Breaking CAPTCHAs on the Dark Web

Using neural networks to enable scraping

RP #62, Kevin Csuka & Dirk Gaastra

Supervisor: Yonne de Bruijn, Fox-IT

6 February, 2018

University of Amsterdam

Introduction

Scraping the Dark Web

Useful for threat intelligence companies

Scraping the Dark Web

Useful for threat intelligence companies

... sometimes hard to get to.

Scraping the Dark Web

Useful for threat intelligence companies

... sometimes hard to get to.

Mainly the blockades, such as CAPTCHAs, is an issue for the scrapers.

CAPTCHA

Figure 1: CAPTCHA example

- Completely Automated Public Turing test to tell Computer and Humans Apart

CAPTCHA

Figure 1: CAPTCHA example

- Completely Automated Public Turing test to tell Computer and Humans Apart
- Test to determine whether the user is human or not

Main question

How would a scraper be able to circumvent CAPTCHAs that prevent it from properly scraping dark web websites?

Main question

How would a scraper be able to circumvent CAPTCHAs that prevent it from properly scraping dark web websites?

Sub-questions:

1. Impact of solving CAPTCHAs

Main question

How would a scraper be able to circumvent CAPTCHAs that prevent it from properly scraping dark web websites?

Sub-questions:

1. Impact of solving CAPTCHAs
2. Solve CAPTCHAs by using Optical Character Recognition (OCR)?

Main question

How would a scraper be able to circumvent CAPTCHAs that prevent it from properly scraping dark web websites?

Sub-questions:

1. Impact of solving CAPTCHAs
2. Solve CAPTCHAs by using Optical Character Recognition (OCR)?
3. Solving CAPTCHAs by using Machine Learning (ML)

Related Work

Related Work

1. Lawrence et al. created their own dark web scraping tool, D-miner; CAPTCHAs were solved by human labor [1]

Related Work

1. Lawrence et al. created their own dark web scraping tool, D-miner; CAPTCHAs were solved by human labor [1]
2. Ryan Mitchell demonstrated how to solve CAPTCHAs using Optical Character Recognition with Tesseract [2]

Related Work

1. Lawrence et al. created their own dark web scraping tool, D-miner; CAPTCHAs were solved by human labor [1]
2. Ryan Mitchell demonstrated how to solve CAPTCHAs using Optical Character Recognition with Tesseract [2]
3. Torch has previously been used to train a neural network to solve CAPTCHAs by Arun Patala [3]

Methods

Methods

Two methods to solve the questions:

1. Categorizing dark web websites
2. Breaking CAPTCHAs

1. Categorizing websites

1. Categorizing websites

Analysis of 633 dark web websites

1. Categorizing websites

Analysis of 633 dark web websites

- Which ones are up?

1. Categorizing websites

Analysis of 633 dark web websites

- Which ones are up?
- Are there any duplicates?

1. Categorizing websites

Analysis of 633 dark web websites

- Which ones are up?
- Are there any duplicates?
- Which ones block scraping?

1. Categorizing websites

Analysis of 633 dark web websites

- Which ones are up?
- Are there any duplicates?
- Which ones block scraping?
- What kind of blockade are they using?

2. Breaking CAPTCHAs

2. Breaking CAPTCHAs

There are 3 common approaches to defeat CAPTCHAs:

2. Breaking CAPTCHAs

There are 3 common approaches to defeat CAPTCHAs:

1. Using a service which solves CAPTCHAs through human labor

2. Breaking CAPTCHAs

There are 3 common approaches to defeat CAPTCHAs:

1. Using a service which solves CAPTCHAs through human labor
2. Exploiting bugs in the implementation that allow the attacker to bypass the CAPTCHA

2. Breaking CAPTCHAs

There are 3 common approaches to defeat CAPTCHAs:

1. Using a service which solves CAPTCHAs through human labor
2. Exploiting bugs in the implementation that allow the attacker to bypass the CAPTCHA
3. Character recognition software to solve the CAPTCHA

