気になるこの用語

第31回

消費生活相談の周辺用語を取り上げ、やさしく解説します。

松原 仁 Matsubara Hitoshi 人工知能研究者
東京大学大学院情報理工学系研究科AIセンター教授。公立はこだて未来大学特任教授。元人工知能学会会長。著書に『AIに心は宿るのか』(集英社インターナショナル、2018年)など

人工知能 ディープラーニングの弱点

ディープラーニングは対象物に対して人間とは違う特徴に注目する可能性があります。そのため思いがけない問題が生じます。画像認識を例にどのような問題が起きるのか説明します。

ディープラーニングは人間と違う見方をする ― 画像認識の落とし穴

前回まで述べてきたようにディープラーニングは非常に強力な道具です。従来の人工知能では不可能だった、任意の対象物の特徴を自ら見つけて識別や判定ができます。これはすごいことなのですが、このことによって思いがけない問題が生じます。ディープラーニングによる画像認識を例にして説明しましょう。

ディープラーニングによる画像認識の能力が高いことは既に説明しましたが、人間が画像を認識するしくみとディープラーニングが画像を認識するしくみは同じではありません。ディープラーニングは人間とは異なる場所に注目する場合があります。

ある研究者が「動物が含まれている写真」と「動物が含まれていない写真」をそれぞれ大量に学習させた結果、ディープラーニングはその2つを高い精度で分類できるようになりました。しかしその研究者が詳しく調べてみたところ、ディープラーニングは動物が含まれているか含まれていないかで分類をしているのではなかったのです。何と、背景がぼやけている写真を「動物が含まれている写真」と分類していました。「動物が含まれている写真」は動物にピントが合っているので、背景はぼやけています。「動物が含まれていない写真」として見せていたものは、背景の景色がぼやけずに写っている写真でした。景色だけが写っているので景色にピントが合っています。そのためにディープラーニングは背景がぼやけている写真を「動物が含まれている写真」、背景がはっきりしている写真を「動物が含まれていない写真」と見なすように学習してしまったのです。人間とは違う所をディープラーニングは見ていたことになります。ディープラーニングとしては、結果的に人間と同じ分類ができればよいのであって、その分類を行う方法が人間と同じとは限らないということです。これではディープラーニングへの教え方がまずかった、ということになります。

動物が含まれていない写真として「建築物が含まれている写真」や「自動車が含まれている写真」などを見せて学習させれば、ディープラーニングは「背景がぼやけているのが、動物が含まれている写真」とは分類しません(どの写真も背景はぼやけているため)。しかし、また別の、動物とは異なる違いに注目して学習してしまう可能性はあります。

また、ディープラーニングによる画像認識を「だます」こともできてしまいます。一例として、人間が見たら明らかにスクールバスなのですが、ディープラーニングはダチョウと判断する画像を人工的に作ることができます。画像の一部の画素に細工を施すと、人間にはその変化がまったく分からない(元のスクールバスに見える)のに、ディープラーニングにはダチョウと

して見えてしまうということが起きます。

人間に錯視という現象があります。同じ長さなのに、長さが異なるように見えるというような例です。ディープラーニングにも特有の「錯視」があることになります。これはディープラーニングが“自ら特徴を習得できる”という性質によって起きています。人間がスクールバスをスクールバス、ダチョウをダチョウと認識するときに用いている特徴と、ディープラーニングがそれぞれを認識するときに用いている特徴は、ある程度は同じですが、一部は異なるのです。人間にはスクールバスに見える画像に、ディープラーニングにはダチョウを示す特徴が存在しているわけです。これはかなり困った問題です。それは、悪意があればいろいろな不正行為ができてしまうからです。例えば、人間には明らかに別人と分かるのに、ディープラーニングは同一人物と判断してしまう、ということが可能になります。こういうことが起きないようにするための対策が研究されています。

学習データに含まれる偏りの問題

ディープラーニングによる画像認識は、人間が持っている偏見や差別をあらわにしてしまうという問題もあります。意識的な偏見だけでなく、無意識的な偏見も顕在化します。世界中で広く市販されているディープラーニングの画像認識システムは、白人男性の顔をより正確に認識できる傾向を持っています。有色人種や女性の顔は間違いやすいのです。そうなってしまう原因はディープラーニングに学習させた膨大なデータに白人男性の写真が多いからです。ある調査によれば、顔の画像認識に使われている標準データにおいて、白人が占める割合は84%、そのうち男性が占める割合は78%だそうです。インターネットに載っている顔写真は有名人やお金持ちが多く、その大半が白人男性だからです。これは画像認識に限った話ではなく、ディープラーニングに共通する問題です。多様なデータを集める必要はあるのですが、無意識的にデータが偏ってしまう危険があります。

第三者によるラベリングの問題

ディープラーニングによる画像認識は人間とは異なる特徴に注目する場合があると説明しましたが、それがもたらすショッキングな研究結果が最近発表されました。ディープラーニングによる顔の認識で、その人がどの政党を支持するかかなりの精度で分かるという研究がアメリカで発表されたのです。アメリカなので民主党支持者か共和党支持者に大別されますが、顔写真だけで、どちらかがかなりの精度で分かるというのです。人種、性別、年齢などが違えば両党の支持率が違うことは当然あるでしょうから、それらの情報から分かるのは意外ではありません。この研究結果が意外なのは、人種、性別、年齢を一致させてもディープラーニングは顔写真だけから支持政党を70%以上の確率で当てることができたということです。偶然当たる確率はほぼ50%なので、70%以上というのはかなりの精度です。この発表をした研究者は、以前、ディープラーニングによる顔の認識でその人の性的指向がかなりの精度で分かるという研究も発表しました。人間にはその人の顔を見ただけでは、支持政党や性的指向は分かりませんが、ディープラーニングは(それなりに)分かってしまうということです。骨相(こっそう)学は科学的に否定されたはずなのに、ディープラーニングがどうして分かるのかは今後の研究を待たないといけません。骨格ではなく顔の向きや表情などに特徴が表れているのではないかという見解もあります。背景として、アメリカでは支持政党や性的指向をSNSなどで公表している人が多く、その情報と公開された顔写真とひも付けて、ディープラーニングで学習させることが可能なのです。日本ではその種の情報はあまり公開されていませんので同様の研究は難しいと思われます。

次回は、人工知能と創造性について考えてみましょう。