nicht, was man da machen kann, damit das weniger Forschung und Entwicklung wird. Also man kann, das passiert ja auch, immer mehr Nischen abdecken durch immer mehr Nischenlösungen, immer mehr Startups konzentrieren sich auf immer weitere Bereiche. Das ist definitiv irgendwo natürlich der wirtschaftliche Weg, aber auch vielleicht die Lösung, weil ich weiß leider nicht, wie man das auch glaskugeltechnisch, keine Ahnung tatsächlich, wie man so eine Forschungsentwicklungsaufgabe nachträglich, nachhaltig lösen kann. Ich habe keine Ahnung. Also es ist sehr schwierig.

A3

Okay, ja, also wir haben uns dazu auf jeden Fall Gedanken gemacht. Es kann schon grob in die Richtung gehen. Wir wollen KI-Modelle vergleichbar machen miteinander, sodass quasi ein Auswahlprozess vereinfacht wird. Ich würde dir gerne mal zeigen, was du vielleicht auch schon gesehen hast auf der Webseite. Und zwar sind das unsere schönen KI-Label. sieht man das?

I7

Ich sehe gerade nur schwarz. Ah jetzt

A3

Jetzt sieht man was. Ok. ich kann es aktuell noch nicht so ganz auf dein Problem übertragen aber ohne dich jetzt weiter zu primen und irgendwie zu beeinflussen sagt mir doch gerne mal was du hier spontan siehst was ist deine Reaktion auf dieses KI-Label?

I7

Also die Reihenfolge, in der ich das jetzt gelesen habe, kann ich auch direkt sagen, das ist ja sehr spannend für den Designer, glaube ich. Also ich habe als allererstes natürlich dieses A, B, C, D, E, das fällt sofort ins Auge und dann oben angefangen, ja, dieses AI-Level habe ich nicht wirklich gelesen, dann direkt MobileNet V3, dann dieses Infer Image-Net und dann bin ich runtergesprungen zu der Batterie, halt wie viel Energie das verbraucht, schätze ich mal. Das Rechts daneben verstehe ich nicht. Corrupted Robustness. Unten links Top Accuracy. Und unten rechts Running Time per inference. Ja, also oben links, unten rechts ist mir sofort ersichtlich. Unten links müsste ich wissen, worauf sich die Accuracy bezieht. Und oben rechts, tatsächlich sagt mir gar nichts, muss ich sagen. Also, äh...

A3

Oben rechts die Robustness, meinst du? Okay. Ja, ich meine, die Accuracy ist, also wir haben hier einen Bildklassifikator, MobileNet V3, ImageNet. Ich weiß nicht, ob du schon mal mit ImageNet gearbeitet hast.

I7

Ich glaube schon, ja.

A3

Ja, aber das ist ja ein riesiger Bilddatensatz mit tausend Klassen, auf denen quasi immer unterschiedliche Objekte zu sehen sind. Und ja, das MobileNet V3 soll halt eben klassifizieren, was auf einem spezifischen Bild zu sehen ist. Ist da ein Hund drauf oder eine Katze oder ein Eis am Stiel oder irgendwie sowas. Und die Accuracy bezieht sich in dem Fall auf, im Testdatensatz, wie viel das MobileNet V3 eben richtig erraten hat, was auf dem Bild zu sehen ist. Also 63% Rate. Du kannst dir denken, wenn wir jetzt random einfach eine von diesen tausend Klassen wählen würden, dann hätten wir eine eintausendste Accuracy. Und die Robustness, das ist eine Metrik, die steht dafür, wie reliabel letztendlich dieses Modell entscheidet. Das heißt, man attackiert quasi die Bilder und ändert Pixel baut ein bisschen Neues ein.

I7

Ah ja das kenne ich

A3

und dann schaut man inwieweit sich das Ergebnis des mobilenets das Klassifikationsergebnis eben ändert und wie robust es quasi gegenüber dieser Pixelangriffe ist so genau das wäre das zu der Robustness. Hast du hast du noch weitere Fragen hast du irgendwelche was ist dir gerade noch reingekommen jetzt auch während der Erklärung vielleicht oder durch die Erklärung?

I7

Also jetzt, wo du Robustness erwähnt hast, Corrupted Robustness, da fällt mir… es ist wahrscheinlich bezogen auf diesen Angriff mit diesem Paper, dass man beispielsweise durch das Verändern von einigen wenigen Pixeln die komplette Klasse ändern kann, so wie man möchte. Dann sagt man, er kennt gerade einen Autor, aber das soll eine Ente sein. Für den Menschen ist das visuell nicht anders, aber der kriegt am Ende, ich glaube, den Angriff meinst du wahrscheinlich.

A3

Also, ich weiß nicht, XXXXXXX, ich habe das so verstanden, dass das hier tatsächlich durch eine Random-Pixel-Attacke passiert ist, aber es gibt, das was du meinst, ist natürlich diese Inference-Attacks, dass man wirklich spezifisch Pixel ändert, um die Klasse dann zu ändern.

A1

Genau, also man muss auch dazu sagen, dass der Wert tatsächlich gerade nur hypothetisch ist, die anderen sind reale Messwerte. Ich arbeite gerade daran, nochmal eine Test-View zu bauen, wo wir mit Auto-Attack praktisch adversarial(?) Robustness auch end-to-end durchtesten. Genau, also es gibt halt einmal so Targeted-Attacks, wo man wirklich versucht, den perfekten Angriff zu finden, um beispielsweise einen Labelflip hervorzurufen. Das ist natürlich nochmal ein bisschen schwieriger, dagegen robust zu sein, weil das dann auch in der Regel versucht, die Ausgaben des Modells auf anderen Bildern oder die Gradienten oder so zu benutzen, um halt irgendwie den Input praktisch zu vergiften, also so Poison-Attacks. Was wir hier tatsächlich testen werden, ist eher so eine komplett generische, wir schmeißen da verschiedene Verzerrungen oder Gorschen-Noise(?) oder was weiß ich was drüber mit einer bestimmten Norm und dann weiß man halt, okay, gegeben diese Angriffe, dann haben wir immer noch die und die Robustheit. Aber ja, also Robustness-Testing und Adversarial-Testing ist auch wieder so ein Rabbit Hole, also man kann das halt beliebig komplex machen.