2. Breaking CAPTCHAs

There are 3 common approaches to defeat CAPTCHAs:

1. Using a service which solves CAPTCHAs through human labor
2. Exploiting bugs in the implementation that allow the attacker to bypass the CAPTCHA
3. **Character recognition software to solve the CAPTCHA**

2. Breaking CAPTCHAs - Dataset

Testing two common types of CAPTCHA:

Figure 2: CAPTCHAs set 1, generated using PHP

Figure 3: CAPTCHAs set 2, generated with Python

2. Breaking CAPTCHAs

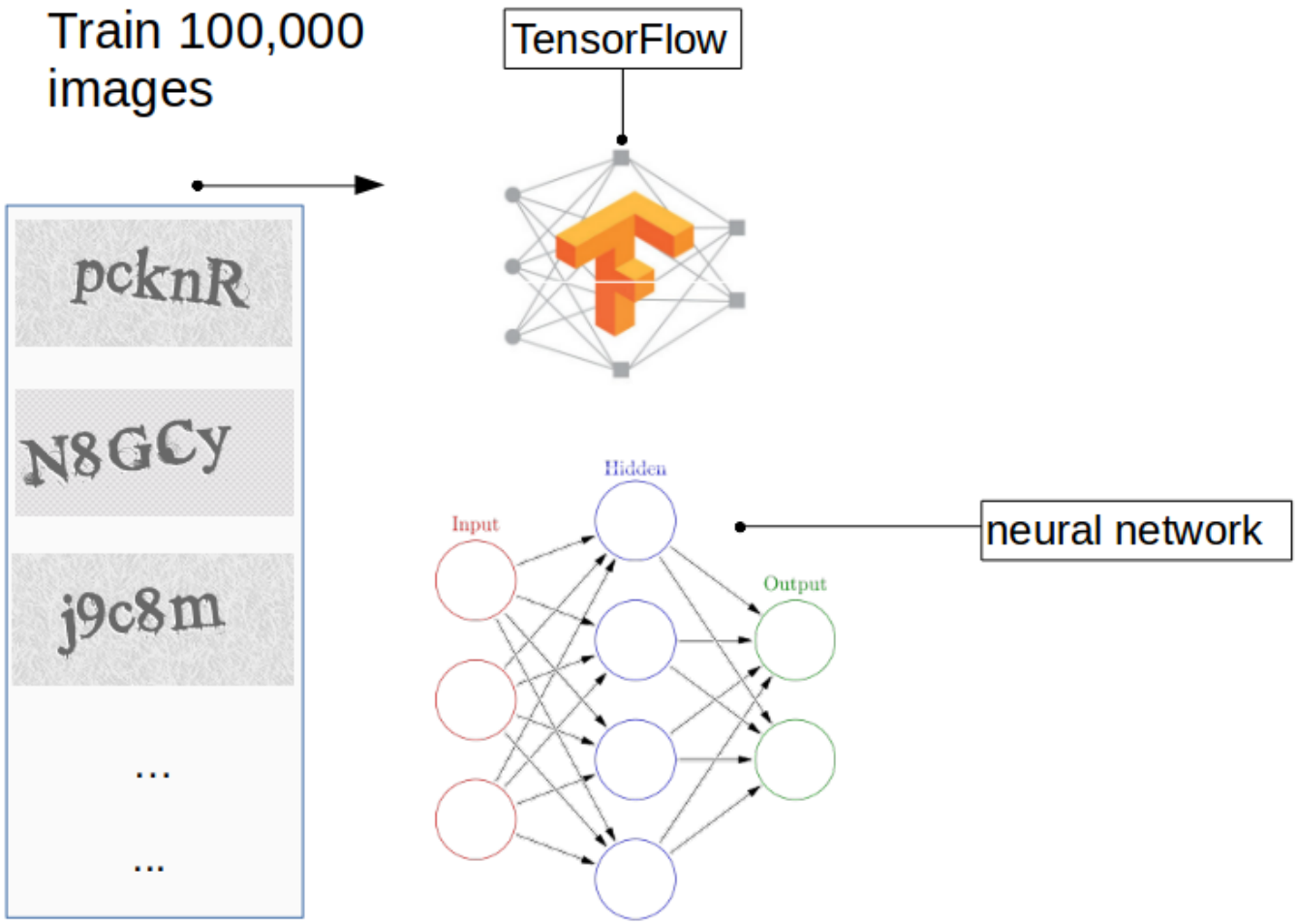


Figure 4: Training the neural network

Figure 5: Login web page with generated CAPTCHA

