BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2017
http://excel.fit.vutbr.cz

Rozpoznávání CAPTCHA kódů

Radek Pazderka

Abstrakt
Tento článek se zaměřuje na rozpoznávání CAPTCHA kódů pomocí dvou odlišných klasifikátorů. Jedná se o histogramový klasifikátor a konvoluční neuronovou síť. U obou klasifikátorů uvádíme stručný postup jejich trénování a testování. V závěru článku jsou oba tyto klasifikátory porovnány s již existujícími přístupy k rozpoznávání CAPTCHA kódů. V diskuzi pak porovnáváme výhody a nevýhody jednotlivých metod.

Klíčová slova: CAPTCHA kód, strojové učení, konvoluční neuronová síť, LeNet, Caffe, histogramový klasifikátor

Přiložené materiály: Demonstrační video

*xpazde14@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Úvod

Každý, kdo chce v dnešní době vytvořit webovou stránku, buď pro sebe, nebo pro někoho na zakázku, měl by dbát na její bezpečnost. Hrozbu představují nelegální programy, které se na webové stránce vydávají za běžného uživatele, ale na rozdíl od něho odesílají větší počet a libovolný typ požadavků na webovou stránku. Webová stránka zpravidla slepě odpovídá na každý požadavek, který jí přijde, a nezjistí, že se nejedná o člověka, ale o nelegální program. Webový útočník, který vytvoří tento program, se zajímá pouze o profit nebo získání důležitých informací o systému a tím celou webovou stránku může znehodnotit.

Jako obrana proti těmto webovým zločincům vznikl CAPTCHA kód [1]. Hlavní myšlenka CAPTCHA kódů byla rozlišit, zda se jedná o reálného uživatele nebo o program, který se za reálného uživatele pouze vydává. Ochrana pomocí CAPTCHA kódů spočívá v tom, že je po uživateli vyžadován přepis textu z obrázku nebo odpověď na jednoduchou otázku, popř. identifikace předmětu v obrázku. Na tento triviální úkol člověk dokáže bez problémů odpovědět, ale pro program je tento úkol značně obtížnější. Postupem času se ale zvyšuje schopnost programů rozpoznávat větší množství druhů CAPTCHA kódů, proto je pro CAPTCHA kód důležité stále přicházet s novými bezpečnostními prvky, pomocí kterých se CAPTCHA kód stane na určitý čas pro programy nerozpoznatelný.

2. Motivace

V tomto článku se zaměříme na dva možné způsoby rozpoznání textových CAPTCHA kódů. První způsob rozpoznání je založen na podobnosti histogramů jednotlivých znaků. Druhý způsob je založen na klasifikaci znaků pomocí konvolučních neuronových sítí.

Protože existuje mnoho druhů textových CAPTCHA kódů a není možné získat všechny jejich datové sady, je nutné vytvořit aplikaci, která si natrénuje znaky z daných CAPTCHA kódů umístěných na webové stránce a následně je dokáže klasifikovat. Pro splnění tohoto účelu jsem vytvořil histogramový klasifikátor, který tento problém řeší.

Druhý způsob rozpoznávání textových CAPTCHA kódů je založen na klasifikaci znaků pomocí konvo-

lučních neuronových sítí. Pro klasifikaci je důležité mít připravenou datovou sadu znaků z CAPTCHA kódů, které se mají natrénovat. Jedná se o zdlouhavější postup trénování, než je u histogramového klasifikátoru, ale úspěšnost správného rozpoznávání u tohoto klasifikátoru je značně vyšší.

V rámci mé bakalářské práce byla vytvořena aplikace s grafickým uživatelským rozhraním, dále jen GUI, která podporuje oba způsoby rozpoznání.

3. Existující řešení

3.1 Optical Character Recognition

Optical Character Recognition, dále jen OCR [2], slouží k převodu skenovaného textu do digitální podoby. Na základě této technologie vzniklo několik programů, které rozpoznávají CAPCHA kódy. Jeden z nejlepších uváděných placených programů pro rozpoznávání kódů CAPTCHA je GSA Captcha Breaker, který umožňuje rozpoznání více než 600 druhů CAPTCHA kódů [3].