I7

Ja, verstehe ich. Also was mir noch dazu aufgefallen ist oder eingefallen ist, ist wenn wir damals irgendwelche Modelle bewerten wollten. Was wir auch oft gemacht haben, ist, wir haben einfach verschiedene Teams verschiedene Ansätze verfolgen lassen, als ich damals noch im Ausland gearbeitet habe. Und um die am Ende zu vergleichen auf den Bildern war eigentlich die für uns Aussagekräftigste Metrik, waren immer die False-positives, False-negatives, dass man das irgendwie in der Übersicht bekommt. Und das war auch irgendwie das Ziel, das immer zu minimieren. Und tatsächlich war das auch so, wenn wir das minimiert bekommen haben, da war natürlich auch die visuelle, also das war immer natürlich diese quantitative Geschichte, aber diese qualitative, irgendwo fast schon subjektive Erkenntnis, wie gut die KI funktioniert, war direkt zusammen mit der, mit der qualitativen Metrik, äh, quantitativen Metrik von False-positives, False-negative. Irgendwo natürlich klar, aber es gibt auch Metriken, die man irgendwie optimiert und da ändert sich visuell eigentlich nichts. Aber das schien mir immer eine ziemlich gute Metrik zu sein, das war auch immer das Erste, worauf wir geschaut haben. Fällt mir dazu ein, wenn ich das jetzt sehe. Und eine andere Sache, die mir auch einfällt, muss ich sagen, ist, ich sehe das Label jetzt und ich, zum Beispiel das jetzt beispielsweise A bewertet. Ich denke mal, damit ist gemeint, overall das ganze Modell ist ein sehr, sehr gutes, gutes Modell, deswegen hat es das Rating A bekommen. Also ich, ich schätze mal, dass es damit gemeint ist und nicht irgendwie Energieeffizienz A. Wenn das so ist, dann denke ich mir, ja, ich würde ein bisschen, vorsichtig bei so einem Label sein, so auch in ein paar Jahren, weil was heißt A für mich von vor sieben Jahren, von vor drei Jahren, von vor zwei Jahren, im Vergleich zu heute, da ist ja A eine ganz andere Geschichte. Und, ähm, also ich müsste irgendwie noch wissen, zeitlich gesehen, wann war das denn A? Also wenn man irgendwie so ziemlich groß, ich weiß, da steht 2012, damit ist das Modell gemeint, nehme ich mal an.

A3

Und sogar hier, also das haben wir tatsächlich auch issued, das steht für wann das Label quasi entstanden ist. Also Juli 24, quasi dieser Monat.

I7

Ja genau, das ist irgendwie so, wenn das einen Ticken prominenter ist, weil ich finde, diese zwei Informationen gehören auf jeden Fall, also für mich jetzt, das ist jetzt meine Sicht, das ist jetzt nur meine Idee dahingehend, also die gehören absolut zusammen und müssen fast ähnlich prominent dargestellt werden, finde ich, damit man das Label wirklich auf einen Blick nutzen kann. Also 2012 müsste ähnlich groß sein oder 2024 müsste ähnlich groß sein, beispielsweise wie das A, B, C, D, E, wäre jetzt meine Intuition dahingehend, aber das ist mein erster Gedanke da.

A3

Okay, ja, super. Genau, dieses Gesamtrating hier, das speist sich quasi aus unter anderem diesen Metriken und ich glaube noch so sechs weiteren. Ich habe die ganze genaue Zahl nicht im Kopf und das kommt alles aus einer Studie von XXXXXXX, die vor ein paar Jahren publiziert wurde oder letztes Jahr, wo mehrere Netze, die die gleiche Aufgabe lösen, also in dem Fall hier Klassifikation von Bilddaten, miteinander verglichen wurden und aus diesem Vergleich hat sich dann jeweils eben gespeist, hey, ist das jetzt zum Beispiel, wenn man sich hier die Batterie ansieht, ist das im Vergleich zu anderen Modellen besonders energieeffizient oder ist es besonders akkurat und all diese Metriken zusammen ergeben dann eben den Compound Score, quasi das, was du hier oben siehst. Ja, so ist das entstanden.

I7

Ich sag mal, wenn ich das sehe, dann die größten, größten, größten Vorteile, die ich davon nutze, ist auch eine Story, den nicht nur Entwicklern, weil Entwickler, ja, um ehrlich zu sein, die kriegen, oder ich würde das jetzt auch, und ich denke, viele andere, die klar auch, würden das auch ohne so ein Label hinbekommen. Und auch viele Informationen, die darüber hinaus, also dann, wenn das so das Übersichtslabel ist, und dann hättet ihr noch so Zusatzlabels oder irgendwie so tiefere Informationen, dann würde das vielleicht doch schon sehr sinnvoll sein. Ja, aber den größten Nutzen da, den ich jetzt, ich sag mal, auf Anhieb sehe, ist gerichtet an die Entscheidungsträger und die Kunden, die dann am Ende, also Kunden im Sinne von Kunden für einen KI-Entwickler, das Unternehmen möchte jetzt KI einführen, und wir können denen dann diese verschiedenen Labels zeigen. Ich sag mal, wo ich den allergrößten Nutzen sehe, ist, dass man dem dann zeigen kann, mit zwei Labels, schau mal, wir haben die Lösung, dass man das einmal mit beispielsweise einem vortrainierten multimodalen LLM lösen kann, deine Objekte zu finden. Das ist jetzt, Entwicklungsaufwand ist da ziemlich gering, du promptest das ein bisschen, was er da finden soll, und dann kriegt er halt diese Inferenz dadrauf hin, aber schau mal, der Energieverbrauch ist davon im Vergleich zu einem spezifischen Modell, was genau die gleiche Aufgabe löst, unverhältnismäßig viel höher. Darüber hinaus ist die Zeit, die das LLM braucht, unverhältnismäßig viel länger, das heißt, wir können schon mal [unverständlich] knicken beispielsweise, als Beispiel. Und wenn man dann diese zwei Label hinhalten kann, das eine hat so ein D oder so und das andere hat ein A, dann sollte eigentlich dem Entscheider sofort klar sein, boah, eigentlich ist so ein spezifisches Modell echt viel besser als so ein generisches Gigantenmodell, was sehr abstrakt viele Aufgaben lösen kann. Darüber hinaus bin ich mir sicher, dass die Accuracy auch bei so einem spezifischen Modell sehr hoch sein wird. Also alles in allem sich da den größten Nutzen tatsächlich hat, wenn ich das den beiden vorhalten kann und sagen kann, hey, schau mal, das würden wir gerne dir vorstellen.

A3

Oh ja, sehr guter Punkt. Ich würde gerne auch einfach mal ein zweites Label dazu holen. Genau, das ist genau ja auch unter anderem ein Fall, den wir uns gedacht haben mit diesen KI-Modellen, dass man eben zwei Modelle zum Beispiel zeigt, so wie das hier ist. Hier haben wir einmal das MobileNet V3 auf der rechten Seite, das, was wir gerade schon hatten. Und jetzt haben wir einfach das EfficientNet B4 dazu geholt. Ist natürlich auch ein Image-Klassifikator, also jetzt kein Large-Language-Model, wie du es gerade in deinem fiktiven Beispiel genannt hast. Aber siehst du hier, also gerade jetzt, wenn du nochmal so einen Vergleich hast, vielleicht auch eine Brücke zu deiner alltäglichen Arbeit, würdest du das nutzen?

I7

Ich muss sagen, als erstes, was ich erkenne, ist, mir war vorher gerade eben nicht klar, dass die Icons unten color-coded waren. Jetzt im Nachhinein ergibt es natürlich Sinn, aber ich dachte, es ist einfach ein grünes Batteriesymbol und ein rotes Accuracysymbol und das sind die Farben davon gewesen. Jetzt erst erkenne ich, dass das eine Signifikanz hat, welche Farbe das hat. Umso besser ist das jetzt, wenn man das verstanden hat, dass es so funktioniert. Aber das war jetzt gerade nicht so ersichtlich für mich. Ob das sinnvoll ist, zwei KI-Modelle so zu vergleichen?

