arunpatala / captcha.irctc

Watch 13 Star 137 Fork 24

Code Issues 11 Pull requests 0 Pulse Graphs

Reading irctc captchas with 98% accuracy using deep learning

29 commits 1 branch 0 releases 1 contributor

Branch: master New pull request Find file Clone or download

arunpatala Update README.md Latest commit 8ed60cd on Mar 5

data	Initial commit	6 months ago
models	resnet and data aug	6 months ago
MultiCrossEntropyCriterion.lua	Multi	6 months ago
README.md	Update README.md	6 months ago
data.lua	Initial commit	6 months ago
main.lua	cleaning up	6 months ago
mainaug.lua	resnet and data aug	6 months ago
mainr.lua	resnet and data aug	6 months ago
models.lua	resnet and data aug	6 months ago
train.lua	resnet and data aug	6 months ago
transforms.lua	resnet and data aug	6 months ago

README.md

Reading irctc captchas with 98% accuracy using deep learning

Captcha is a system used on popular websites to stop bots or automatic software to access the site. IRCTC is a popular travel website in india where people book travel tickets on trains. Because of high demand of tickets, booking during peak hours (tatkal) has a captcha image containing letters that humans have to enter to book the ticket. This supposedly stops ticket booking through automated software.

Update

Using the residual networks (34 layers deep) which won the Imagenet challenge and some data augmentation improves the captcha accuracy to 98%. I have added the relevant code. Run mainr.lua or mainaug.lua.

Demo

Demo upload your own file

Here we show that the current captcha system used by IRCTC is vulnerable. Especially we see how to use deep learning to read captchas with 95% accuracy (better than me). Machine learning can be seen as trying to find a function given examples

of input and output of that function. So if y=f(x), here x can be the captcha image and y can be text to be read from the image. We are trying to fit a function f using machine learning to read text from the image. The great thing about machine learning is that if we provide the pairs of image and text, the ml is generic enough and tries to figure out the underlying function. Here we use deep neural networks as out machine learning algorithm. So the two main steps are to generate the pairs of images and text which is one of the main tasks in machine learning applications

Dataset

To use deep learning we need sufficient training set. For this use this url to download sample captcha images and you need to provide the appropriate labels in the format used in example data folder in this repository. I have used around 10000 samples to acheive 95% accuracy (test set 1000 samples). The example data uploaded has only 100 samples which is insufficient for training this model.

Neural network

We use torch to train the neural network and a VGG based deep neural network architecture. We train on a GPU Nvidia 780 Titan. Otherwise it takes a lot of time to train. You can check similar set up for CIFAR dataset at this blog post. The main difference is the criterion, as the output of a image is a sequence of characters which begs us to use RNN. But that is also not necessary here as we use our custom MultiCrossEntropyCriterion and is sufficient to get good enough results. I have committed the code to github repository here. We assume the dataset to be in a folder (named data). The shown image size is 50×170. The network can be changed accordingly to fit any other captcha.

What it means

Captcha as a system is vulnerable and should be stopped as noted here. There are other means to stop automated software. This is a demonstration that the current captcha system used by IRCTC is vulnerable and doesnt do its intended purpose of stopping automated booking. Thats one reason why a lot of automated software is puportedly available. I have tried contacting IRCTC to let them know about this. Atleast this will make them stop using captchas which doesnt stop bots and even inconvenice legit users.

Requirements

Install torch software along with CUDA and cutorch, csvigo remove cuda calls if not using CUDA in main.lua and train.lua set batchsize, data folder etc in main.lua th main.lua

Contact

If you are in need of help in any interesting projects in machine learning or deep learning, please contact arunpatala@gmail.com