Mit Künstlicher Intelligenz und dem ChatGPT-Algorithmus auf der Jagd nach Cyberkriminellen

Die Verwendung von mehrschichtigen, neuronalen Netzen hat die Leistung des maschinellen Lernens in vielen Bereichen, etwa Bilderkennung, maschinelle Übersetzung und Malware-Klassifizierung, erheblich verbessert. Mit zunehmender Skalierung werden neuronale Netze immer besser und sind somit durchaus in der Lage, echte „Game Changer“ zu sein. Das macht momentan der Chatbot ChatGPT mehr als deutlich. Basis für das Programm ist das ebenfalls von OpenAI stammende Sprachverabeitungsmodell GPT-3.5. Im Supercomputing-Maßstab und im Zusammenspiel mit maschinellem Lernen, nutzt Sophos AI diese fortschrittliche Technologie, um noch bessere Sicherheitsanwendungen zu realisieren. Denn insbesondere im Bereich der Cybersicherheit sticht die enorme Leistungsfähigkeit im Gegensatz zu bisherigen, kleineren Modellen hervor. Das machen verschiedene, erfolgreiche Testreihen klar, die die Sophos-Experten im letzten Jahr noch mit GPT-3 gemacht haben

GPT-3 bietet enorme Potenziale für die IT-Security
GPT-3 ist ein vortrainiertes, umfangreiches Sprachmodell, dessen Flexibilität und Genauigkeit durchaus bahnbrechend sind. Und genau an dieser Stelle ist die menschliche Kreativität von Sophos AI gefragt, nämlich wo und wie sich diese Technologie sinnvoll im Kampf gegen Cyberkriminalität einsetzen lässt. Denn wenn Eingabe- und Ausgabedaten in Text umgewandelt werden können, sind die Anwendungsmöglichkeiten von GPT-3 auch in diesem Bereich endlos. Zum Beispiel könnte man GPT-3 bitten, aus einer Funktionsbeschreibung funktionierenden Python-Code zu schreiben oder eine Klassifizierungsanwendung mit nur wenigen Beispielen erstellen.

Die Experten von Sophos AI sehen in GPT-3 enorme Potentiale. Beispielsweise ist es relativ einfach, einen unmarkierten Datensatz zu finden; allerdings ist es meist sehr zeitaufwändig und schwierig, einen markierten Datensatz für das Training eines herkömmlichen maschinellen Lernmodells zu erstellen. Herkömmliche maschinelle Lernmodelle, die mit wenigen Beispielen trainiert werden, weisen häufig Probleme mit der Überanpassung auf. Sprich, sie lassen sich nicht gut auf zuvor nicht existente Beispiele verallgemeinern. Mit dem GPT-3 „Few-Shot Learning“ hingegen benötigt Sophos AI nur wenige kommentierte Trainingsbeispiele und übertrifft damit herkömmliche Modelle. Da GPT-3 selbstüberwacht und in großem Umfang trainiert wurde, hat sich gezeigt, dass es bei mehreren Klassifizierungsproblemen mit nur wenigen Beispielen gut abschneidet.

Zwei Beispiele für die konkrete Anwendung:

Spam-Erkennung

Es ist eine Herausforderung, ein leistungsstarkes Spam-Klassifizierungsmodell mit nur vier unkritischen und vier Spam-Beispielen zu trainieren. Herkömmliche Klassifizierungsmodelle benötigen oft einen großen Trainingsdatensatz, um genügend Signale zu lernen. Da GPT-3 jedoch ein Sprachmodell ist, das mit einem großen Textdatensatz trainiert wurde, kann es die Intension einer Klassifizierungsaufgabe erkennen und die Aufgabe mit wenigen Beispielen lösen.

Beim Lernen mit wenigen Beispielen ist das Prompt-Engineering, bei dem das Format der Eingabedaten für Textvervollständigungsaufgaben entworfen wird, ein wichtiger Schritt. Abbildung 1 zeigt den Prompt für eine Spam-Klassifizierungsaufgabe.