A3

Ja, und ob du eine Brücke für deinen Arbeitsalltag siehst, wo du es nutzen könntest.

I7

Also ich mach so, das ist das, was ich vorhin ganz am Anfang meinte ein bisschen, quantitative Metriken und qualitative Metriken, beziehungsweise jegliche Metrik, die in einer Zahl ausgedrückt wird versus ich hab das Modell ausgetestet, um zu schauen, wie das wirklich mein Problem löst. Und also ich hab bis jetzt noch keine Metrik gesehen, die das sehr gut zeigen konnte. Wie gesagt, False Positives, False Negatives kamen dem schon am nächsten, aber auch selbst dort hatten wir mal Modelle, die wesentlich niedrigere False-Negatives hatten, aber trotzdem augenscheinlich mehr falsche Resultate zeigten. Also am Ende heißt es, ich würde beide trotzdem probieren müssen und ich würde natürlich intuitiv anfangen bei denen, die A geratet sind, damit ich vielleicht mehr Arbeit spare, wenn das ziemlich gut läuft. Ich würde aber trotzdem durch viele durchtesten und mich auch erfahrungsgemäß auf so ein Label nicht verlassen können, weil das ist einfach eine sehr spezifische Sache. Also als Entwickler, das ist der wichtige Punkt, glaube ich, an dieser Stelle.

A3

Okay, würdest du dich gar nicht auf das Label verlassen oder würdest du sagen, vielleicht hilft es für eine Vorauswahl oder einfach um einen Überblick vielleicht sich zu verschaffen oder sagst du wirklich, nee, ich muss alles austesten, das hat keinen Mehrwert für mich?

I7

Also sieh es mal so, wenn ich anfangen würde zu testen und ich würde sehen, die ersten Modelle sind sehr top, die so gut bewertet sind mit dem Label und je weiter runter ich gehe nach dem Label, werden die wirklich wesentlich schlechter, dann würde ich irgendwann aufhören, dann würde ich nicht noch bis E durchprobieren, denke ich mal also genau

A3

Okay, also mal so ein bisschen an das Label gewöhnen, so sich ran tasten.

I7

Ja, also schon das Level so ein bisschen unterstützend nutzen, zu wissen, wann ich aufhören muss zu testen.

A3

Okay, alles klar. Was fällt dir denn an dem Label jetzt generell gefragt, fällt dir hier irgendwas besonders negativ vielleicht auf, was dir aufstößt, was du denkst, was man besser machen könnte, um es hilfreicher zu gestalten für dich?

I7

Also außer den Sachen, die schon angesprochen wurden, dass das Jahr vielleicht ein Ticken größer sein muss, damit man versteht, weil ich denke jetzt, okay, wenn wir das dieses Jahr benutzen, das Label, oder nächstes Jahr, okay, aber wenn ich auf einen Schlag sehen kann, boah, das ist von 2013 ein A-Rating beispielsweise, also wenn man jetzt mal in die Zukunft guckt, von 2024 ein A-Rating, wir sind im Jahr 2034, dann das würde ich auf einen Blick gerne auch sehen wollen, denke ich. Damit das, ja genau. Das Zweite ist, wie gesagt, ich habe nicht erkannt direkt, dass die Icons unten color-coded sind. Erst nachdem ich das zweite Label gesehen habe, ist mir das aufgefallen. Aber da steht eigentlich so das Wichtigste mit drauf. Wie gesagt, dieses Corrupted Robustness war für mich bis jetzt nicht relevant, glaube ich. Und der ist die Top One Accuracy, ob man da vielleicht Metriken, oder ob man da vielleicht andere Metriken reinmachen kann, ich weiß das nicht, kann ich jetzt nicht sagen. Aber alles in allem sieht das optisch sehr gut aus, also sieht sehr ansprechend aus. Genau. Als Übersicht.

A1

Vielleicht einmal kurz als Einschub, weil das vielleicht auch für dich relevant ist. Also die Software, ich habe halt eine Software entwickelt in den letzten Jahren, um solche Label komplett generisch zu generieren. Du kannst im Endeffekt ein Pandas DataFrame reinladen mit dieses Modell auf diesem Datensatz, hat diese Eigenschaften und entsprechend viele Rows dann, wie viele Datensätze und Modelle du getestet hast. Und theoretisch kannst du, also ich habe versucht, das so, es ist halt Work-in-Progress-Class, kein Produkt oder so, aber du kannst halt auch selber entscheiden, welche Icons dafür genutzt werden sollen und so ein Kram. Das ist im Moment alles noch ein bisschen dirty, weil es halt irgendwie hauptsächlich ein Forschungssoftware ist und nicht für den produktiven Einsatz bisher entwickelt wurde. Aber die Idee ist halt, dass auch du solche Label generisch selber generier... Also du kannst gegeben deiner eigenen Evaluationen und Benchmarks und so, die du gemacht hast und halt auch genau die Metriken, die für dich halt relevant sind, so ein Label generieren, um anderen Leuten zu vermitteln, was das Modell, was du vielleicht speziell trainiert hast, irgendwie kann. Oder die Modelle, die du irgendwie dir mal Benchmark-mäßig angeschaut hast, was die können. Und wie du auch am Anfang schon gesagt hattest, ist das vor allem halt relevant, oder zumindest hoffe ich, dass das vor allem relevant ist, um halt diesen Knowledge-Gap zu überbrücken. Jetzt in deinem Fall hast du den Vorteil, du bist Management und KI-Experte, das heißt, du kannst Modelle selber trainieren und bewerten und gegeneinander vergleichen und Entscheidungen fällen. Aber in ganz vielen Fällen, in ganz vielen Firmen sitzen auch Leute in Entscheidungspositionen, die von dieser Materie kein tiefes Verständnis haben. Und das Gesamtscoring ist natürlich auch immer so ein bisschen gegeben der eigenen Prioritäten. Also wenn man sagt, man will ein sehr energieeffizientes Modell oder man hat harte Constraints auf der Laufzeit, das muss innerhalb von so und so vielen Millisekunden irgendwie die Kamerabilder auswerten, dann sind das natürlich auch wieder so Domänen- oder Use-Case-spezifische Einschränkungen. Und das Label kann man natürlich auch, und das Gesamtscoring kann man da demnach natürlich auch anpassen und sagen, diese Metriken sind besonders wichtig. Das heißt, das ist so ein bisschen ein Open-Source-Gedanke, der hier hinter steckt. Und ja, insofern ist das auch ganz anpassbar auf individuelle Fälle.