3.2 Rozpoznávání CAPTCHA kódů založené na levné pracovní síle

Existuje řada webových stránek, které nabízí rozpoznání CAPTCHA kódů za určitý obnos peněz. Cena se v dnešní době pohybuje okolo dvou dolarů za správné rozpoznání 1000 CAPTCHA kódů. Většina společností umožňuje i registraci a možnost výdělku určité sumy peněz při rozpoznávání CAPTCHA kódů. Za správné rozpoznání přibývají na účet velmi malé obnosy peněz.

Výhodou této služby je nízká cena za rozpoznání CAPTCHA kódů. Na druhou stranu zákazník musí počítat s faktem, že se jeho CAPTCHA kód bude rozpoznávat v řádech jednotek až desítek vteřin.

Této služby využívaly programy, které stahovaly videa z určitých serverů a každé video bylo chráněno CAPTCHA kódem. Valnou většinu zákazníků tvoří lidé, kteří se nechtějí zdržovat s opisováním CAPTCHA kódů, nebo weboví útočníci, kteří útočí na webové stránky chráněné velice silným CAPTCHA kódem, na které neexistuje volně dostupný program, který by s velkou úspěšností daný druh CAPTCHA kódu rozpoznal.

4. Vlastní řešení

V této kapitole si popíšeme klasifikační algoritmy, které jsou v této práci použity. Konkrétně se jedná se o konvoluční neuronovou síť LeNet a histogramový klasifikátor využívající vzorce cosine distance s mean normalizací.

4.1 Histogramový klasifikátor

Klasifikace se provádí pomocí podobnosti histogramů. Vždy se porovnává histogram klasifikovaného znaku z CAPTCHA kódu a histogram znaku z trénovací sady. Základem pro klasifikaci je funkce pro porovnání dvou histogramů. Jedná se o funkci cosine distance s mean normalizací z knihovny OpenCV [4][5]. Tato funkce dokáže určit shodnost dvou stejně širokých histogramů. Algoritmus klasifikace jsem musel modifikoval tak, aby byl schopen ohodnotit shodnost dvou histogramů obsahující libovolný počet sloupců. Modifikace algoritmu spočívá v umělém rozšíření histogramů o nové sloupce tak, aby vznikly dva histogramy obsahující stejný počet sloupců. Algoritmus si však pamatuje, který histogram původně obsahoval méně sloupců a tento histogram posouvá pod původně mohutnějším histogramem. Při každém posunutí se spočítá skóre podobnosti histogramů pomocí vzorce (1). Po vyzkoušení všech možností se vybere možnost s nejvyšším skóre podobnosti.

$$d(H_1,H_2) = \frac{\sum_{i=0}^{I}(H_1(i)-\overline{H_1})(H_2(i)-\overline{H_2})}{\sqrt{\sum_{i=0}^{I}(H_1(i)-\overline{H_1})^2 \sum_{i=0}^{I}(H_2(i)-\overline{H_2})^2}} \quad (1)$$

kde:

- $\overline{H_k}$ vyjadřuje průměr všech hodnot v daném histogramu:
  - $\overline{H_k} = \frac{1}{N}\sum_{j=0}^{N} H_k(j)$.
  - $k = 1,2$.
  - N udává počet sloupců v histogramu.
- I udává počet sloupců histogramů H_1 a H_2.
- $d(H_1,H_2)$ je vzorec pro cosine distance s mean normalizací. Vrací skóre podobnosti dvou histogramů, kde $d \in <-1,1>$.

Datová sada pro histogramový klasifikátor se skládá z několika CAPTCHA kódů, ve kterých se vyskytují všechny alfanumerické znaky, které jsou v daném druhu CAPTCHA kódu použity. Při vytváření této datové sady je důležité dbát na to, aby se znaky v CAPTCHA kódu nepřekrývaly. Pokud by tomu tak bylo, aplikace by znaky z daného CAPTCHA kódu nemohla použít pro trénování a celý CAPTCHA kód by přeskočila.

Trénování histogramového klasifikátoru probíhá v několika fázích. V první fázi se musí načíst datová sada CAPTCHA kódů do aplikace. Každý CAPTCHA kód z datové sady by měl být pojmenovaný tak, aby jeho název odpovídal textu na CAPTCHA kódu. V případě, když jednotlivé CAPTCHA kódy nejsou korektně pojmenované, aplikace poskytuje rozhraní

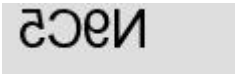
pro přejmenování a následné natrénování všech CAPTCHA kódů z datových sad. V druhé fázi může uživatel zvýšit rychlost a přesnost klasifikace tím, že CAPTCHA kódy rozdělí do skupin podle druhů CAPTCHA kódů. Ve třetí fázi uživatel vytvoří název pro soubor s trénovacími daty a znaky z CAPTCHA kódů nechá natrénovat. Časová náročnost trénování všech znaků z jednoho CAPTCHA kódu se pohybuje v jednotkách milisekund.

