

Wo war das nochmal?

KI errät Geodaten zum Foto

Geolokalisierung, die Ortsbestimmung zu einem Foto ohne Metadaten – Google erforscht und nutzt sie hinter den Kulissen und ein Leibniz-Institut in Hannover hat ebenfalls eine Lösung entwickelt. Sein KI-System kann jetzt öffentlich ausprobiert werden.

Von Arne Grävemeyer

Ein Foto ohne Metadaten, wo wurde es wohl geschossen? Eventuell sind auf dem Bild markante Gebäude oder Bergkämme zu erkennen, aber nur wenige Globetrotter werden den Ausschnitt damit einem Ort zuweisen können. Auch die Websuche nach Bildhinweisen ist mühsam und misslingt oft. Wer eine Privataufnahme nicht mehr einer bestimmten Reise zuordnen kann oder wer glaubt, eine Fake News über ein Foto entlarven zu können, für den gibt es mit Geolokalisierung durch künstliche Intelligenz ein schnelles Hilfsmittel. Etwas ehrlicher ist allerdings die Bezeichnung Geo-Estimation, übersetzt also Geografie-Schätzung.

An einem Mittwoch im September kam es auf der European Conference on Computer Vision (ECCV) 2018 in Mün-

chen zum Showdown. Unabhängig voneinander präsentierten zwei Forschungsgruppen ihre KI-Systeme zur Foto-Geolokalisierung. Auf der einen Seite der Datenriesen Google und auf der anderen Hannoveraner Wissenschaftler der Leibniz-Gemeinschaft.

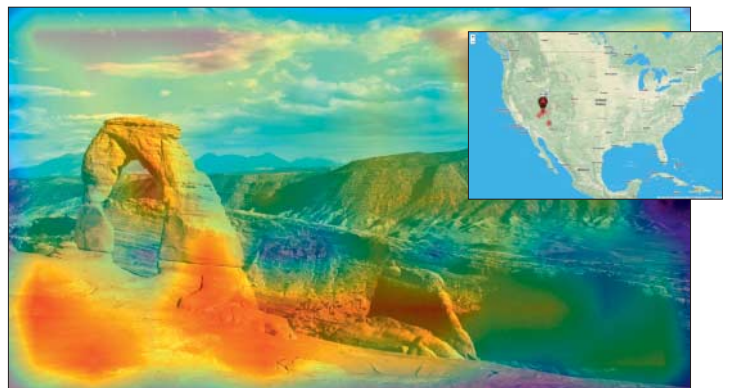
Den 2016 entwickelten PlaNet-Ansatz aus dem Google-Konzern hat ein Team um Tobias Weyand und Paul Hong-suck Seo verfeinert und als CPlaNet neu vorgestellt. Bei beiden Ansätzen klassifiziert ein neuronales Netz die eingehenden Fotos auf die Maschen eines Netzes, das um den gesamten Erdball gespannt wird. Wurde der erste Ansatz noch mit sagenhaften 91 Millionen Fotos trainiert, so verringerten die Forscher beim CPlaNet-Ansatz die Zahl der Trainingsbilder auf etwa 30 Millionen. Das Google-Tool ist offenbar für den internen Einsatz bestimmt, über seine Aufgaben innerhalb der Google-Produkte verrät der Konzern auf Anfrage nichts. Wird es einen Zugang für Privatanutzer geben? Kein Kommentar.

Der Herausforderer kommt aus Hannover. Professor Ralph Ewerth leitet die Forschungsgruppe Visual Analytics am Leibniz-Informationszentrum Technik und Naturwissenschaften (TIB – Technische Informationsbibliothek). Die TIB dient als zentrale Fachbibliothek für Technik, betreibt aber auch eigene Forschung,

etwa zusammen mit dem Forschungszentrum L3S der Leibniz Universität Hannover. Zwei Promovierende aus der Gruppe Visual Analytics, Eric Müller-Budack und Kader Pustu-Iren, haben eine eigene künstliche Intelligenz zur Geolokalisierung entwickelt. Auch bei diesem Ansatz wird allein aus der Bildinformation ein Tipp auf den Aufnahmeort abgeleitet. Die TIB-KI ist mit nur etwa fünf Millionen Fotos trainiert worden, danach gelang ihr die Überraschung: Das System hat die Google-Konkurrenz beim Benchmark mit Standard-Testdaten übertrumpft.

