

#10: Rozpoznávání CAPTCHA kódů

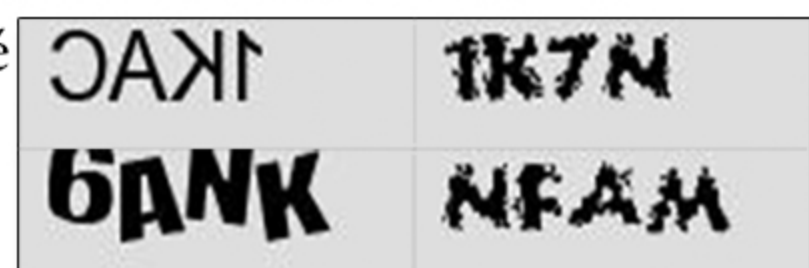
pomocí konvolučních neuronových sítí a histogramového klasifikátoru

Datové sady

Datová sada pro CNN se skládá z 224 333 vzorků alfanumerických znaků. Obsahuje znaky získané z CAPTCHA kódů, znaky různých fontů, ručně psané znaky a znaky vyseparované z fotek.



Datová sada pro histogramový klasifikátor se skládá z několika CAPTCHA kódů, ve kterých se vyskytují všechny alfanumerické znaky, které jsou v daném CAPTCHA kódu použity.



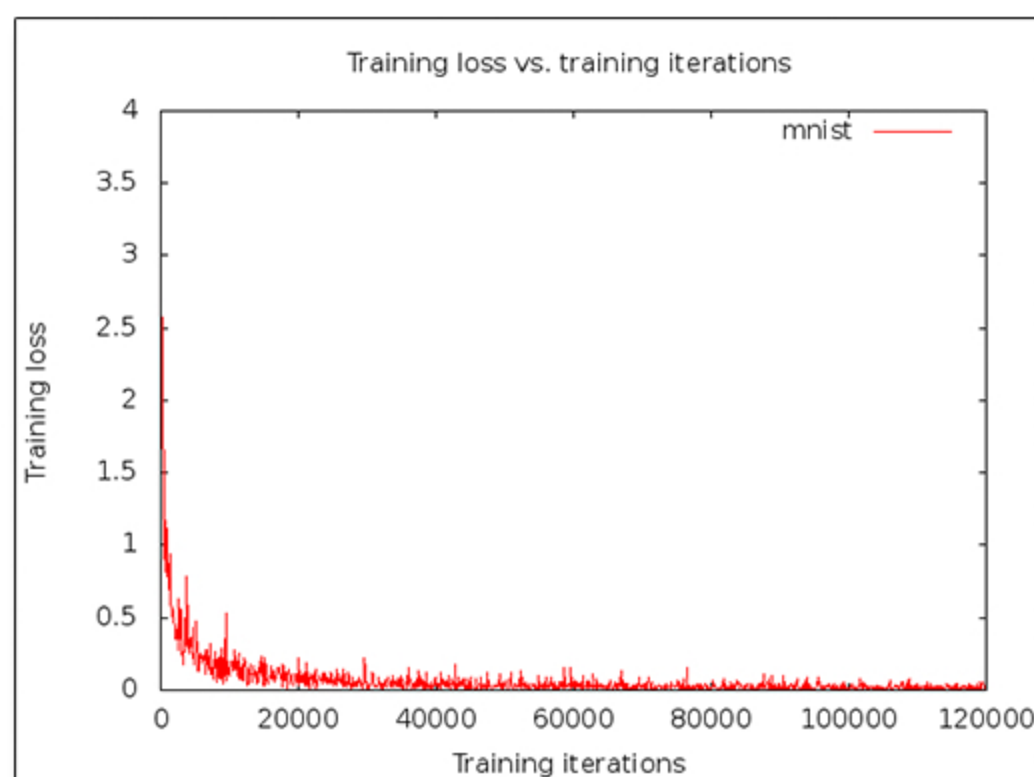
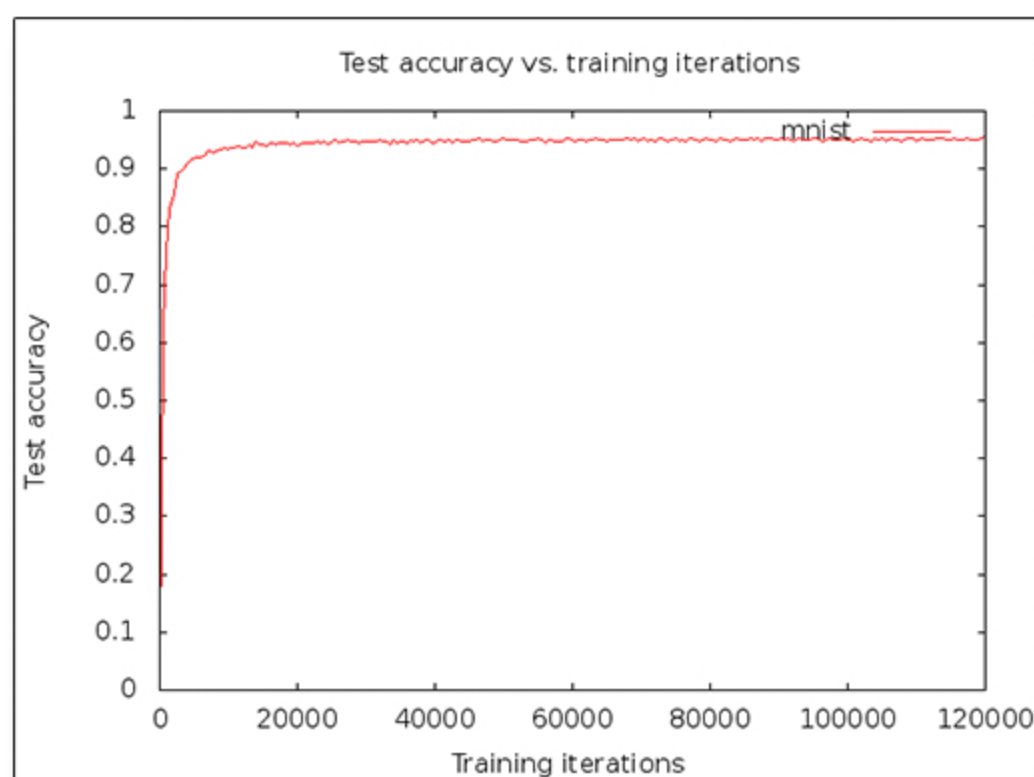
Segmentace znaků