2. Breaking CAPTCHAs

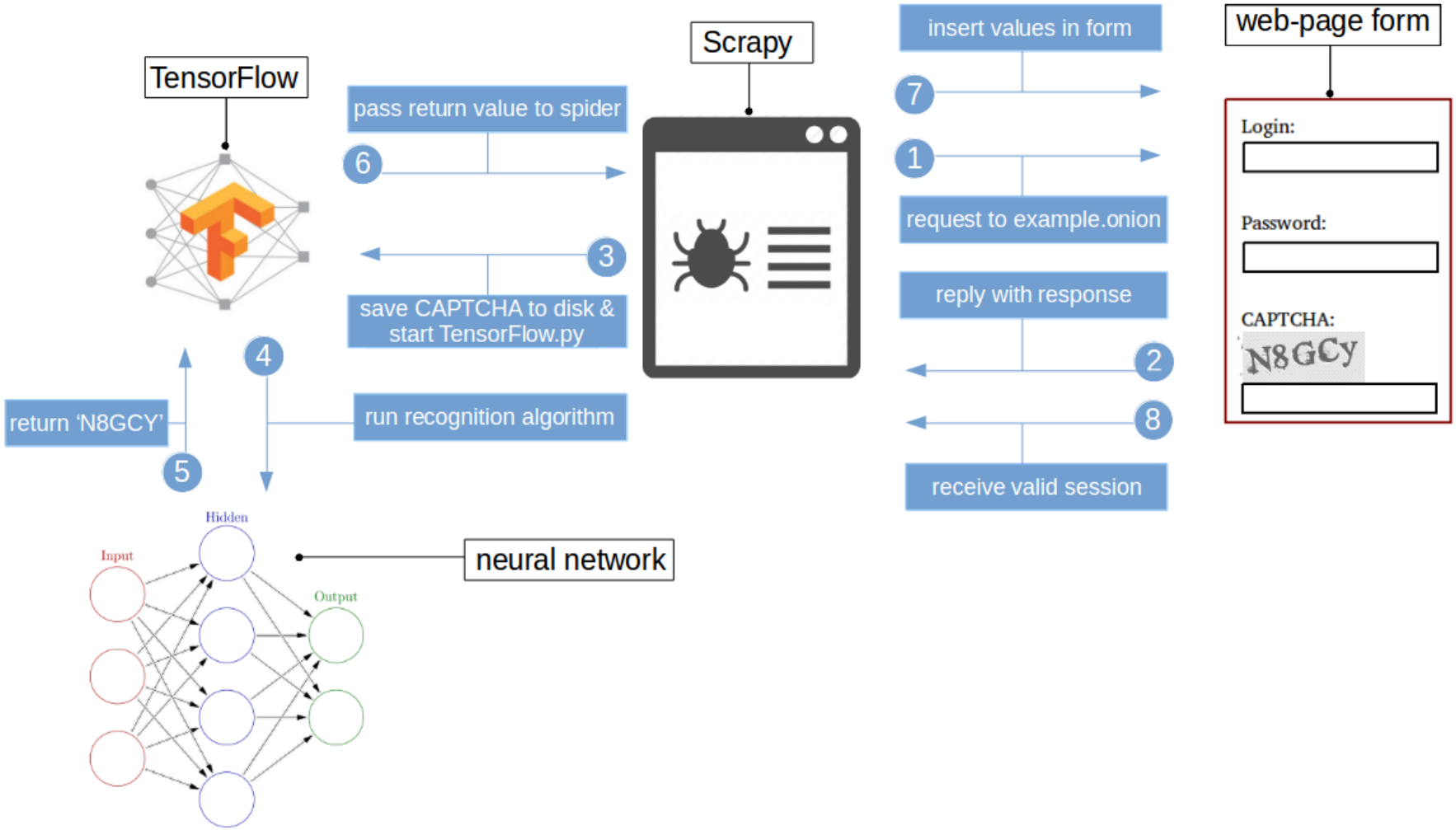


Figure 6: Workflow of solving CAPTCHA with TensorFlow via Scrapy

Results

1. Categorizing websites

1. Categorizing websites

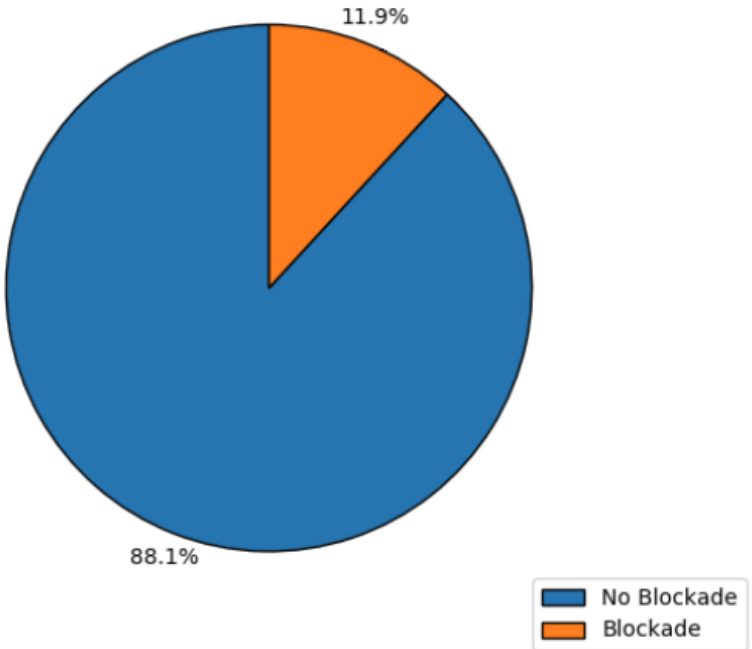


Figure 7: Percentage of scraping blockade using CAPTCHAs (n = 465)