Lokalisierer jetzt im Web

Seit Ende Januar kann das Geolokisierungstool der TIB mit beliebigen Fotos ausprobiert werden (siehe ct.de/yn7a). Dabei wird auch deutlich, wie die Software vorgeht. „Unser Ansatz entwickelt die erste Google-Lösung PlaNet weiter. Auch unser Tool teilt die Erdoberfläche in Parzellen ein und nimmt für jedes Foto eine Klassifizierung auf diese Zellen vor“, berichtet Müller-Budack. Allerdings haben die Forscher eine wichtige Verfeinerung vorgenommen: Ein Filter unterteilt die eingehenden Fotos gleich zu Beginn in Stadtszenen, Innenaufnahmen oder Naturbilder. Damit gliedert sich das Tool zur Geolokalisierung in vier spezialisierte Subnetze, nämlich den Vorfilter sowie drei



Nicht hundertprozentig lokalisiert: Der Grand Canyon liegt, wie auf der Karte angezeigt, größtenteils in Arizona. Der imposante Delicate Arch auf dem Foto steht allerdings doch weiter westlich in Utah.



Hamburg, Speicherstadt: Die Class Activation Map (rechts) markiert die für die Klassifizierung entscheidenden Bildelemente. Beim Zoom in die Ergebniskarte zeigt sich, dass die KI diese Ansicht sogar auf zwei Straßenzüge genau lokalisieren kann.

separate Schätzer für entweder Stadt-, Innen- oder Naturaufnahmen. „Am stärksten sind neuronale Netze, wenn sie auf eine Aufgabe spezialisiert sind“, betont Müller-Budack. Bei sich überschneidenden Aufgabengebieten für ein entsprechend größeres neuronales Netz wären hingegen schwächere Ergebnisse zu erwarten.

Beim öffentlichen Einsatz im Web lädt die TIB-Geolokalisierung zunächst zum Ratespiel ein. Der Besucher der Webseite kann nicht nur eigene Fotos hochladen und einschätzen lassen, er kann auch persönlich gegen die KI antreten. Zu dem Zweck stehen derzeit 48 Fotos zur Auswahl, die selbstverständlich nicht Bestandteil des Trainingssets der KI waren. Der Besucher kann sich eines aussuchen und dazu dessen Herkunft auf der Weltkarte schätzen. Danach versucht es der Web-Dienst selbst. Am Ende werden die Differenzkilometer zum tatsächlichen Aufnahmeort verglichen. Ein schöner Partyspaß, der nach 48 Versuchen allerdings zu Ende ist.

Krokodile in London

Schon dabei zeigt sich, dass die KI starke Momente hat, aber auch ganz schwache Versuche. Einmal wird eine Brücke in Venedig exakt erkannt und das nächste Mal eine Landschaft im Umland von Rom an die spanische Atlantikküste verlegt. Beim Foto eines Krokodils aus Südostasien tippte die KI der Hannoveraner gar auf London; allerdings nicht wegen der aktuellen Debatten im Unterhaus, sondern weil es im Norden des Regent's Parks einen gut sortierten Zoo gibt. Ewerth erläutert, dass für die Trainingsphase des zugrundeliegenden neuronalen Netzes etwa fünf Millionen flickr-Fotos mit eindeutigen Geo-

tags genutzt worden sind. „Dabei hatten wir eine höhere Abdeckung in Europa und Nordamerika und eine geringere Abdeckung etwa afrikanischer oder südostasiatischer Landstriche.“

Die TIB-Webseite bietet Einblick in die Entscheidungsvorgänge der KI. Die sogenannte Scene Classification gibt einen Score-Wert aus, der in der Regel mit klarem Votum bestimmt, ob das Foto eine Stadtszene, Natur oder eine Indoor-Aufnahme zeigt. Noch informativer ist die Class Activation Map, also eine Heatmap, die auf dem Eingangsfoto farblich markiert, welche Bildbereiche für die Klassifizierung relevant waren.