I7

Ich muss auch wirklich sagen, jetzt wo ich drüber nachdenke, auch als wir am Ende unser Modell hinbekommen haben, also für einen Kunden, da waren das wirklich auch die wichtigsten Aspekte, dass wir wirklich die Zeit unglaublich runterbekommen haben, also Inference und die Accuracy unglaublich hochbekommen haben. Und das Lustige ist, also hätten wir damals die Möglichkeit gehabt, unsere vortrainierten oder unsere Rates/Waits(?), sag ich mal, mit so einem Dings auch intern bewerten zu können, dass wir direkt auf einen Blick sehen, welche Rates/Waits(?) sind wie gelabelt, also Label jetzt in diesem Sinne. Das wäre schon, das wäre schon auch sinnvoll so eine Übersichtstabelle oder Datenbank so. Ja gibt natürlich auch Alternativen, dass man so Benchmarking Listen hat und so, aber ja, also wäre schon irgendwo relevant gewesen, denke ich.

A3

Das ist ein sehr interessanter Use Case, den du da aufmachst. Das ist ja quasi während der Entwicklung eigenentwickelte Modelle auch schnell mal zu labeln, um einfach besser kommunizieren zu können, besser einen Überblick zu haben über das, was man eben selber gerade da entwickelt hat. Ja, da machst du auch einen guten Punkt noch auf. Welche anderen Kommunikationsformen kennst du denn überhaupt? Wie stellt ihr euch KI-Modelle untereinander dar, um sie zum Beispiel im Entwicklerteam zu kommunizieren oder auch an Kunden zu kommunizieren? Oder auch wie trefft ihr eine Auswahl? Auf welche Informationen greift ihr da zurück?

I7

bei uns das so… Sorry ich wurde gerade angerufen du hast gefragt welche Informationen wir nutzen um untereinander die KI-Modelle vorzustellen und zu kommunizieren. genau also lustigerweise wir hatten halt natürlich Loss/Losts und der Loss war schon sehr involviert und wenn man den loss runter bekommen hat das schon einiges ausgesagt aber, und es ist ein großes aber, am Ende des Tages was immer immer immer die visuelle Abnahme auf unseren typischen Test Videos weil also am Ende des Tages ist wirklich eine quantitative Metrik und das nochmal selber gesehen zu haben, das ist wirklich super wichtig und das interessante ist auch es bringt mir nichts wenn das Modell sehr viel quantitativ rausholt weil es einige bestimmte Szenarien super optimiert bekommen hat aber das was der Kunde eigentlich am meisten zu sehen bekommt ziemlich blöde Fehler rein macht dann wird der Kunde bei der Abnahme nur das sehen und dann ist das auch egal wieviel punkte das Ding intern gemacht hat aber natürlich ist da eine super starke Korrelation zwischen der quantitativen Metrik und dem qualitativen Ergebnis. Wie wir es am Ende gemacht haben war dann wirklich so, dass wir geguckt haben, welches visuell die besten Ergebnisse Liefert. Und wir haben den auch immer behalten. Wir hatten dann leider kein vernünftiges System. Wir hatten dann solche Sachen wie beispielsweise Modellnamen One model to rule them all oder so ein schwachsinn, damit wir wussten okay das ist das Modell was alles andere toppt oder so. Ja aber das haben wir dann auch irgendwann abgekürzt, also die Abkürzung One Model Rule them all, das war dann die ersten Buchstaben haben wir davon genommen und das war dann immer 1, 2, 3 versioniert und ja, das ist ein bisschen chaotisch gewesen, wir hatten da kein vernünftiges System. Und das ist auch der Grund hinter dem Startup ein bisschen gewesen, dass man so ein Benchmarking tab hat, dass man sofort sehen kann, wie gut deine Modelle abschneiden, aber auch das würde ich persönlich nur dann auch als Indikator nutzen. Das ist nur mein Erfahrung.

A3

Ja, alles klar. Ja, sehr interessant. Guter Punkt auch mit dem Benchmarking. Wir haben uns mal auf die Suche begeben im weltweiten Internet und haben sechs verschiedene Kommunikationsformen identifiziert, wie so KI-Modelle irgendwie kommuniziert werden können. Und zwar wären das zum einen natürlich die wissenschaftlichen Publikationen, aus denen ja oft KI-Modelle irgendwie hervorgehen. In dem Fall wäre das das MobileNet V3-Paper, was von einem Google-Team geschrieben wurde. Natürlich sehr viel Text, sehr viel, wie es dazu gekommen ist, aber hier gibt es natürlich auch Benchmarkings ganz viel in solchen Papern. Dann hätten wir die Model Cards, die sind in dem Fall von Google, die auch zum Beispiel sowas ein bisschen kürzer dargestellt beinhalten. Was macht das Netz denn überhaupt? Was sind die Inputs? Wie groß ist das Netz? Und dann hier auch wieder ein paar Benchmark-Ergebnisse, wie schnell es ist, unter welchen Umgebungsvariablen diese Benchmarking-Tests durchgeführt wurden. Dann haben wir Papers with Code. Ich weiß nicht, ob du das schon mal benutzt hast.

I7

Natürlich, ja.

A3

Ja, super, okay. Gut, dann kennst du das ja quasi alles, was MobileNet V3 zitiert, wenn man hier jetzt runter scrollt, findet man da eine super Auflistung, sowas wie towards Data science aber auch Medium, einfach diese Blockartikel, die mittlerweile von echt vielen Leuten benutzt werden, wenn sie gerade nicht weiterwissen, ja ein bisschen weniger wissenschaftlich, bisschen mehr journalistisch fast schon. Dann natürlich die Python Dokumentationen, die auch wirklich spezifisch angeben, welche Parameter eben gesetzt werden sollen und slags/flags(?), wenn man eben so ein MobileNet V3 eben benutzt. Und zu guter Letzt haben wir noch die Factsheets eigentlich so ein bisschen wie das Pendant von Google nur das hier ist IBM. IBM hat immer eine relativ ausgiebige Auflistung von Informationen über die KI-Modelle. Zum Beispiel haben die hier auch den Punkt Bias den ich auch sehr interessant finde, also wirklich verzerrungen des Modells. Kann auch in die Richtung Diskriminierung gehen durch KI Modelle, fairness. Und ja sehr ausgiebig diese Informationen aber halt eben nur für IBM publizierte Modelle. Ja genau haben wir so identifiziert. Wenn wir das jetzt mal so nebeneinander stellen auch mit dem KI Label, was fällt dir da so spontan ein? Was für Vor und Nachteile siehst du hier im Vergleich?

I7

Ja, es ist natürlich allen voran die Übersichtlichkeit, dass man auf einen Schlag die wichtigsten Aspekte sehen kann. Vor allem, wenn man das Label dann irgendwann kennt, dann würde man ja auch sehen, da kennt man ja die vier Zeichen und dann weiß man in welcher Farbe die vier Zeichen sind, dann muss man nicht mal genau lesen und weiß, wie gut die sind. Also man sieht das ein bisschen schon auf den ersten Blick alles.