Testování. Histogramový klasifikátor jsem testoval na CAPTCHA kódech v odstínu šedé. V tomto CAPTCHA kódu se písmena mohou dotýkat nebo mírně překrývat. I přes to, že CAPTCHA kód obsahuje málo bezpečnostních prvků, se často tento druh CAPTCHA kódu objevuje na mnoha webových stránkách. Toto usnadňuje práci programům, které se snaží rozpoznávat CAPTCHA kódy.

Zvolený CAPTCHA kód se rozděluje do tří skupin. Jedná se o druh s deformovanými hranami, (Obrázek 1), zrcadlově otočenými písmeny, (Obrázek 2), a tučně psaným textem, (Obrázek 3).

Histogramový klasifikátor jsem testoval na 200 náhodně stažených CAPTCHA kódech. Testy ukázaly, že histogramový klasifikátor dokázal rozpoznat všechny CAPTCHA kódy se **100%** úspěšností. Při testování rychlosti rozpoznání jednoho CAPTCHA kódu se čas pohyboval v jednotkách milisekund.

Obrázek 1. CATPCHA kód s deformovanými hranami.



Obrázek 2. CAPTCHA kód se zrcadlově otočeným písmem.

Obrázek 3. CAPTCHA kód s tučným písmem.

4.2 Konvoluční neuronové sítě

Konvoluční neuronové sítě (Convolutional Neural Networks, dále jen CNN) jsou velmi populárním nástrojem pro detekci nebo klasifikaci objektů v obraze [6]. V této práci jsem použil konvoluční síť LeNet, která byla navrhnuta pro klasifikaci jednoho znaku ze vstupního obrazu. Konvoluční síť LeNet jsem trénoval pomocí frameworku Caffe.

Datová sada použitá v mé aplikaci pro CNN obsahuje 224 333 obrázků alfanumerických znaků. Z těchto obrázků je 180 000 znaků získaných z CAPTCHA kódů. Od každého znaku jsem získal 5 000 vzorků. Dále se datová sada skládá z 36 567 obrázků znaků napsaných různými fonty, 1 979 obrázků ručně psaných znaků a 5 787 obrázků obsahující vyseparované znaky z fotek. Různé formáty datových sad jsem vybíral proto, aby CNN byla schopná klasifikovat co nejvyšší možné množství alfanumerických znaků. Na Obrázku 4 je uvedena ukázka dvou vzorků z každé uváděné datové sady.

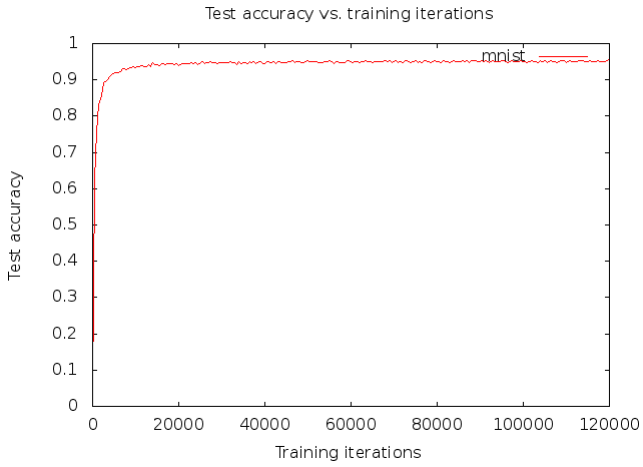
Použil jsem datovou sadu Chars74K [7].

Obrázek 4. Ukázka datové sady pro CNN.

