ICH SEHE WAS, DAS DU NICHT SIEHST

VISUAL QUESTION ANSWERING HEUTE & IN ZUKUNFT

Thomas Eiter, Nelson Higuera, Johannes Oetsch

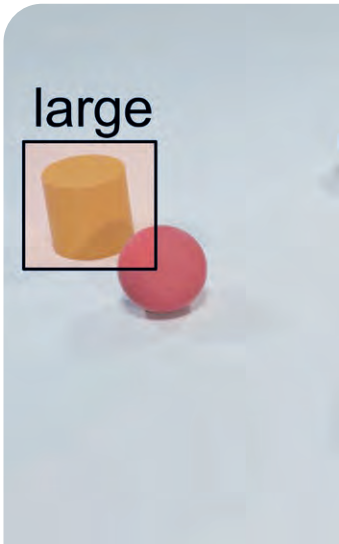
Fragen zu Bildern zu beantworten, stellt Maschinen vor mehrere Herausforderungen. In Verbindung mit anderen KI-Ansätzen können hier Large Language Models (LLMs) zu großen Fortschritten beitragen.

In Science-Fiction-Filmen können KI-Systeme, wie z.B. in Stanley Kubricks Klassiker „2001: A Space Odyssey“ der omnipräsente Computer HAL, visuellen Input wahrnehmen und sich mit Menschen darüber austauschen. Was spielerisch aussieht, benötigt verschiedene kognitive Fähigkeiten, die für sich gesehen komplexe Problemstellungen bergen und deren Verbindung schwierig ist. Im Gebiet des Visual Question Answerings (VQA) beschäftigt man sich damit, konkrete Fragen in natürlicher Sprache zu einem Bild oder Video maschinell zu beantworten, z.B. also mit der Frage, welche Größe der Zylinder in Abb. 1 links vom braunen Metallobjekt hat, das links von der großen Kugel ist. Für die Antwort muss man intuitiv Objekte erkennen, die Frage verstehen und die Antwort unter Bedacht auf semantische Begriffe wie „links von“, „braun“ und „groß“ bilden.

Es gibt unzählige VQA-Datensets und -Benchmarks unterschiedlicher Charakteristik, zudem wurden viele Architekturen für VQA-Systeme entwickelt. Alle benutzen neuronale Netzwerke und Deep Learning, um oben genannte Subprobleme lösen zu können. Bei End-to-End-Architekturen erfolgt dies implizit in einem (komplexen) neuronalen Netzwerk, während modulare Architekturen spezielle Komponenten aufweisen. Zu Letzteren gehören neuro-symbolische Ansätze, die Komponenten mit symbolischer Wissensdarstellung und -verarbeitung verwenden (Abb. 2). Dabei wird ein explizites semantisches Modell einer Szene wie in Abb. 1 erstellt, das die Basis für die Ermittlung der Antwort auf die Frage bildet. Neben der Modularität, die einen Austausch von Komponenten erlaubt, ist vor allem die Verfügbarkeit eines symbolischen Modells ein großer Vorteil.

Abb. 1: Visual Question Answering: Beispielfrage (Q) in einem Szenario mit Antwort (A) und einer Erklärung, welche Änderung eine andere Antwort (F) bewirken würde (rechts)

(Q): Welche Größe hat der Zylinder links vom braunen Metallobjekt, das links von der großen Kugel ist? (A) klein (F) Wann wäre die Antwort „groß“?