A3

sonst irgendwelche Nachteile die du vielleicht auch im KI-Label siehst gegenüber einer anderen Form die du schon kennst mit der du schon gearbeitet hast

I7

Ich muss sagen, die, die ihr jetzt gerade alle aufgeführt habt, die, die ich hauptsächlich immer genutzt habe, waren, also natürlich ins Paper geschaut und geguckt, wie die mit den Benchmarks und den anderen Modellen, ja, wie die sich vergleichen lassen. Aber was ich mit und auch am meisten zum Schluss hin genutzt habe, war dieses Papers with Code, dass ich mal gesehen habe, wer in diesem Benchmark ziemlich gut abschneidet, das Paper mir angeguckt, das zweitbeste Paper mir angeguckt, das drittbeste Paper mir angeguckt. Aber doch tatsächlich sehr viele Papers immer wieder angeguckt, muss ich sagen. Ja, diese Model Sheets, oder wie die hießen, die es bei Hugging Face gibt am Anfang, die muss ich sagen überspringe ich immer. Ich lese die fast gar nicht durch. Weil, wenn es meine Aufgabe löst, dann, das ist das Allerwichtigste, dann mache ich zur Not auch zwei Server hin. Oder dann gucke ich, ob es so wichtig ist, wirklich Realtime zu sein oder jede Sekunde einmal Inferenzen zu bekommen. So drumherum zu bauen, ist eher machbar, wenn es die Aufgabe zumindest lösen kann. Aber alles in allem muss man sagen, natürlich ist irgendwie klar, dass das Rechte, also euer Label, ziemlich übersichtlich dargestellt ist und auf einen Blick wirklich das zeigt, was für einige Leute relevant sein kann.

A3

Obwohl du ja dann eigentlich der typ bist der so richtig sich die paper durchliest das ist ja auch nicht selbst macht ja auch nicht jeder genauso denn das würde für mich bedeuten du bist ja eher so ein Mensch der dann richtig detailwissen haben möchte aber dann aber du siehst trotzdem einen Vorteil darin

I7

Also durchlesen tue ich mir die Paper nicht, also wenn, dann überfliegen.

A3

Ja gut, keiner liest Paper so richtig.

I7

Wollte ich gerade sagen.

A3

Okay, verstehe.

A1

Das ist eine klassische Aussage. First year PhD-Student: keiner Least Paper. Finde ich sehr gut XXXX, sag das nicht deinem Doktor-Vater.

A3

Nein, auf keinen Fall.

A1

Schöner Money Quote von unserem Interview.

A3

Oder?

I7

Nee, aber also es ist halt so, ich glaube, wie Paper aufgebaut ist und wie man Paper liest oder wie ich Paper lese, gerade für ein Modell, ist halt ganz anders. Ich schaue mir sofort die visuellen Resultate an, dann wie sich das im Benchmarking vergleicht, scroll ich hoch, wenn es ein Appendix gibt, der sehr detailliert ist, den gucke ich mir am mitunter meisten an und dann, wenn ich denke, okay, wie haben die das denn jetzt gemacht, dann versuche ich mir den Abstrakt kurz durchzulesen, um zu gucken, dann, was haben die, worauf baut das auf, wie haben die das ungefähr gemacht und wenn das noch spannend ist, dann gehe ich wirklich rein und schaue mir in der Methodik, was die jetzt anders gemacht haben, welche Module die anders genutzt haben und ja, das wäre so die Reihenfolge, wie die jetzt den Versuchsaufbau gemacht haben und welche quantitativen Resultate genau am Ende rauskamen und wie die das gemessen haben, das interessiert mich meistens gar nicht. Dafür habe ich die Zeit wirklich nicht.

A3

Ja, verständlicherweise. Okay, ja, sehr gut. Dann habe ich glaube ich jetzt verstanden, was zum Beispiel so ein KI-Label für dich an Mehrwert bringen könnte. Dann wäre noch eine, ja, ein bisschen auch wieder glaskugelmäßige Frage, ja, wer sollte denn ein solches Label ausstellen? Kannst du dir das vorstellen, wie das irgendwie in der realen Welt funktionieren könnte?

I7

Also das, was XXXXXXX gerade gesagt hat, hat mich sehr überzeugt, dass man sagt, das ist so ein bisschen Open Source mäßig das heißt wenn ich jetzt mal die Idee einfach weiterspinne, ich weiß nicht ob es genau das ist, was gesagt werden sollte, aber wenn man so eine Software Suite bereitstellen kann, dass das reproduzierbar, nachweislich das Label besitzt, wie man dann auch angibt, dass es das Label besitzt das wäre die einzige Sache, die ich akzeptieren würde. Alles andere was nicht reproduzierbar ist selbst und alles andere was eine Entität ausstellt und man muss der jetzt glauben das würde ich nicht… vielleicht würde man es am Ende trotzdem benutzen wenn es wirklich sehr gut ist, aber würde ich so nicht begrüßen, glaube ich also

A3

Auch nicht sowas wie ein TÜV, das haben ganz viele schon genannt, so eine unabhängige Prüfsache, die auch andere Sachen ja heute schon testet oder eine Behörde.

I7

Ich muss ehrlich sagen so wie ich die Community aktuell einschätze, die ganze AI-Community, die Leute sind ja stolz auf ihre Modelle und stolz auf ihre GitHub-Seiten und stolz auf ihre was auch immer. Die werden das Label selber generieren wollen, denke ich mal. Was man definitiv machen kann, ist natürlich, man kann so eine Stelle anbieten, wo man da so ein Docker-Container-mäßig sein Modell hochladen kann und der testet das da für dich durch und produziert dann das Label, nachweislich diese Instanz hat das Label authentifiziert dargestellt, weil man dann auch diese Authentifizierung dabei hinbekommt, dann würde es wieder Sinn ergeben, so eine Instanz zu haben und zu sehen, wer hat das denn authentifiziert, statt dass irgendein random Typ im Internet das für sich selbst ausgestellt hat. Ich würde wahrscheinlich, sofern ich noch nicht den Fall gesehen habe, dass da jemand mit betrügt, auch dem Label so glauben, natürlich würde ich am Ende eh testen, aber ich glaube, dieses Authentifizierungs-Ding, was ich zum Ende nur, was mir so ein bisschen eingefallen ist, während ich geredet habe, hat auch nochmal einen anderen Mehrwert, glaube ich. Ja, so eine TÜV-Stelle, ja. Kann schon sein.

A3

Okay, ja auch, also ich höre das zum ersten Mal, diesen Vorschlag, ich finde das auch sehr schlau, also wirklich das als Docker-Container zu machen und dann authentifiziert sich dieses Label einfach aber selber generieren zu lassen, ja, ohne dass man da jetzt unbedingt noch ein TÜV zum Beispiel braucht. Was viele auch gesagt haben, ist, dass ein TÜV gar nicht dahinter herkommen würde, so schnell wie KI-Modelle eben weiterentwickelt werden, überhaupt entwickelt werden.

I7

Ja klar, also das ist der Grund, warum ich das mir gar nicht erst vorstellen konnte, muss ich ehrlich sagen. Also das ist eine Sache, die muss schnell gehen, darüber hinaus muss die auch in jeder Version, also ich würde so eine Sache in eine CI/CD-Pipeline integrieren, das heißt, wenn das Modell irgendwie weiterentwickelt wird und dann nochmal trainiert wird, dass das dann automatisch so dieses TÜV Zertifikat oder was auch immer an dieser Webseite verifiziert, anstoßt, dass sich das dann baut, das Label generiert und das neueste Label mit auf dem Readme angezeigt wird. Also alles vollautomatisch, dass da jetzt ein TÜV das händisch macht, so ein Team, das kann ich mir beim Besten nicht vorstellen.