Und das geht so: Sowohl das Tool der TIB als auch das von Google setzen für die Foto-Klassifikation auf sogenannte Convolutional Neural Networks (CNN), eine Form von neuronalen Netzen, die schon bei vielen Projekten in der Erkennung von Bildinhalten ihre Stärken bewiesen haben. CNN heben in den ersten Verarbeitungsschichten zunächst einfache Bildelemente wie Linien und Kanten hervor. Darauf aufbauend werden dann in tieferen Schichten komplexere Strukturen in die Klassifikation einbezogen. Um die Entscheidungsfindung eines derartigen neuronalen Netzes nachträglich zu erklären, werden sogenannte Class Activation Maps zu einer speziellen Fotoklassifizierung berechnet. Im Prinzip wird die letzte Schicht eines neuronalen Netzes bei dieser Methode rückwärts durchlaufen und überprüft, welche Bildbereiche die typischsten Merkmale einer bestimmten Region enthalten. Ein einfacher Farbcode von kaltem Blau für wenig Einfluss über Gelb bis leuchtend Rot für große Auswirkungen kann dann das Foto überdecken. So erkennt der Betrachter schnell, welches Zwiebelturm-

chen, welche Fensterform, welche Einrichtungsgegenstände oder welche Pflanzenart letztlich für die getroffene Lokalisierungsentscheidung ausschlaggebend war.

Damit lassen sich Fehlentscheidungen zumindest nachvollziehen. So kann ein ungewöhnliches Ornament an einer Fassade die KI auf eine falsche Fährte locken. Ebenso kann ein Gartenfoto falsch verortet werden, wenn dort eine ortsfremde Pflanze zu sehen ist. Die Heatmap erklärt dann wenigstens, welche Merkmale zur Irritation führten. Ganz generell gelingt die Lokalisierung am besten bei Stadtsichten, die Analyse von Innen- und Naturaufnahmen liefert nicht so gute Trefferquoten.

Test bringt Überraschungen

Beim c't-Test mit eindeutigen Ortsaufnahmen zeigten sich genau diese Probleme recht klar. Eine stimmungsvolle Aufnahme aus der Hamburger Speicherstadt wurde von der TIB-KI eindeutig zugeordnet und sogar auf wenige Straßenzüge genau lokalisiert. Aber schon eine auffällige Felsformation am Colorado River, der sogenannte Delicate Arch, ein eigentlich sehr beliebtes Fotomotiv, wird zwar korrekt in der Nähe des Grand Canyon verortet, aber dann doch mit erstaunlicher Bestimmtheit ein paar Hundert Meilen nach Westen verlegt.

Noch krasser ist die Fehleinschätzung einer Aufnahme des japanischen Fuji. Obwohl auf dem Foto sogar eine Pagode abgebildet ist, verwechselt die KI den Vulkan mit Mount Shasta im Norden Kaliforniens – immerhin auch ein Vulkan.

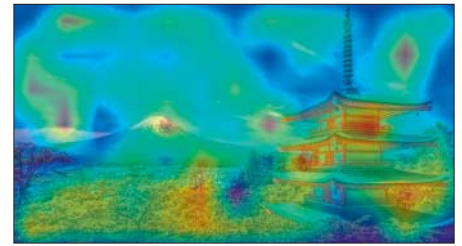
Man muss die Ergebnisse der Geolokalisierung also mit Vorsicht genießen, aber die Fortschritte auf diesem Gebiet

sind faszinierend. Bereits 2008 hatte ein Team um James Hays an der Carnegie Mellon University in Pittsburgh mit Im2GPS (Image to GPS) ein System der Bildvergleiche vorgestellt. Die Forscher hatten dazu sechs Millionen Fotos indiziert. Zu einem Eingangsfoto wurde dann aufgrund seiner Bildelemente das ähnlichste Foto des Ursprungsdatensatzes ermittelt. Die Geodaten des ähnlichsten Fotos bildeten das Ergebnis der Geolokalisierung. Die Forscher erkannten erfreut, dass sie damit etwa 30-fach besser lagen als mit zufälligen Raten.

Immerhin: Der von ihnen genutzte Im2GPS-Testdatensatz gilt Forschern heute noch als Standard. Und die Ergebnisse neuerer Tools zeigen rasante Steigerungen. Konfrontiert mit dem Im2GPS-Testfotosatz ordnete der erste Google-Ansatz PlaNet ganze 8,4 Prozent der Fotos auf einen Kilometer genau zu. Im 25-Kilometer-Radius gelang die Lokalisierung von fast 25 Prozent der Fotos. Hier hatte man also immerhin schon die richtige Stadt getroffen. Maximal 200 Kilometer Abstand zum Aufnahmeort wiesen die Einschätzungen bei 38 Prozent der Fotos auf. Das richtige Land sozusagen, nämlich einen Umkreis von 750 Kilometer traf PlaNet bei 54 Prozent der Fotos. In einem Umkreis von 2500 Kilometer und damit wenigstens auf dem richtigen Kontinent bewegten sich die Treffer bei sieben von zehn. Das ist besser als nichts, aber der Begriff Geo-Estimation trifft es eben doch besser als Geolokalisierung.