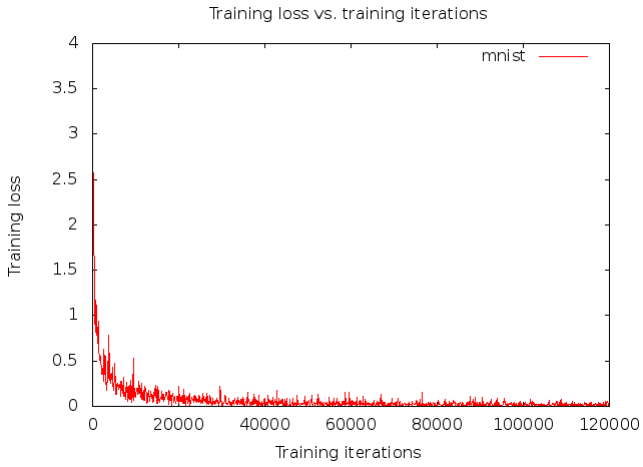
Trénování CNN se rozděluje do několika částí. V první části se musí vstupní datová sada znaků rozdělit na testovací a trénovací datovou sadu v doporučeném poměru 1:5. Tyto dvě datové sady se musí transformovat do LMDB[1] databází. Dále se musí pomocí transformovaných datových sad vytvořit soubor, který udává váhy bodů v datových sadách. V další části se pomocí konfiguračních souborů a frameworku Caffe začne trénovat CNN. Trénování trvalo 9 hodin[2]. Výstupem trénování CNN jsou dva grafy. Na Obrázku 5 je uveden graf závislosti přesnosti na počtu iteracích CNN. Podle tohoto grafu můžeme zjistit, že úspěšnost CNN na testovací sadě se od hranice 20 000 iterací téměř nemění a stále se drží okolo 95 %. Na Obrázku 6 je uveden graf závislosti výstupu chybové funkce na iteracích CNN. Chybová funkce udává míru nepřesnosti v nastavení vah sítě. Cílem je nalézt takovou iteraci, ve které je výstup chybové funkce minimální.

Do aplikace jsem vybral caffemodel s počtem iterací 55 000, který má 95,22% přesnost na testovací datové sadě.

[1] Lightning Memory-Mapped Database.
[2] Na CPU Intel Core(TM) i7-3630QM, 2.4GHz

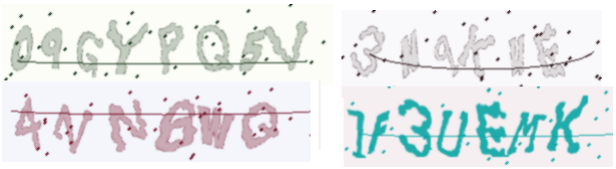


Obrázek 5. Graf závislosti přesnosti CNN na počtu trénovacích iterací.



Obrázek 6. Graf závislosti chybové funkce na počtu trénovacích iterací.

Testování CNN proběhlo na CAPTCHA kódech vygenerovaných pomocí oficiálního Python generátoru CAPTCHA kódů [8]. Tento generátor využívá mnoho webových stránek z důvodu možnosti vygenerování CAPTCHA kódu z vlastních fontů písma a tím si vytvořit unikátní CAPTCHA kód. Tento generátor u CAPTCHA kódu deformuje celé znaky, mění jim velikost, úhel postavení, barvu a její intenzitu. Jako další bezpečnostní prvek používá vodorovnou čáru a rušivý šum v pozadí. Na Obrázku 7 je ukázka použitého CAPTCHA kódu.



Obrázek 7. Ukázka CAPTCHA kódů z testovací sady pro CNN.

Testování proběhlo na 600 náhodně vytvořených CAPTCHA kódech, ve kterých je proměnlivý počet znaků. Nastavil jsem rozmezí 4-8 znaků na CAPTCHA kód. Aplikace dosáhla **53,5%** úspěšnosti, což několikanásobně překonalo minimální 1% úspěšnost, která je oficiální hranicí pro úspěšné rozpoznání jednoho druhu CAPTCHA kódu [9]. Průměrný čas potřebný k rozpoznání jednoho CAPTCHA kódu se pohybuje okolo 2 vteřin.

5. Přínos

Výhoda histogramového klasifikátoru spočívá ve vysoké rychlosti trénování a klasifikace znaků, která se pohybuje v jednotkách milisekund. Další velkou výhodou je možnost stažení datové sady CAPTCHA kódů přímo z webové stránky, na které se daný CAPTCHA kód nachází. Pomocí této nově stažené datové sady si uživatel může vytvořit klasifikátor, který bude daný CAPTCHA kód rozpoznávat. Rozpoznávat se dají i CAPTCHA kódy, které mají spojené znaky. Na tento případ je histogramový klasifikátor připravený a podle natrénovaných znaků pozná, že jsou 2 nebo více znaků spojených a pokusí se je rozdělit a následně klasifikovat. Naopak nevýhodou je jeho nízká úspěšnost při rozpoznávání CAPTCHA kódů obsahující špatně čitelné znaky. V tomto případě je možné v aplikaci změnit klasifikaci na CNN.