Algoritmus semínkového vyplnění. Tento algoritmus slouží k separaci stejnobarevného tvaru v rastrovém obrázku. Postupně se vybírají sousedé zvoleného bodu stejné barvy a přenáší se do nového obrázku, čímž se oddělí jeden tvar od ostatních. Varianty tohoto algoritmu se liší podle výběru sousedních bodů.

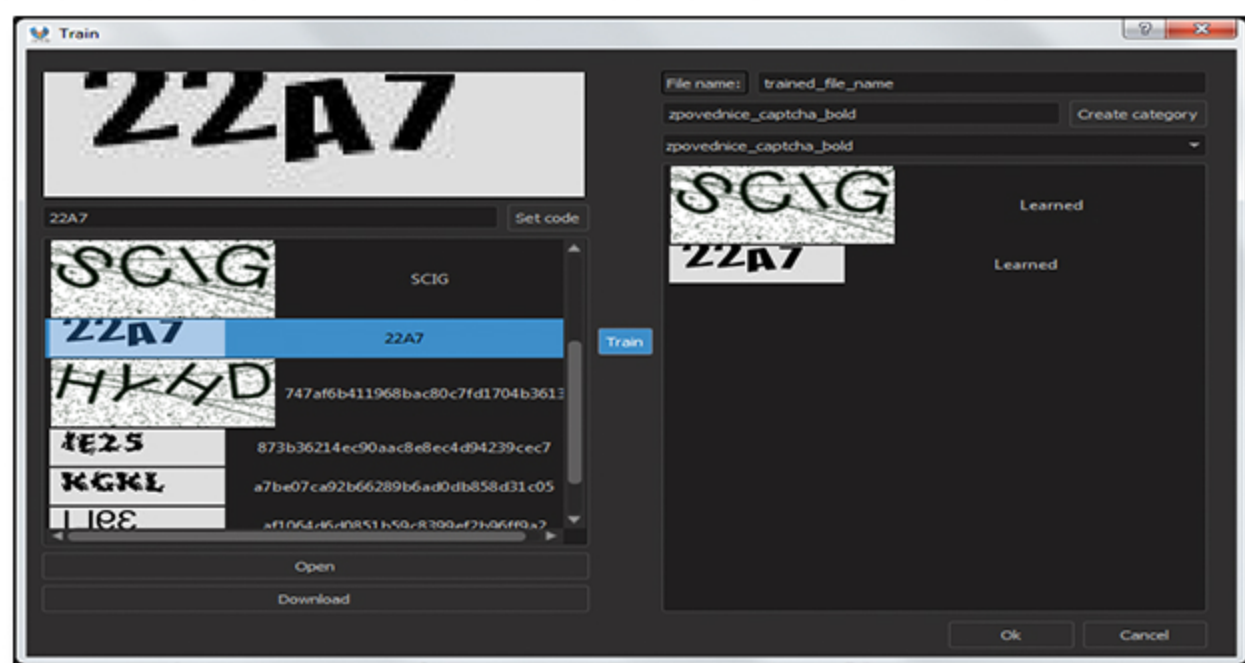
Histogramový segmentační algoritmus. Tento algoritmus slouží k segmentaci objektů z obrazu pomocí histogramu. Nejprve získá horizontální histogram ze vstupního obrazu součtem intenzit pixelů v jednotlivých sloupcích. Následně se vertikálními řezy rozděluje vstupní obraz podle nejnižších hodnot v histogramu.

Trénování klasifikátorů

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 databází. Dále se musí pomocí LMDB vytvořit soubor, který udává váhy bodů v datových sadách. V