A1

Ja, das finde ich tatsächlich, das finde ich einen total guten Punkt. Ich habe das irgendwie auch schon mal überlegt gehabt, dass man natürlich, so wie man irgendwie Code-Doku und in Keras hast du ja auch Tabellen, wo drinsteht, wie gut ist das Modell auf der und der Hardware, das heißt, ich gehe davon aus, dass das tatsächlich nicht CI-technisch, also klar, die Doku wird CI-technisch erzeugt, aber ich glaube nicht, dass die irgendwie CI-mäßig im Hintergrund die Modelle, die Weights, die die hochladen, durchtesten, sondern das wird irgendwer mal aus dem Paper rauskopiert haben, händisch in die Doku eingetragen haben und jetzt landet das im Readme. Und wenn dann halt irgendein Developer aus Versehen die Weights resettet und das trotzdem auf den Master pusht, weiß ich nicht, ob die Software-Tests haben, die das abfangen und gegenchecken. Es wäre tatsächlich natürlich super cool, wenn man wirklich an einem konkreten Modell arbeitet, darüber ein Paper schreibt oder das halt irgendwie bereitstellt, dazu ein GitHub anlegt, dass dann einfach für jedes Modell, was in dem GitHub-Repo ganz oft so efficientNet oder so, hat man ja auch verschiedene Versionen, verschiedene Größen von dem Modell und so. Wenn dann solche Label generiert werden würden, wir haben das auf der und der Hardware durchgetestet, da verbraucht es die und die Laufzeit, hat die und die Benchmark-Ergebnisse, wäre an sich natürlich total cool. Also...

I7

Also was ich mir bei diesem Label wünschen würde, definitiv wäre, ähnlich gut designt noch ein… Also das ist das Hauptlabel sage ich mal und auf der rechten Seite hast du dann noch so ein größeres Nebenlabel für die, die dann Interesse haben, dass man noch andere relevante Informationen aufzeigt, vielleicht vergleichbare Modelle, wie die im Gegensatz dazu abschneiden etc. Also so ein ergänzendes Label daneben. Also wenn es das nicht gibt, dann muss es in einer anderen Form darunter gemacht werden von dem Autor des Modells in irgendeiner Art und Weise. Ich denke, wenn man sowieso so ein Label ausstellt, dann liegt das auch ziemlich nah, ein Übersichts-Label zu haben und rechts ein detailliertes, größeres Label mit verschiedenen Graphen, die relevant sein können. Ich sage mal, am Ende des Tages, das ist jetzt ich sage mal eine allgemeine Meinung, das wird man wahrscheinlich immer diese Nischen Sachen niemals abdecken können. Da müsste man so ein Riesen-Label mit allem Erdenklichen. Und wir haben ja auch unsere eigenen Metriken in dem Unternehmen gehabt, die wir selbst konstruiert haben. Das wird in so einem Label niemals dargestellt werden können, auf so eine Art und Weise allgemein. Dass das aber allgemein gut nutzbar sein kann, als Übersicht, auf einen Blick zu sehen, oh, das ist schnell und ist A-Gerated allgemein, so beispielsweise, ist schon interessant auf jeden Fall.

A1

Ja, gerade wo du sagst, dass das halt extremst individuell ist, die Needs, die Metriken, die ganzen verschiedenen Lernaufgaben, es gibt ja inzwischen auch, also wir reden ja hier die ganze Zeit nur über einen Bildklassifikator, es gibt natürlich unfassbar vielseitige Lernaufgaben und entsprechend Modelle und Lösungen dafür. Aber da wäre natürlich der große Vorteil, wenn das nicht eine unabhängige Third-Party-Instanz macht, sondern dass er so einen Open-Source-Gedanken hat, dass halt jeder sich die Software schnappen kann, die Metriken und die Icons, die für ihn relevant sind, irgendwie bereitstellen kann und dann ist das mehr so ein grobes Framework mit ein paar Beispielen, wie es zu benutzen ist, aber man kann das dann für die eigene Needs halt anpassen. Das wäre halt der Vorteil. Wenn dann eine Firma sagt, wir bauen superspezifische Modelle für den und den Anwendungsfall, können die sich trotzdem solche Label generieren.

I7

Jetzt wo du das so sagst, was super spannend wäre, wäre, wenn man generell diese Datenbank an Labels hätte und die sind jetzt, sagen wir mal, durchgetestet, es gibt dieses allgemeine Label und wenn man dann irgendwie, wenn das so ein JavaScript-Widget wäre oder was auch immer, wenn man seine Konfiguration, die für jemanden relevant ist und die wird dann encoded in einen kleinen String oder so, wenn man die einfach da in dieses Inputfeld reinmachen kann und wenn man Enter drückt und dann baut der dir automatisch diese Übersicht, die für dich relevant wäre. Oder wenn du eine Tabelle hättest mit 100 Modellen, die verglichen werden, wenn du da deine Konfiguration oben reinmachst, entweder, wie gesagt, als encoded String und dann macht dir direkt das Interface so, dass du alle Labels deiner Ansicht relevant dargestellt bekommst, das wäre super spannend. Das würde so eine Sache sein, wo ich mir auch vorstellen kann, dass man das auch intensiv nutzen würde, also ich das intensiv nutzen würde, wenn ich schon alles durchgetestete an Labels so umkonfiguriert bekomme, dass die, ja, für mich relevant wären, die ganze Information. Auch da natürlich muss man sagen, es gibt Metriken, die müssen einfach mal durchgerechnet werden, die kann man so nicht anzeigen in einem vorgenerierten Ding, aber das wäre schon mal ein richtig guter Schritt, glaube ich, in die richtige Richtung. Ich weiß nicht, ob das klar wurde, was ich meinte mit dem konfiguriertem String und dann hat man diese Labels für sich konfiguriert. Ich habe mir das ein bisschen vor Augen vorgestellt wie das aussehen kann.

A1

Ich glaube, ich weiß, wie du das grob meinst. Tatsächlich, wo du auch ganz so sagst, so weitergehende Informationen wären total praktisch. Ich schicke mal gerade einen Link hier rein. Das ist auch Work in Progress gerade. Ist jetzt ein anderer Anwendungsfall. Geht jetzt hier um Forecasting-Modelle. Aber ich habe halt auch an so einem Tool gearbeitet, was einem das so interaktiv visualisiert. Und hier ist jetzt zum Beispiel Running Time versus eine bestimmte Fehlermetrik. Genau, XXXX kann das mal eben kurz aufmachen. Das ist noch einfacher. Wenn du jetzt über einen konkreten Point hier hoverst, in diesem Scatterplot, dann generiert der dir halt eine tabellarische Zusammenfassung und das passende Label dafür. Das braucht ein bisschen lange, weil das ist kostenlose Renderinstanz und irgendwie 0,1 CPU und so.

I7

Ist das sowas wie… I st das Dash?

