Andrea Trinkwalder

Künstliche Ortskenntnis

Googles neuronales Netz PlaNet erschließt sich die Welt: Es lernt Orte und Landschaften kennen

Neuronale Netze können Hot Spots des Tourismus wie Eiffelturm oder Freiheitsstatue schon fast im Schlaf identifizieren. Jetzt haben Google-Forscher ein neuronales Netz entwickelt, das über die einfache Objekterkennung hinauswächst: Es verortet Landschaften und Stadtszenen anhand ihrer typischen Atmosphäre im passenden Teil der Welt.

Als der Computer lernte, Hund, Katz und Maus zu unterscheiden, war die Welt begeistert. Für die Forscher blieb das trotz aller Euphorie aber immer nur der Anfang. Denn wirklich nützlich wird Bilderkennung erst, wenn der Computer Szenen beschreibt oder die Einzelteile zu einem großen Ganzen zusammenfügt.

Einen Ansatz dafür haben die Google-Forscher Tobias Weyand, Ilya Kostrikov und James Philbin jetzt gefunden [1]: Ihr neuronales Netz namens PlaNet lernt anhand von Trainingsdaten nicht nur, Orte an prominenten Objekten oder Gebäuden wie etwa dem schiefen Turm von Pisa zu erkennen. Es soll sich anhand subtilerer Merkmale auch von Landschaften ein Bild machen, um etwa eine Island- von einer Australien-Szene unterscheiden zu können. Diese Aufgabe ist wesentlich schwieriger als die Objekterkennung und nicht immer eindeutig lösbar, weil bestimmte Landschaften in verschiedenen Ländern oder gar Erdteilen auftreten und in einem Foto auch untypische Merkmale dominieren können. So sind auch außerhalb von London rote Telefonzellen und Doppeldecker-Busse zu finden. Doch ganz wie einem weltgewandten Menschen mit solider Allgemeinbildung soll es PlaNet eines Tages gelingen, aus verräterischen Hinweisen in der Umgebung – etwa Straßenschildern oder Architektur – die richtigen Schlüsse zu ziehen.

Die Gruppe um Weyand behandelt die Lokalisierung als Klassifizierungsproblem: Die Bilder – also die Matrix der Pixelwerte – werden in ein neuronales Netz eingespeist, woraufhin dieses für jede Region auf der Weltkarte eine Wahrscheinlichkeit dafür berechnet, dass das Bild dort aufgenommen wurde. Dabei kommt ein Convolutional Neural Network (CNN) zum Einsatz – eine in der Bild- und Spracherkennung bewährte Architektur [2]. CNNs sind so konstruiert, dass sie nach Training mit Millionen Bildbeispielen genau diejenigen Merkmale in Bildern finden, die für bestimmte Kategorien charakteristisch sind, beispielsweise das Schema eines Menschen, eines Tieres oder einer Pflanze.

Welt-Klassifizierung

Bei der Lokalisierung besteht die erste Herausforderung darin, sinnvolle Kategorien zu definieren. Dafür finden die Forscher im Internet zwar Millionen Fotos mit GPS-Koordinaten, müssen sie aber möglichst automatisiert so zu Regionen zusammenfassen, dass die Maschine sinnvolle Schlüsse daraus ziehen kann. Weyand und seine Kollegen haben sich dafür entschieden, die Welt so aufzuteilen, dass jeder Bereich (entspricht einer Kategorie) eine gleich große Menge an Trainingsbildern enthält. Dicht besiedelte Regionen bestehen also aus sehr vielen kleinen Zellen, während ländlichere Gegenden weiträumig abgesteckt werden.