další části se pomocí konfiguračních souborů a frameworku Caffe začne trénovat CNN. Výstupem trénování CNN jsou dva grafy. Graf závislosti přesnosti CNN na počtu iterací CNN. Druhý graf udává závislost výstupu ztrátové funkce na iteracích CNN. Ztrátová funkce udává míru nepřesnosti v nastavení vah sítě. Cílem je nalézt takovou iteraci, ve které je výstup ztrátové funkce minimální.



Trénování histogramového klasifikátoru zajišťuje vytvořená aplikace. Umožňuje získání všech histogramů znaků z CAPTCHA kódů a následné uložení do externího souboru.



Předzpracování obrazu

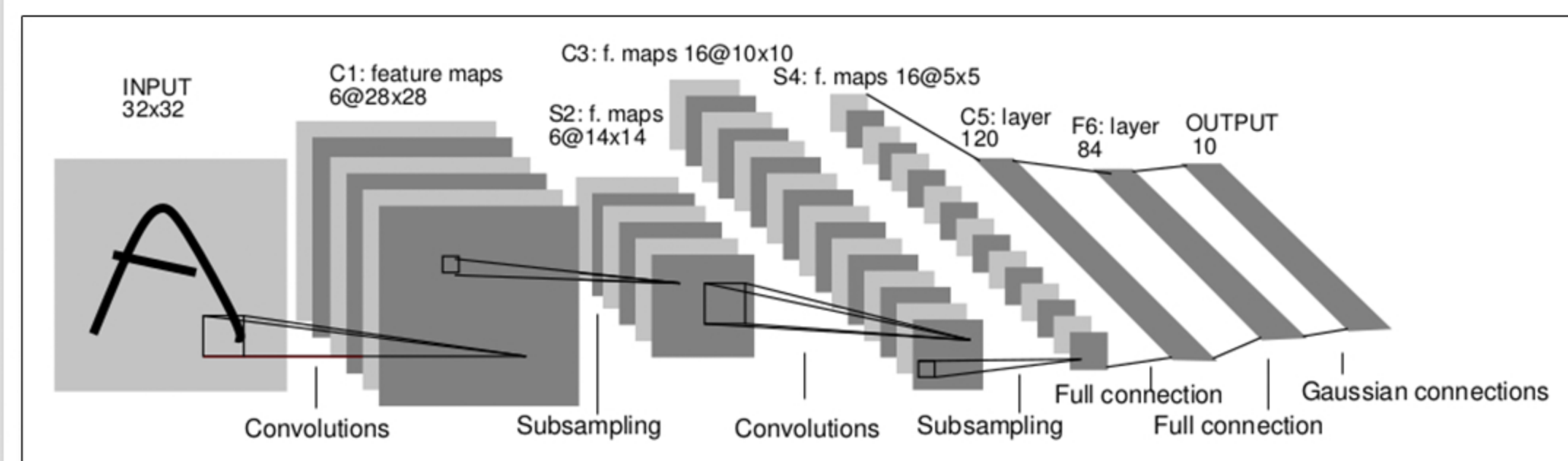
Algoritmus prahování. Jedná se o základní algoritmus segmentace obrazu založený na hodnocení jasu každého pixelu. Základem algoritmu je hodnota prahu, podle které se určuje, na jakou barvu se má daný pixel nastavit.