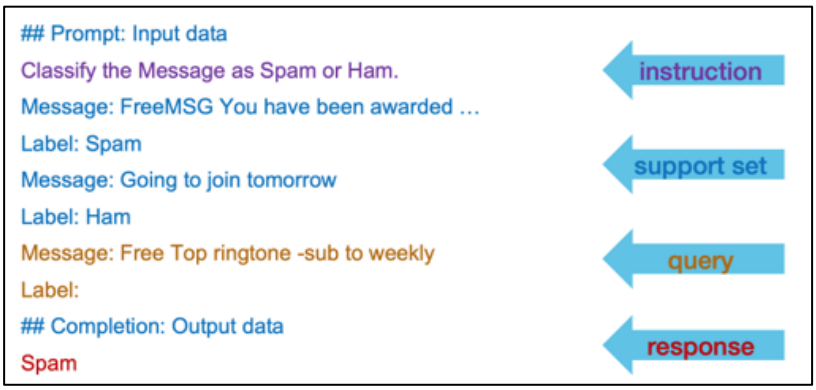

Abb. 1

Der Prompt enthält eine Anweisung und einige Beispiele mit ihren Beschriftungen als Support-Set, und im letzten Abschnitt ist ein Abfragebeispiel angefügt. Anschließend wird GPT-3 aufgefordert, aus der Eingabe eine Antwort als Label-Predication zu generieren.
Vergleicht man die Klassifizierungsergebnisse zwischen traditionellen ML-Modellen und dem „Few-Shot-Learning“ mit GPT-3, ist schnell zu erkennen, dass es die traditionellen ML-Modelle wie logistische Regression und „Random Forest“ deutlich übertrifft. Dies liegt daran, dass das „Few-Shot-Learning“ die Kontextinformationen der gegebenen Beispiele nutzt und das Label des ähnlichsten Beispiels als Ausgabe auswählt. Dadurch erfordert GPT-3 kein erneutes Training, sondern ermöglicht es, ein leistungsstarkes Klassifizierungsmodell mit einfachem Prompt-Engineering zu erstellen.

Lesbaren Erklärungen für schwer zu entzifferndem Code
Das Reverse Engineering von Befehlszeilen ist selbst für Sicherheitsexperten eine schwierige und zeitraubende Aufgabe. Noch schwieriger ist es, „Living-off-the-Land"-Befehle zu verstehen, denn diese sind lang und enthalten schwer zu analysierende Zeichenketten. Angreifer nutzen hierfür Standard-Apps und Standard-Prozesse auf dem Computer ihrer Opfer, um beispielsweise Phishing-Aktivitäten zu tarnen. GPT-3kann eine Befehlszeile in eine verständliche Beschreibung übersetzen – zum Beispiel aus einer gegebenen Beschreibung des Codes einen funktionierenden Python- oder Java-Code schreiben. Es ist auch möglich, GPT-3 zu bitten, mehrere Beschreibungen aus einer Befehlszeile zu generieren, und die ausgegebenen Beschreibungen werden mit Token-Wahrscheinlichkeiten auf Wortebene versehen, um den besten Kandidaten auszuwählen. Der Ansatz von Sophos AI zur Auswahl der besten Beschreibung aus mehreren Varianten ist die Verwendung einer Rückübersetzungsmethode, die diejenige Beschreibung auswählt, die die ähnlichste Befehlszeile zur Eingabebefehlszeile erzeugen kann.

„GPT-3 ist ein Meilenstein für die Cybersicherheit, da es Spam erkennen und komplexe Befehlszeilen mit wenigen Beispielen analysieren kann“, so die Experten des Sophos AI Teams. „Die Flexibilität von GPT-3 eignet sich sehr gut für den Kampf gegen die sich ständig weiterentwickelnden Cyber-Bedrohungen. Wir gehen davon aus, dass in Kürze auch die noch schwierigeren Cybersicherheitsprobleme mit entsprechend größeren neuronalen Netzwerkmodellen adressiert werden können.“

Social Media von Sophos für die Presse
Wir haben speziell für Sie als Journalist*in unsere Social-Media-Kanäle angepasst und aufgebaut. Hier tauschen wir uns gerne mit Ihnen aus. Wir bieten Ihnen Statements, Beiträge und Meinungen zu aktuellen Themen und natürlich den direkten Kontakt zu den Sophos Security-Spezialisten.
Folgen Sie uns auf und

LinkedIn: https://www.linkedin.com/groups/9054356/
Twitter: @sophos_info

Pressekontakt:

Sophos
Jörg Schindler, PR-Manager Central & Eastern Europe
joerg.schindler@sophos.com, +49-721-25516-263

TC Communications
Arno Lücht, +49-8081-954619
Thilo Christ, +49-8081-954617
Ulrike Masztalerz, +49-30-55248198
Ariane Wendt +49-172-4536839
sophos@tc-communications.de