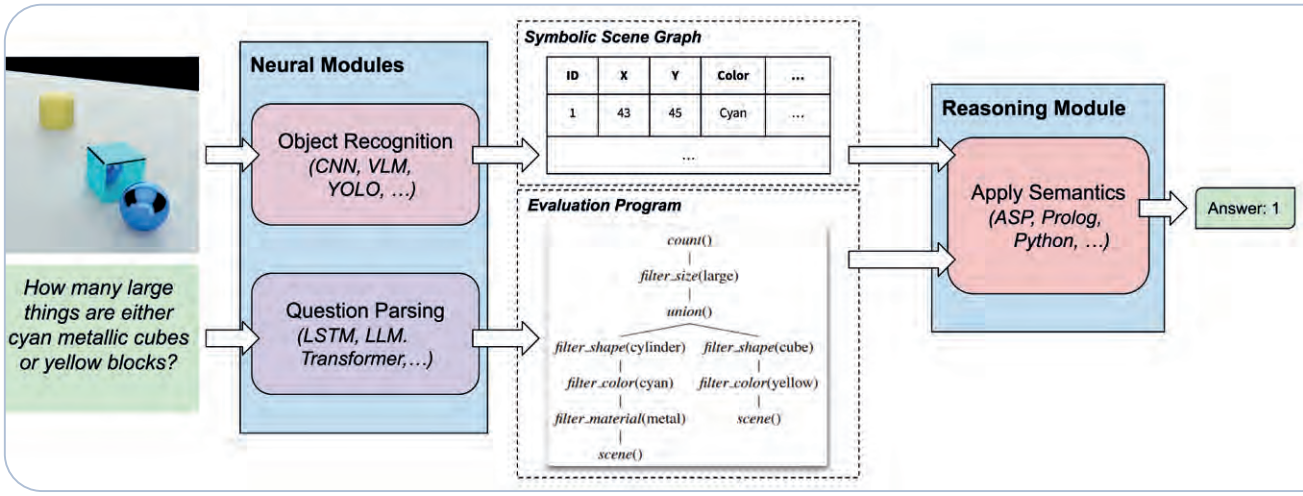
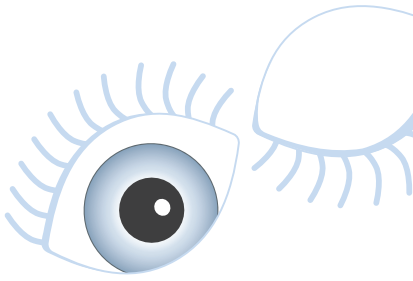


Abb. 2: Neuro-symbolische Architektur für VQA

Ein symbolisches Modell begünstigt die Einsicht in das System und kann für die Erstellung von Erklärungen genutzt werden, womit sich Transparenz und Systemvertrauen erhöhen lassen. Insbesondere haben Modelle in deklarativen, logik-basierten Sprachen den Vorteil, dass sie für Schlussfolgerungen in verschiedene Richtungen (von Eigenschaften zu Konsequenzen oder von Beobachtungen zu möglichen Ursachen) flexibel nutzbar sind. So lassen sich z.B. Erklärungen wie „Unter welchen Änderungen würde die Antwort in Abb. 1 ‚groß' lauten?" unter Verwendung von Techniken wie Abduktion und kontra-faktuellem (hypothetischem) Schließen elegant lösen. In einer Kooperation mit dem Bosch Center for Artificial Intelligence in Renningen, Deutschland, verfolgen Forscher:innen am Institute of Logic and Computation der TU Wien den neuro-symbolischen Ansatz, der für Anwendungen in Gebieten wie Produktion, autonomes Fahren, Gesundheitswesen u.a.m. von großem Interesse ist.

Fotos: © TU Wien

Der rasante Aufstieg von LLMs für Sprache wie etwa GPT oder PaLM eröffnen auch für das VQA neue Möglichkeiten. Auf LLMs basierte Programme wie ChatGPT oder Bard ermöglichen eine barrierefreie Kommunikation zwischen dem System und dem Benutzer, die sich zu einem komplexen Dialog entspinnen kann. Ein solches Szenario bringt neue Herausforderungen: Fragen sind nicht wie üblich an ein enges Vokabular gebunden, und Antworten erfolgen im Kontext vorheriger Fragen und Antworten sowie Rückmeldungen. Es sind Lösungen gefragt, die auf allgemeinere Fragebereiche übertragbar sind und mühelos (im maschinellen Lernen als Zero-Shot Solving bekannt) angepasst werden können. Weiters ist eine Einschätzung des Gegenübers sowie von dessen Kenntnissen und Erwartungshaltungen hilfreich. Dafür müssen Benutzermodelle entwickelt werden, aus denen fundierte Schlüsse gezogen werden können. Commonsense Reasoning, d.h. Schließen mit Allgemeinwissen, z.B. ein verdecktes Objekt ist nicht sichtbar, existiert aber – eines der schwelenden Probleme der KI –, und auch ethische Aspekte spielen hierbei eine wichtige Rolle. Neuro-symbolische Ansätze sind hierzu naturgemäß prädestiniert.



VQA hat in den letzten Jahren große Fortschritte erlebt. Bis potenziell vielschichtige Fragen wie „Ich sehe was, das du nicht siehst" befriedigend beantwortet werden können, dürfte aber noch einiges an Forschung und Entwicklung nötig sein.