1. Categorizing websites

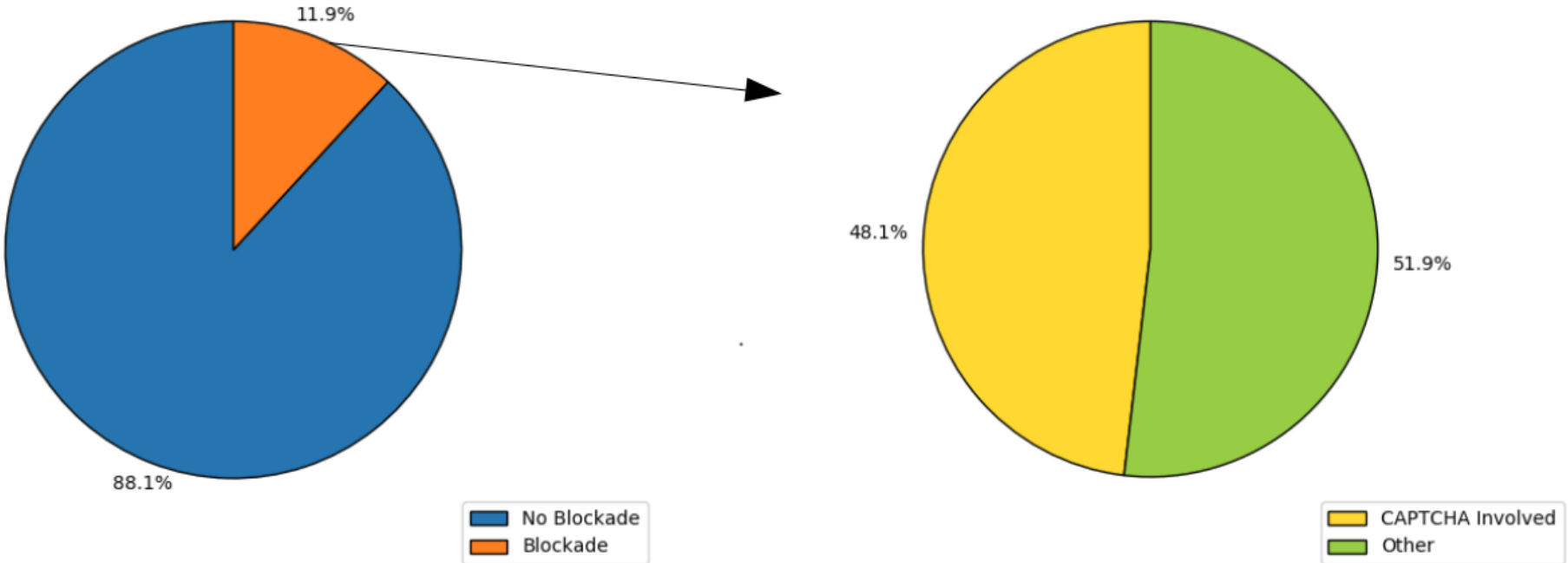


Figure 8: Percentage of scraping blockades using CAPTCHAs ($n = 465$, $n = 55$)