Gaussovo rozostření. Aplikováním Gaussova filtru na obraz získáme obraz rozostřený. Toto rozostření se provádí pomocí Gaussovy funkce. Používá se pro snížení šumu a detailů v obraze. Gaussovo rozostření použijeme jako předzpracování obrazu před Otsu metodou prahování.

Otsu metoda prahování převádí obraz v odstínu šedi na binární. Algoritmus hledá optimální hodnotu prahu, kterým oddělí popředí a pozadí obrazu. Použijeme ho k oddělení textu v CAPTCHA kódu a jeho pozadí. Při hledání prahu iteruje přes všechny možné hodnoty prahu. Při každé iteraci se obraz rozdělí do dvou tříd - popředí a pozadí obrazu a hledá hodnotu prahu, kde je vážený součet těchto tříd minimální.

Klasifikace znaků

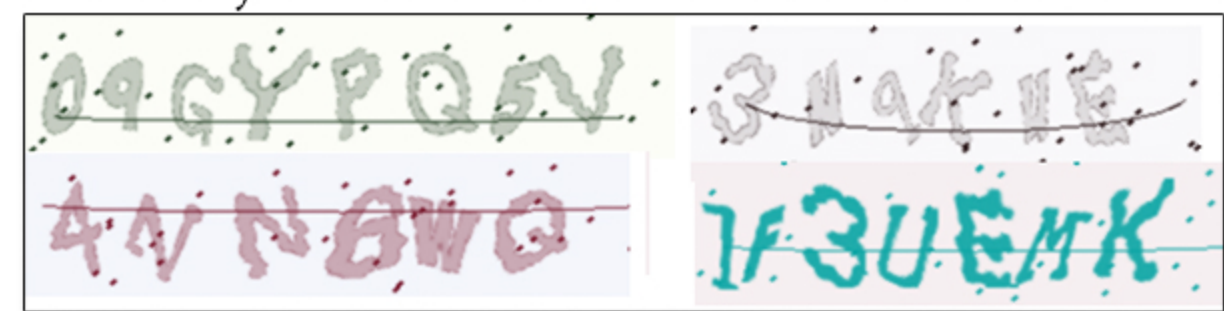
Konvoluční neuronové sítě (Convolutional Neural Networks, dále jen CNN) jsou velmi populárním nástrojem pro detekci nebo klasifikaci objektů v obraze. CNN jsou tvořeny z navzájem propojených vrstev. Každá vrstva v CNN aplikuje filtr na svá vstupní data a výstup posílá do následující vrstvy. Pro trénování a testování vlastní CNN jsem použil síť LeNet ve frameworku Caffe.



Histogramový klasifikátor. Klasifikace se provádí pomocí podobnosti histogramů. Vždy se porovnává histogram klasifikovaného tvaru a histogram tvaru z trénovací sady. Základem pro klasifikaci je funkce pro porovnání dvou histogramů se stejným počtem sloupců z knihovny OpenCV. Algoritmus klasifikace jsem však modifikoval tak, aby byl schopen procentuálně ohodnotit shodnost dvou histogramů obsahující libovolný počet sloupců.