Trefferquote verdoppelt

Mit dem nachfolgenden CPlaNet-Ansatz gelang Google 2018 eine klare Steigerung. Jetzt trafen fast doppelt so viele Einschätzungen auf einen Kilometer genau, nämlich 16,5 Prozent. In der richtigen Stadt lagen demnach schon 37 Prozent und im richtigen Land 62 Prozent. Und das, obwohl weit weniger Trainingsbilder genutzt worden waren. Der Unterschied: Der PlaNet-Ansatz klassifiziert Fotos schlicht auf ein Netz mit über 25.000 Feldern auf der Erdoberfläche. Dabei sind die Felder übrigens nicht alle gleich groß, da für jedes Feld eine genügende Anzahl von Trainingsfotos eingesetzt werden muss. Entsprechend sind Innenstadt-Parzellen recht eng geschnitten, ländliche Gegenden eher weitläufig eingefasst und entlegene Gebiete wie Ozeane oder die Polkappen bleiben sogar vollständig ausgespart.



Achtung, es ist nur eine Schätzung: Der Fujiyama wird von der KI offenbar mit Mount Shasta in Kalifornien verwechselt. Die Ergebniskarte zeigt leichtere Aktivierungen auch in Japan.



Beim verfeinerten CPlaNet-Ansatz ist zum einen der Trainingsdatensatz besser gepflegt. Er wurde um zahlreiche Indoor-Fotos und unscharfe Aufnahmen bereinigt sowie auch um einander ähnliche Fotos. Zudem wird hier nun nicht mehr nur ein Netz um den Erdball gelegt. Stattdessen werden gleich fünf unterschiedliche Netze aufgespannt und damit im Grunde fünf Klassifizierer parallel genutzt, jeder mit einem eigenen Fotodatenatz trainiert. Die Maschen dieser Netze sind nicht deckungsgleich, wodurch eine Vielzahl an Schnittmengen entstehen. Am Ende ermittelt die Gesamt-KI für jedes Eingangsfoto im Grunde fünf Klassifizierungsergebnisse. Dort, wo sich die Ergebnisfelder überschneiden, gilt die Lokalisierung als besonders valide.

Vorfilter lohnt sich

Der Paukenschlag gelang dem TIB-Team auf der ECCV 2018 in München, als es seine Testergebnisse parallel zum Google-Forscherteam veröffentlichte. Ihre KI verortete sogar 16,9 Prozent der Fotos aus dem Im2GPS-Testfotosatz auf einen Kilometer genau, 43 Prozent wurden wenigstens der richtigen Stadt zugeordnet und sogar 67 Prozent dem richtigen Land. Der Kniff mit der Scene Selection und der Verteilung auf drei thematisch spezialisierte neuronale Netze hatte sich als überlegen erwiesen.

Die TIB-Wissenschaftler dachten bereits über eine Ausweitung ihres Systems

nach. Zusätzlich zu drei Szenen gibt ein in der Wissenschaft existierender Benchmark einen Katalog von 16 Unterszenen und sogar 365 Kategorien vor. Tag oder Nacht, Straßenleben oder Baudenkmäler, Tierwelt, Pflanzen oder Gesteinsformationen – die Scene Selection ließe sich noch ausbauen. Aber dieser Schritt erscheint den Forschern kaum umsetzbar, zumal dann für jeden nachfolgenden Klassifizierer eine große Menge passender Trainingsfotos gefunden werden müsste.

Stattdessen arbeiten die Forscher an der TIB derzeit an einem Tool, das in kombinierten Beiträgen eine Text-Bild-Schere erkennt und zudem noch eine Personenerkennung mit in die Analyse integriert. Wobei Ewerth betont, dass die KI immer nur einen Anhaltspunkt liefern kann, der Mensch müsse einen Verdachtsfall stets überprüfen.

Das Sympathische an der TIB-Forschung ist, dass ihre Ergebnisse als Open Source im Web zugänglich sind. So steht nicht nur die Geo-Estimation mit Foto-Upload-Möglichkeit im Web zur Verfügung. Alle Modelle der TIB-Forschungsgruppe Visual Analytics sind zudem auf GitHub verlinkt (siehe ct.de/yn7a) und können damit für beliebige weitere Projekte eingesetzt werden.

(agr@ct.de) **ct**

Geolokalisierung im Web und GitHub-Zugriff auf die Modelle der TIB Hannover: ct.de/yn7a