A1

Ja, das ist Dash Plotly. Du kannst dann halt irgendwie das Label dir in groß natürlich auch anschauen. Das ist jetzt nicht perfekt gepolished. Das ist auch noch die alte Energy Label Skala, die jetzt hier ist. Das sind eher so Energie Label, angelehnt an die Energie Label aus der EU. Aber im Endeffekt die gleiche Idee. Und dann den Plot, diesen 2D-Sketter-Plot, den man sieht, da kann man tatsächlich dann halt auch customisen, was möchte man auf der X- und der Y-Achse haben und so. Und dann hast du halt genau diesen Vergleich zwischen verschiedenen Modellen zum Beispiel, hast du halt genau darüber auch gegeben. Und du hast die Properties, die ganzen Metriken nochmal auch als detaillierte Tabelle mit den echten Messwerten und mit den relativen Scorings und dem Rating, was dann am Ende farbcodiert wird. Und auch der Gewichtung für das Gesamtrating. Das sieht man halt alles da. Das ist auch Teil dieser Open-Source-Library, an der ich arbeite.

I7

Das ist natürlich super spannend, wenn man so ein Label hat und man scannt das, so wie da oben in der QR-Code, man kommt auf so eine Seite, wo man wirklich genau sieht, wie dieses Label zustande kommt. Das ist super spannend. Und wie gesagt, wenn man dann noch, ohne viel rumgucken zu müssen, irgendwie seine Konfiguration da schnell reinbekommt, dass man sofort die Ansicht bekommt, die jemandem relevant ist eigentlich.

A1

Also die Konfiguration, damit meinst du jetzt konkret?

I7

Ach so, ich habe das irgendwie intern in meinem Kopf zu Ende, so ein paar Schritte gedacht, also vielleicht kommuniziere ich noch zu Ende, also ich stelle mir vor, du hast so eine Seite, wo du wirklich die ganzen detaillierten Ansichten hast und dann kannst du jetzt mit verschiedenen Schiebern und Reglern und hin und her schieben, dir das Dashboard und die Ansicht so zurechtlegen, dass genau die Sachen angezeigt werden, die für dich relevant sind. Du hast so Sachen ausblenden kannst, die nicht wichtig sind, Sachen, die gerade

A1

Wir reden bei die Sachen, die Sachen, damit meinst du zum Beispiel diese ganzen verschiedenen Metriken und du sagst das

I7

Genau, genau, genau. Metriken, auch vielleicht Vergleiche mit anderen Modellen, die ähnlich sind. Genau, solche Sachen, genau. Und dann die Ansicht, wie man das dargestellt bekommt, etc. Wenn das so eine Software-Library wäre zum Beispiel und ich kriege einfach den Link vom Label rein und ich kann mir mein eigenes Dashboard bauen, das wäre so eine Sache. Oder das andere wäre, wenn man sich das dann einmal auf der Webseite zurechtgerückt hat, dass dann die Webseite so diese Ansicht-Teilen-Button hat und dann hast du so einen Code, der da automatisch generiert wird. Und wenn du den dann irgendwo anders bei einem anderen Modell eingibst, dass er automatisch wieder diese Konfiguration dir zeigt, sodass du immer in dieser Konfiguration dir deine Modelle angucken kannst beispielsweise. Ich weiß nicht, ob das klar wird, was ich meine. Ja.

A1

Es wird klar, was du meinst. In meinem Kopf explodiert direkt Entwicklungsaufwand. Aber an sich ist das natürlich ... Also wir wollen nicht drüber reden, dass das technisch möglich ist. Es ist nur dann wieder, das geht dann halt wieder Richtung Produkt, was halt für meine Promotion bis nächstes Jahr nicht unbedingt die oberste Prio ist. Aber klar, an sich wäre das cool. Ich habe auch schon mal überlegt, was richtig geil wäre. Da ist tatsächlich so ein Button dafür. Wenn jemand jetzt dieses Label hat, wäre natürlich auch mega cool, wenn man daraus jetzt automatisch so eine Modelcard als PDF generieren könnte. So auf zwei Seiten, Beschreibung des Modells, wichtige Benchmarks, Vergleiche und so. Dass man das alles end-to-end miteinander verlinkt. Aber das kollidiert alles in dem Moment, wo du mit diesem Framework möglichst viele verschiedene Modelle und Lernaufgaben und alles supporten willst. Also das erzeugt einfach so eine, also so eine Modelcard ist ja super abhängig davon, wie, also welches Problem du überhaupt lösen willst. Und ja, es ist halt nicht so einfach, das so wirklich zu generalisieren. Also das ist jetzt soweit generalisiert. Den interessiert, der weiß halt auch nicht, was die, was die Semantik hinter Training Time ist. Für den ist das nur irgendeine Zahl und die heißt halt Training Time. Und dann hast du noch so Späße, dass halt irgendwie manche Metriken auch irgendwie anders gebounded sind. Das heißt, du müsstest irgendwie, ne, Accuracy willst du maximieren, Running Time willst du minimieren. Das ist jetzt der einfachste Fall. Dann gibt es aber auch Metriken, wo Zero zum Beispiel kein zulässiger Wert ist. Du kannst halt Zero Accuracy haben, aber wenn du Zero Runtime hast, dann musst du da eigentlich auch wieder Catches drum machen, weil es ist halt einfach, dann ist irgendwas wirklich ganz kaputt und so. Und also, dann hast du irgendwie Metriken, die skaliert sind. Also es ist nicht ganz, alles nicht ganz trivial, das zu vereinheitlichen.

I7

Also wie gesagt, ich glaube insgesamt das Fazit dazu, wenn ich mir das alles angeguckt habe, ich wiederhole ja immer wieder diesen Punkt, dass es sehr individuell immer auf die Nische zugestimmt etc. etc. Aber wirklich den größten Mehrwert nach wie vor sich darin, dass man das Kunden vorlegen kann, dass man denen sagen kann, guck mal, das waren ja wichtige Punkte, wir haben ja das Label so konfiguriert, das sind jetzt die wichtigen Punkte für dich gewesen. Und das Modell, was wir dir vorschlagen, das ist jetzt A geratet. Das Modell, wenn wir das mit einem LLM lösen beispielsweise, dann wäre das hier energietechnisch katastrophal und das wäre dir ja sehr wichtig und dementsprechend ist das eine D oder eine E Lösung. Also das wäre, das ist ja ein richtig krasser Mehrwert. Also in der Kommunikation mit Leuten, die wirklich wenig Ahnung haben und in der Kommunikation mit vor allem Kunden. Und wenn man dann noch irgendwie das hinbekommt. Das ist ja auch sehr viel in Richtung, wie soll ich sagen, Carbon Emission und etc. das ist ja auch sehr wichtig, diese Nachhaltigkeit und ja, dass man das den Unternehmen auch ermöglicht zu sagen, hey, ihr könnt jetzt ein Label auf eure KI hauen, wo es auch zeigt, dass das der Energieeffizienz ist, was ihr da gebaut habt, das Produkt, was von den Kunden genutzt wird. Das könnte für die auch sehr relevant sein. Also muss man halt gucken . Aber das wäre dann ein bisschen wieder was anderes. Das wäre dann nur Energie und Nachhaltigkeit etc.

