One-Pixel Attack Fools Neural Networks

Originalartikel

Backup

<html> <p>Deep Neural Networks can be pretty good at identifying images &#8212; almost as good as they are at attracting Silicon Valley venture capital. But they can also be fairly brittle, and a slew of research projects over the last few years have been working on making the networks&#8217; image classification less likely to be deliberately fooled.</p> <p>One particular line of attack involves adding particularly-crafted noise to an image that flips some bits in the deep dark heart of the network, and makes it see something else where no human would notice the difference. We got tipped with a <a href=„https://www.youtube.com/watch?v=SA4YEAWVpbk“ target=„_blank“>YouTube video of a one-pixel attack</a>, embedded below, where changing a single pixel in the image would fool the network. Take that robot overlords!</p> <figure id=„attachment_303518“ class=„wp-caption alignright c2“><a href=„https://hackadaycom.files.wordpress.com/2018/04/dog_image.png“ target=„_blank“><img data-attachment-id=„303518“ data-permalink=„https://hackaday.com/2018/04/15/one-pixel-attack-fools-neural-networks/dog_image/“ data-orig-file=„https://hackadaycom.files.wordpress.com/2018/04/dog_image.png“ data-orig-size=„736,1146“ data-comments-opened=„1“ data-image-meta=„{&quot;aperture&quot;:&quot;0&quot;,&quot;credit&quot;:&quot;&quot;,&quot;camera&quot;:&quot;&quot;,&quot;caption&quot;:&quot;&quot;,&quot;created_timestamp&quot;:&quot;0&quot;,&quot;copyright&quot;:&quot;&quot;,&quot;focal_length&quot;:&quot;0&quot;,&quot;iso&quot;:&quot;0&quot;,&quot;shutter_speed&quot;:&quot;0&quot;,&quot;title&quot;:&quot;&quot;,&quot;orientation&quot;:&quot;0&quot;}“ data-image-title=„dog_image“ data-image-description=„“ data-medium-file=„https://hackadaycom.files.wordpress.com/2018/04/dog_image.png?w=257&amp;h=400“ data-large-file=„https://hackadaycom.files.wordpress.com/2018/04/dog_image.png?w=401“ class=„size-medium wp-image-303518“ src=„https://hackadaycom.files.wordpress.com/2018/04/dog_image.png?w=257&amp;h=400“ alt=„“ width=„257“ height=„400“ srcset=„https://hackadaycom.files.wordpress.com/2018/04/dog_image.png?w=257&amp;h=400 257w, https://hackadaycom.files.wordpress.com/2018/04/dog_image.png?w=514&amp;h=800 514w, https://hackadaycom.files.wordpress.com/2018/04/dog_image.png?w=161&amp;h=250 161w“ sizes=„(max-width: 257px) 100vw, 257px“/></a> <figcaption class=„wp-caption-text“>We can&#8217;t tell what these are either..</figcaption></figure><p>Or not so fast. Reading the fine-print in the <a href=„https://arxiv.org/abs/1710.08864“ target=„_blank“>cited paper</a> paints a significantly less gloomy picture for Deep Neural Nets. First, the images in question were 32 pixels by 32 pixels to begin with, so each pixel matters, especially after it&#8217;s run through a convolution step with a few-pixel window. The networks they attacked weren&#8217;t the sharpest tools in the shed either, with somewhere around a 68% classification success rate. What this means is that the network was unsure to begin with for many of the test images &#8212; making it flip from its marginally best (correct) first choice to a second choice shouldn&#8217;t be all that hard.</p> <p>This isn&#8217;t to say that this line of research, adversarial training of the networks, is bogus. The idea that making neural nets robust to small changes is important. You don&#8217;t want <a href=„https://hackaday.com/2017/11/03/googles-inception-thinks-this-turtle-is-a-gun/“>turtles to be misclassified as guns</a>, for instance, or Hackaday&#8217;s own Steven Dufresne <a href=„https://hackaday.com/2017/06/14/diy-raspberry-neural-network-sees-all-recognizes-some/“>misclassified as a tobacconist</a>. And you certainly don&#8217;t want speech recognition

software to be <a href=„https://hackaday.com/2018/01/15/fooling-speech-recognition-with-hidden-voice-commands/“>fooled by carefully crafted background noise</a>. But if a claim of &#8220;astonishing results&#8221; on YouTube seems too good to be true, well, maybe it is.</p> <p>Thanks [kamathin] for the tip!</p> <p><iframe class=„youtube-player c3“ type=„text/html“ width=„800“ height=„480“ src=„https://www.youtube.com/embed/SA4YEAWVpbk?version=3&amp;rel=1&amp;fs=1&amp;autohide=2&amp;showsearch=0&amp;showinfo=1&amp;iv_load_policy=1&amp;wmode=transparent“ allowfullscreen=„true“>[embedded content]</iframe></p> </html>

From:
https://schnipsl.qgelm.de/ - **Qgelm**

Permanent link:
https://schnipsl.qgelm.de/doku.php?id=wallabag:one-pixel-attack-fools-neural-networks

Last update: **2021/12/06 15:24**