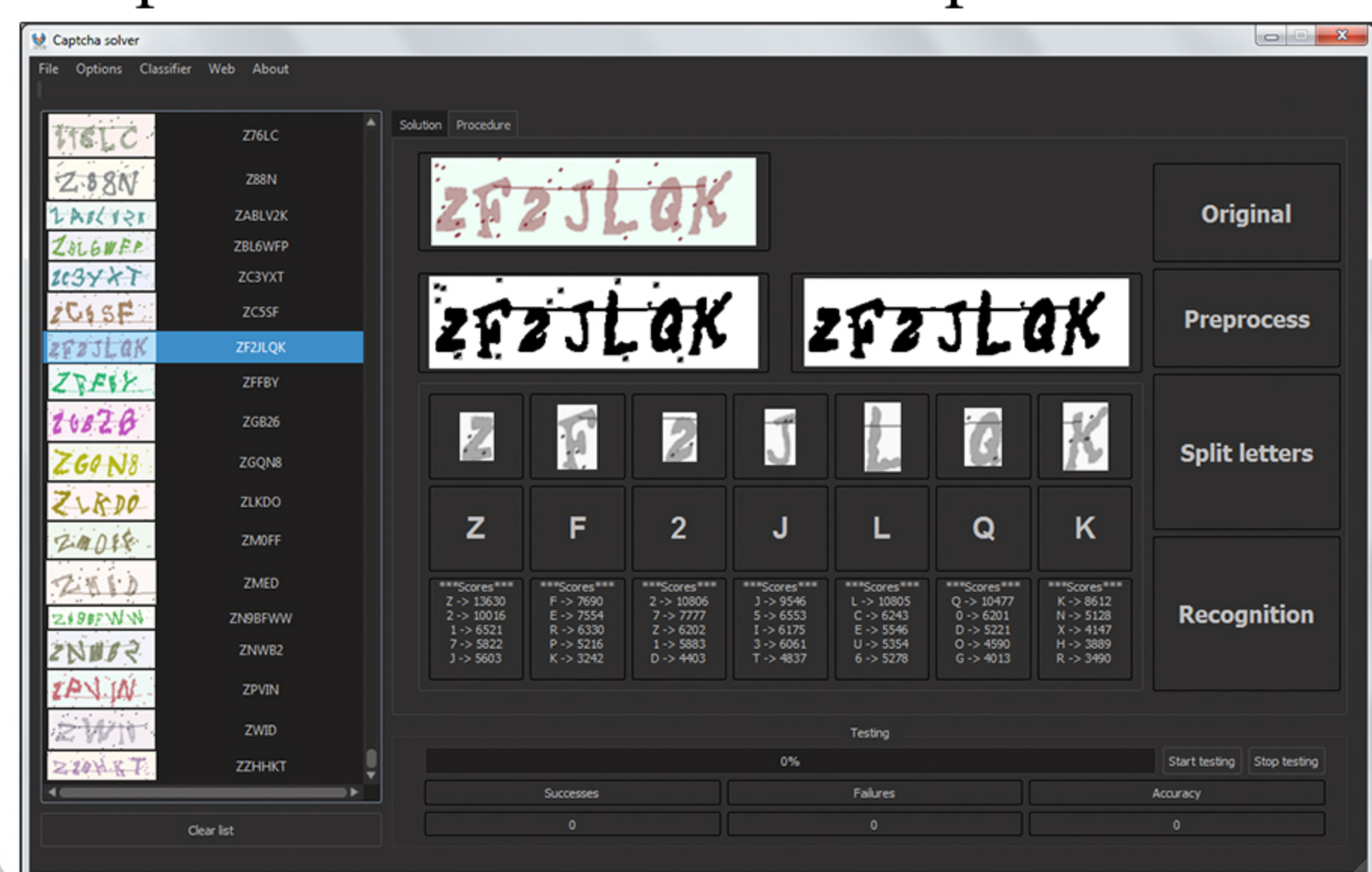
Testování

Testování CNN proběhlo na 600 náhodně vytvořených CAPTCHA kódech, ve kterých se náhodně mění počet znaků. Nastavil jsem rozmezí 4-8 znaků na CAPTCHA kód. Aplikace dosáhla **53,5%** úspěšnosti. Průměrný čas potřebný k rozpoznání jednoho CAPTCHA kódu se pohyboval okolo 2 vteřin.



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.

Rozpoznání CAPTCHA kódu pomocí CNN



Stahování datových sad CAPTCHA kódů