A1

Ja aber das ist ja auch ein großer Bereich gerade. Also ist auch einer der Sachen, sieht man ja, an denen ich sehr interessiert bin wie man halt irgendwie Machiene learning auch ressourcenbewusst machen kann. Das muss nicht unbedingt edge computing sein und ultra low power devices aber zumindest diese awareness dafür zuhaben wie viel Energie wird hier verbraucht und gerade jetzt wo wir LLMs haben und teilweise auch Bildmodelle und so, wo so ganz viel davon wegabstrahiert wird, wo du das Zeug ja nichtmal auf einer eigenen Hardware laufen lässt sondern alles as a Service angemietet, da fällt auch ganz viel unter den Tisch. Und am Ende sind natürlich Firmen eigentlich daran interessiert die wollen unten einmal ein rotes Dollarzeichen wie viel kostet mich das Modell laufen zu lassen? Und ein grünes für wie viel Geld gibt mir mein Modell also wie viel Geld kann man damit machen? Das sind eigentlich die Metriken die für die am relevantesten sind. Aber für die Leute die das umsetzen müssen, also Entwickler und Ingenieure und insbesondere Leute die vielleicht nicht 10 Jahre Machiene Learning und AI Hintergrund haben, für die ist es halt super schwer dieses Profit-Cost-Ding umzumappen auf Metriken von Machiene learning Modellen

I7

Ja ne bin ich voll bei dir

A3

Sehr cool. Ja gut, da hast du eigentlich meine letzte Frage auch schon quasi beantwortet. Ich würde trotzdem gerne nochmal fragen, wie sinnvoll findest du denn eine allgemeine Zertifizierung? Du meintest, dass du dir vorstellen kannst, auf jeden Fall Label irgendwie einzusetzen unter gewissen Voraussetzungen. Wie findest du denn die allgemeine Zertifizierung, wenn das jetzt so eine Pflicht quasi wäre oder ein Standard wäre?

I7

immer bei pflichten sehr sehr sehr vorsichtig weil also wenn man das irgendwie machen möchte unbedingt da muss das so unglaublich unglaublich leicht sein wir haben jetzt schon genug pflichten und jetzt mit dem AI-Act der dazu kommt haben auch viele Panik wie genau das laufen soll in Zukunft. Da haben wir jetzt Workshops zu gemacht also alles was an Regulierungen und so weiter eingeführt wird die sind natürlich wichtig muss gucken dass es seitdem ich angefangen habe ein bisschen auch unternehmerisch tätig zu werden habe ich auch dieses mindset der Unternehmer ein bisschen verstanden es ist so schon super super schwer allen auflagen gerecht zu werden und gleichzeitig dieses Startup auf die Beine zu ziehen also mit Anwälten und Notaren zu reden und dann sich NDAs(?) durch zu lesen auch noch mal Anwälten angucken zu lassen dann sich Verträge durchzusammeln(?), dann dein Impressum muss stimmen, deine Datenschutz Erklärung muss stimmen auf der Webseite und und und und und wenn dann noch ein Label dazukommt für die AI, und ich sage mal du machst viele AI Modelle und das ist auch noch ein bisschen aufwendig das zertifiziert zu bekommen also ich bin gerade dafür dass man mehr also Regularisierungen in dieser Hinsicht die unnötig ist für das Startup abbaut und nicht noch da Sachen auf… Da werde ich immer ein bisschen hellhörig muss man gucken wie die Umsetzung aussieht muss ich ehrlich sagen.

A3

Ja, also eher in Richtung freiwillige Selbstverpflichtung, also sowas wie der Nutri-Score ja auch ist. Da muss ja auch eine Firma nicht unbedingt den Score sich aufdrucken lassen auf die Müsli-Packung oder was. Könnte das ein Weg sein für dich?

I7

Also, ich sag mal so, wenn das eine sinnvolle, also wenn das am Ende zu so einer Sache wird, die echt sinnvoll ist und man kriegt das hin, dass da wirklich die Leute ihren Mehrwert von ziehen, wenn man das freiwillig macht, dann werden das glaube ich auch viele Leute freiwillig machen. Wie gesagt, jemand, der ein KI-Modell entwickelt, dann wird das so eine Standardsache sein, die er da hinzupacken möchte, damit er noch mehr das verkauft bekommt, in Anführungszeichen. Jetzt Nutri-Score finde ich einen schlechten Vergleich, weil für mich ist das halt so eine Sache, die schaue ich nie und die ist für mich komplett, ja gut, das ist meine persönliche Meinung, eine der schlechtesten Metriken, die ich in meinem Leben gesehen habe, die in der Praxis genutzt werden. Also wirklich, naja gut. Das ist meine persönliche Meinung.

A3

Okay, vielen Dank.

I7

Also für mich ist der Nutri-Score absolut irrelevant.

A3

Alles klar, gut. Dann hätte ich auch tatsächlich keine weiteren Fragen mehr. Ich weiß nicht, XXXXXX XXXXXXX, habt ihr noch irgendwas, was euch am Herzen liegt?

A1

Nö, ich fand's sehr cool. Vielen Dank für die Einblicke.

I7

Ach, gern geschehen. Also aus Interesse, muss ich ehrlich sagen, habe ich das gemacht.

A1

Ja, aber es ist ja genau, du bist ja auch genau an der Schnittstelle, die halt interessant für uns ist Ich meine, ich habe auch noch ein Industrieprojekt, wo wir nebenbei arbeiten und es ist halt tatsächlich gerade sehr spannend, wie auch dieses ganze KI-Feld so ein bisschen revolutioniert wird durch LLMs Das heißt, wir haben mit vielen Leuten gesprochen, für die KI einfach 1 zu 1 ein Chatmodell ist Gleichzeitig haben wir auch mit Leuten gesprochen, für die KI gar nicht Deep Learning ist, sondern für die halt, die halt wirklich so einen KI-Transfer machen, für irgendwelche Leute die, weiß ich nicht, teilweise noch eine lineare Regression benutzen wollen, weil die haben so einen, weiß ich nicht, 600 Zeilen Industriedatensatz irgendwie und wollen darauf irgendwie jetzt intelligente Sachen machen, also intelligente Sachen. Im weitesten Sinne. Es ist halt sehr unterschiedlich. Und deswegen ist das ganz cool mit der Studie, kriegt man wirklich so sehr viel Einsicht, was davon in der Praxis wirklich relevant ist. Dafür machen wir das.

I7

Ja. Ich seh grad, du hast kein, äh, also du benutzt exklusiv Plotly, gar kein Matplotlib oder sonst was anderes, ne?

A1

Ich mach gar nichts mehr mit Matplotlib. Alles interaktiv. Naja, ich hab jetzt in letzter Zeit oft angefangen, in Papern immer direkt so ein Dashboard irgendwie mit dazu zu packen, damit man sich das, der Gedanke geht ein bisschen, aber ich glaub, wir können die Aufnahme beenden, ne?