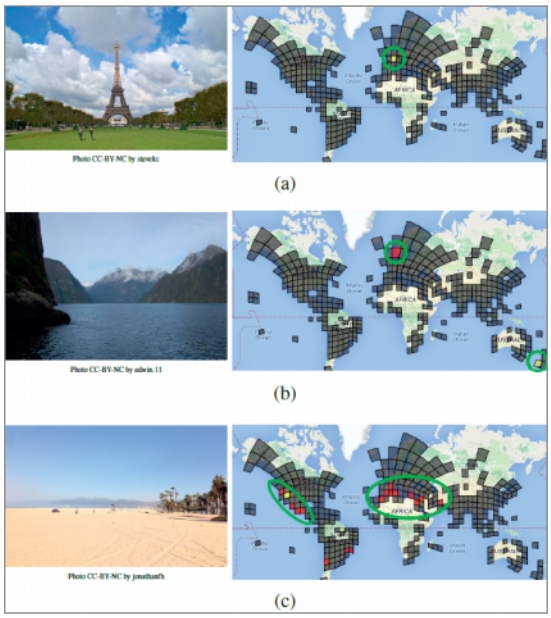

Bild: Google

PlaNet extrahiert typische Merkmale einer Szene und berechnet für jede Region die Wahrscheinlichkeit, dass das Bild dort aufgenommen wurde. Den Eiffelturm verortet es mit Bestimmtheit in Paris, das Fjord-Foto passt zu Norwegen und Neuseeland. Den kalifornischen Strand hat PlaNet ebenfalls korrekt verortet, scheint Ähnliches aber auch vom Mittelmeer und aus Mexiko zu kennen.

Diese Abgrenzung dürfte auch thematisch stimmig sein: Innerhalb einer Stadt muss der Algorithmus die Variabilität einzelner Viertel und Straßenzüge erfassen, während auf dem Land Wiesen und Felder dominieren und in der Wüste eben quadratkilometerweise Sand mit Dünen. Gegenden wie die hohe See oder die Pole, die extrem selten fotografiert werden, wurden ausgespart.

Damit besteht der Modell-PlaNet aus 26 263 Regionen beziehungsweise Kategorien, auf die sich insgesamt 126 Millionen Fotos verteilen, also im Schnitt knapp 4800 Bilder pro Kategorie. Diese Bildersammlung für Training und Test ist den Forschern zufolge übrigens noch recht verrauscht, weil Fotos von Innenräumen, Tieren oder Speisen nicht aussortiert wurden.

Mehr als zwei Drittel der Testfotos (71,3 Prozent) konnte das trainierte System schließlich dem richtigen Kontinent zuordnen, also innerhalb eines Radius von 2500 km verorten, gut die Hälfte (53,6 Prozent) dem richtigen Land (750 km Umkreis), ein Viertel der passenden Stadt und 8,5 Prozent sogar auf Straßenniveau. Und in einem GeoGuessr-Test-Match (siehe c't-Link) gegen welterfahrene Reisende konnte das System 28 von 50 Fotos genauer lokalisieren.

Zukunft des PlaNeten

Mit einem erweiterten System, das ganze Fotoalben im Zusammenhang betrachtet, konnten die Forscher die Erkennungsquote noch um 50 Prozent steigern. Dazu selektierten die Forscher knapp 30 Millionen Google+-Fotoalben und kombinierten das auf Einzelbilder spezialisierte CNN mit einem sogenannten Long-Short-Term-Memory-Netz (LSTM). LSTMs werden beispielsweise in der Sprachverarbeitung eingesetzt, um den Sinn längerer Texte oder gesprochener Wörter deuten zu können – der sich ja häufig erst aus dem Kontext ergibt. Dieses Prinzip funktioniert auch bei der Foto-Lokalisierung: Während das erste Foto nur eine Alm mit Wiese und Kühen zeigt, taucht auf dem zweiten vielleicht schon die Schweizer Flagge oder im Hintergrund das Matterhorn auf – und löst die Mehrdeutigkeit auf.

Die Arbeit von Weyands Gruppe ist ein weiteres Beispiel für die wachsende Bedeutung der Long-Short-Term-Memory-Netze. Kombinationen von CNN und LSTM dienen beispielsweise auch dazu, Bildbeschreibungen zu verfassen. (atr@ct.de)

Literatur

[1] Tobias Weyand, Ilya Kostrikov, James Philbin: PlaNet – Photo Geolocation with Convolutional Neural Networks, http://arxiv.org/abs/1602.05314

[2] Andrea Trinkwalder: Netzgespinste, c't 06/16, S. 130

c't PlaNet-Paper, GeoGuessr: ct.de/yv7n