2. Breaking CAPTCHAs - TensorFlow vs. Tesseract

2. Breaking CAPTCHAs - TensorFlow vs. Tesseract

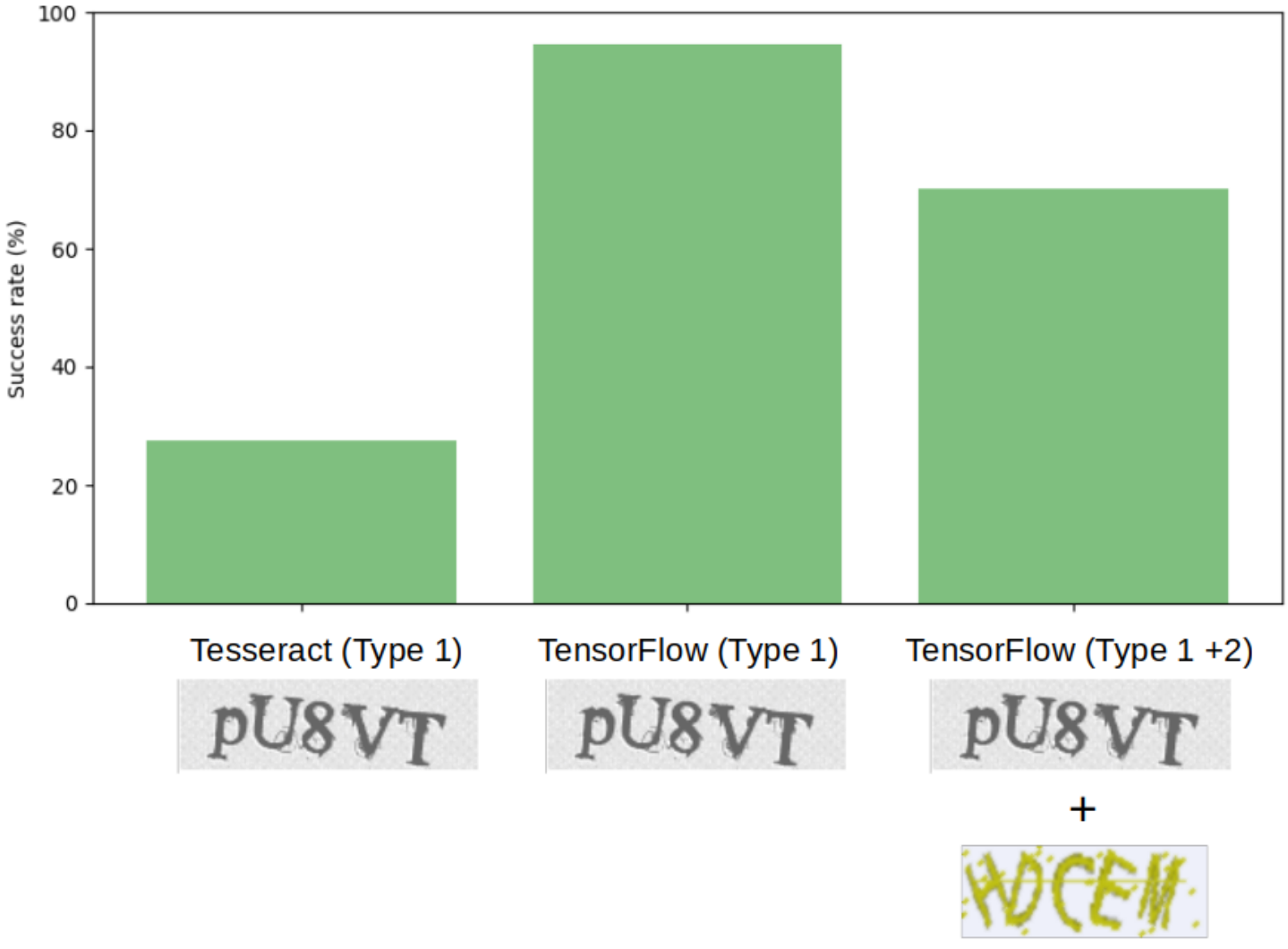


Figure 9: Success rate of Tesseract and TensorFlow (n = 1,000), higher is better

2. Breaking CAPTCHAs - TensorFlow vs. Tesseract

Levenshtein distance: minimal edit distance to get the correct result [5]

E.g. kitten to mitten $= 1$

2. Breaking CAPTCHAs - TensorFlow vs. Tesseract

Levenshtein distance: minimal edit distance to get the correct result [5]