Výhoda CNN klasifikátoru spočívá v možnosti rozpoznávání málo čitelných textových CAPTCHA kódů. CNN umožňuje rozpoznat i takové znaky, u kterých ani člověk nedokáže s jistotou říci, o jaký znak se jedná[3]. CNN podporuje také rozpoznávání ručně psaných znaků. Menší nevýhodou je nižší rychlost rozpoznávání jednoho CAPTCHA kódu, která se pohybuje v průměru okolo 2 vteřin.

5.1 Porovnání vlastního přístupu k rozpoznání CAPTCHA kódů s již existujícími přístupy

Za silné stránky své aplikace považuji možnost její adaptace na různé druhy CAPTCHA kódů. Mnoho aplikací neumožňuje natrénovat si vlastní klasifikátor na daný druh CAPTCHA kódu a následně natrénovaný klasifikátor použít pro např. útok na danou webovou stránku. Další silnou stránkou aplikace je možnost klasifikovat znaky v CAPTCHA kódu, u kterých si ani člověk nemusí být jistý. Dále je možné pomocí aplikace stahovat datové sady CAPTCHA kódů umístěných na libovolných webových stránkách.

6. Závěr

Tento článek popsal dva možné způsoby rozpoznání CAPTCHA kódů. Nastínili jsme stručný postup trénování a testování zvolených klasifikátorů. Při testování

[3] Nejčastěji se jedná o znaky: 0-O, 2-Z, 5-6, apod.

histogramového klasifikátoru na zvoleném CAPTCHA kódu jsem získal 100% úspěšnost. Při testování CNN na obtížnějším CAPTCHA kódu jsem obdržel 53,5% úspěšnost.

Poděkování

Rád bych poděkoval mému vedoucímu panu Doc. Ing. Františku Zbořilovi, CSc za jeho pomoc při psaní tohoto článku.

Literatura

[1] Anjali Avinash Chandavale and A. Sapkal. *Recent Trends in Computer Networks and Distributed Systems Security: International Conference, SNDS 2012, Trivandrum, India, October 11-12, 2012. Proceedings*, chapter Security Analysis of CAPTCHA, pages 97–109. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[2] Piotr Lech and Krzysztof Okarma. *Methods of Natural Image Preprocessing Supporting the Automatic Text Recognition Using the OCR Algorithms*, pages 143–150. Springer International Publishing, Cham, 2016.

[3] 7 Best CAPTCHA Solvers. 7 best captcha solvers. http://www.slideshare.net/marriagenamechange/7-best-captcha-solvers, 2017. [Online; navštíveno 01-04-2017].

[4] Samarth Brahmbhatt. *Introduction to Computer Vision and OpenCV*, pages 3–5. Apress, Berkeley, CA, 2013.

[5] OpenCV. Histograms opencv 2.4.12.0 documentation. http://docs.opencv.org/2.4/doc/tutorials/imgproc/histograms/histogram_comparison/histogram_comparison.html, 2017. [Online; navštíveno 28-03-2017].

[6] L. D. Jackel, M. Y. Battista, J. Ben, J. Bromley, C. J. C. Burges, H. S. Baird, E. Cosatto, J. S. Denker, H. P. Graf, H. P. Katseff, Y. LeCun, C. R. Nohl, E. Sackinger, J. H. Shamilian, T. Shoemaker, C. E. Stenard, B. I. Strom, R. Ting, T. Wood, and C. R. Zuraw. *Neural Network Applications in Character Recognition and Document Analysis*, pages 271–285. Springer US, Boston, MA, 1994.

[7] The Chars74K. Character recognition in natural images. http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/, 2017. [Online; navštíveno 02-04-2017].

[8] Python. A captcha library that generates audio and image captchas. https://pypi.python.org/pypi/captcha/0.1.1, 2017. [Online; navštíveno 02-04-2017].

[9] Elie Bursztein, Matthieu Martin, and John Mitchell. Text-based captcha strengths and weaknesses. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*, CCS '11, pages 125–138, New York, NY, USA, 2011. ACM.