オープンカレッジ

安心安全なAIの実現に向けて

AIが持つ脆弱性とは

名城大学情報工学部
准教授
野崎　佑典

のざき・ゆうすけ　専門は情報セキュリティー。名城大学大学院理工学研究科博士後期課程修了。博士（工学）。1991年生まれ。

現在、生成AI（人工知能）をはじめとしたさまざまなAI技術の利活用が進んでいる。AI技術によって、テキストや画像、音声の認識・推論に加えて、これらの情報を生成することも可能になってきた。一方で、これらのAI技術の利活用において、AI自身に生じる情報セキュリティーの脅威が報告されている。具体的には、「敵対的サンプル」や「モデル抽出攻撃」、「モデル反転攻撃」などの脅威が知られている。

敵対的サンプルは、AIの推論時にノイズを混入させることで意図的に推論結果を誤らせる攻撃である。モデル抽出攻撃は、大量のデータや豊富な計算資源を用いて作成した、貴重なAIモデルを盗み出す攻撃であり、AIサービスを提供する企業にとって重大な脅威となる。モデル反転攻撃は、AIの学習に使用された、通常は秘匿されるべき情報を復元する攻撃であり、プライバシーに対する脅威である。

ここでは敵対的サンプルについて取り上げる。敵対的サンプルでは、AIに与える画像などの入力に対して、「摂動（せつどう）」と呼ばれるある種のノイズ成分を加えることで、特殊な画像を生成する。加える摂動は微小であるため、人の目には生成された画像と元の画像との違いを見分けることはできない。一方で、AIがこの画像を推論する際には、元の画像とは全く違うものに認識されてしまう。これは、AIの学習や推論過程に着目し、AIが誤推論するように意図的に摂動を生成しているためである。

具体的には、AI技術で用いるニューラルネットワークでは、正解クラスに近付くように損失関数の値を最小化するように学習を行い、パラメータを決定するのに対して、敵対的サンプルでは損失関数を最大化するような入力画像を探索することで誤認識を誘発させている。このとき、AIに対してただ単に誤って推論させてしまうだけでなく、意図的に指定したクラスへ推論結果を誘導させることも可能である。

このような画像が自動運転などで用いられる画像推論AIを対象に使用された場合、人物や道路標識などの推論を誤る可能性があり、人命に関わる事故を誘発する恐れがある。そのため、これらの攻撃に対する対策技術を確立する必要がある。AIの頑健性を高める手法として、AIの学習時に通常の画像に加えて、敵対的サンプルも用いる敵対的学習が報告されている。他にも、敵対的サンプルは入力に微小な摂動を混入させていることから、この摂動を取り除くフィルタリングを導入する手法も知られている。

いずれの手法に対しても、対策を打ち破るための攻撃手法が報告されており、イタチごっこの状態となっている。私の研究室では、AIの安全性を高めるためにさまざまな防御手法の研究を進めている。情報セキュリティーの世界では、攻撃と防御がイタチごっことなってしまうが、今後も安心安全な社会の実現に寄与するために研究に取り組んでいきたい。