E.g. kitten to mitten = 1

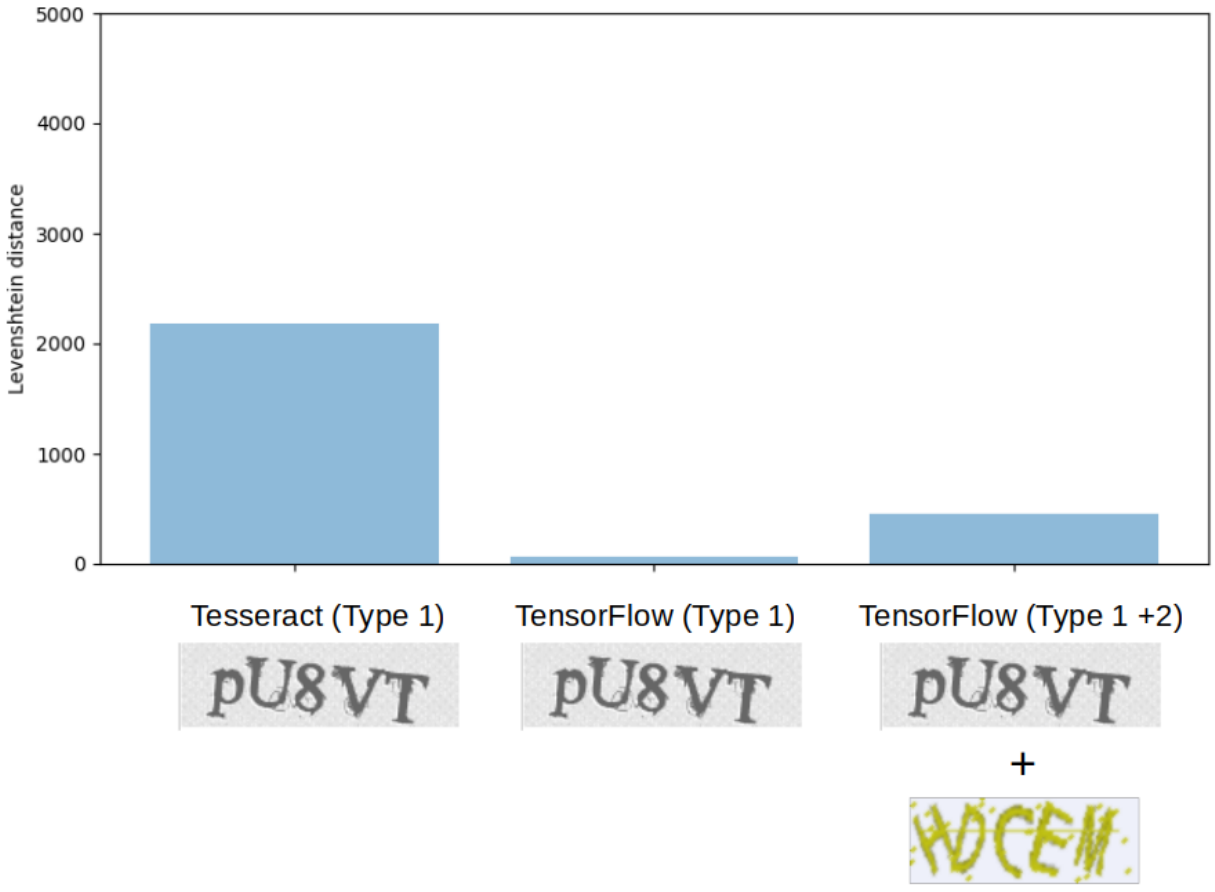


Figure 10: Combined Levenshtein distance, lower is better

Conclusion

Conclusion

- Circumventing CAPTCHAs is necessary to scrape blocked parts of websites

Conclusion

- Circumventing CAPTCHAs is necessary to scrape blocked parts of websites
- Machine Learning is most effective

Conclusion

- Circumventing CAPTCHAs is necessary to scrape blocked parts of websites
- Machine Learning is most effective
- However, if immediacy takes precedent over success rate and accuracy, then Tesseract (OCR) might be a better option

Future Research

Future Research

A more granular analysis of dark web websites:

Future Research

A more granular analysis of dark web websites:

- What content?

Future Research

A more granular analysis of dark web websites:

- What content?
- Any content hidden, due to lack of privileges?

Future Research

Increase readability for Tesseract by "cleaning up" the image

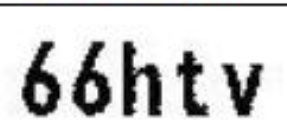
Original	Thresholded	OCR
		66htv

Figure 11: Removing noise from CAPTCHA [6]

Future Research

Achieve a more efficient training model, by using character segmentation

Figure 12: CAPTCHA character segmentation [7]

Future Research

Try more CAPTCHAs:

Future Research

Try more CAPTCHAs:

- Increased difficulty

Future Research

Try more CAPTCHAs:

- Increased difficulty
- If software to generate the CAPTCHAs, including the answers, is not available; send a training set to be solved by human labor. This costs money, $ 1,39 per 1,000 images [8]

Questions

?

References

[1] Lawrence, H., Hughes, A., Tonic, R., & Zou, C. (2017, October). D-miner: A framework for mining, searching, visualizing, and alerting on darknet events. In Communications and Network Security (CNS), 2017 IEEE Conference on (pp. 1-9). IEEE.

[2] Mitchell, R. (2015). Web scraping with Python: collecting data from the modern web. " O'Reilly Media, Inc.".

[3] Arun Patala. https://deepmlblog.wordpress.com/2016/01/03/how-to-break-a-captcha-system/

[4]people.cs.pitt.edu

[5]extremetech.com

[6]ahm3dibrahim.wordpress.com

[7] medium.com

[8] http://www.deathbycaptcha.com/