Sprawozdanie z laboratorium: Komunikacja człowiek – komputer

Temat: *Rozpoznawanie emocji na podstawie mimiki twarzy z użyciem konwolucyjnej sieci neuronowej.*

3 grudnia 2019

Prowadzący: mgr Agnieszka Mensfelt

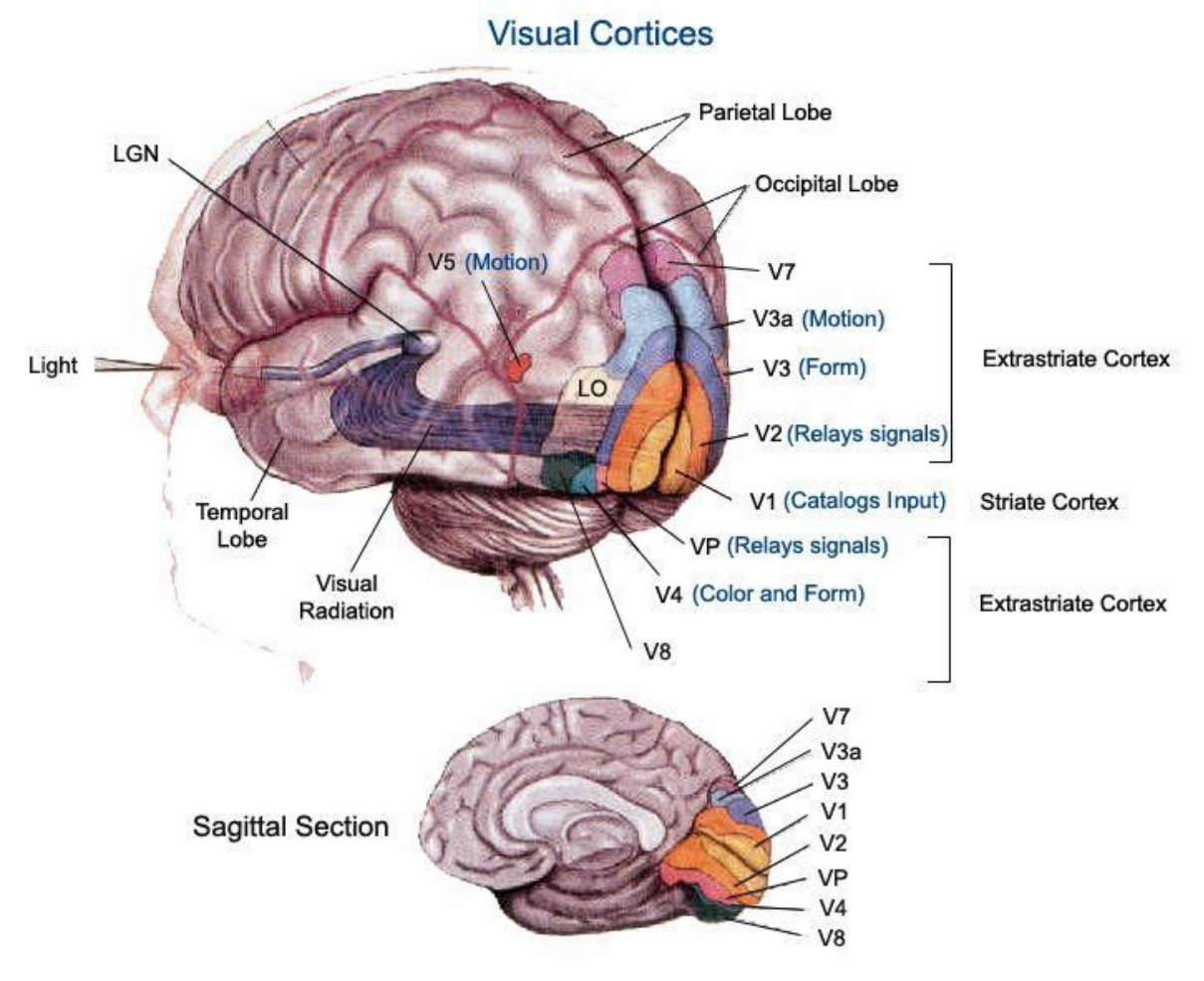
Autorzy: Paweł Kryczka 136748
Łukasz Duhr 136700

Zajęcia: wtorek, 18:30

Oświadczamy, że niniejsze sprawozdanie zostało przygotowane wyłącznie przez powyższych autora/ów, a wszystkie elementy pochodzące z innych źródeł zostały odpowiednio zaznaczone i są cytowane w bibliografii.

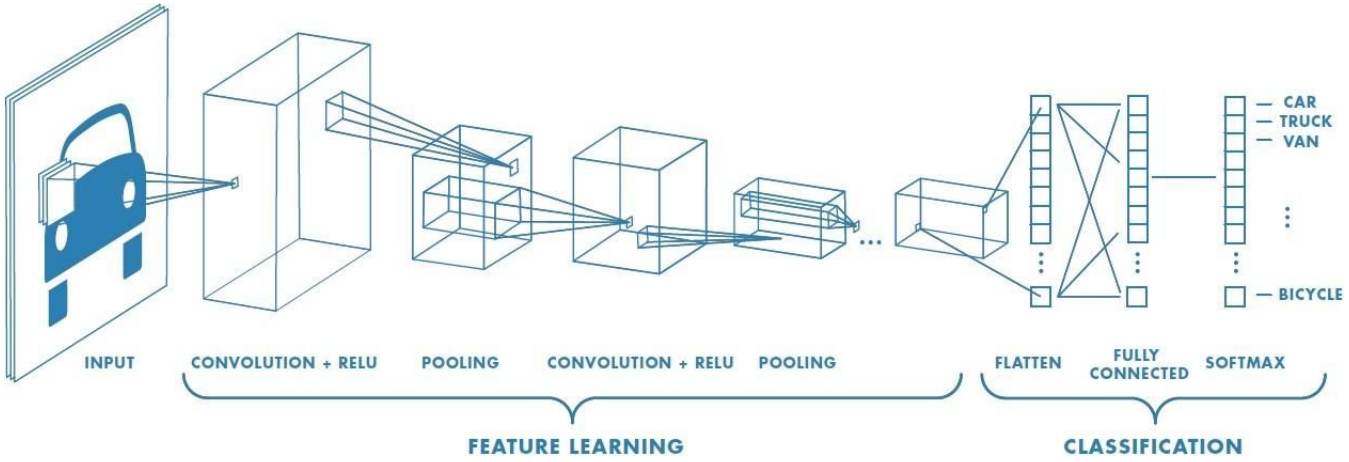
1. Wstęp do sieci konwolucyjnych.

Sieci konwolucyjne to typ sieci neuronowych, które znalazły zastosowanie w dziedzinie przetwarzania obrazów, a dokładniej w rozpoznawaniu obiektów. Ich architektura jest zainspirowana przez działanie kory wzrokowej u ssaków. Organ ten jest odpowiedzialny za przetwarzanie bodźców wzrokowych, mianowicie sygnały z nerwów wzrokowych są bezpośrednio połączone z pierwszą warstwą kory wzrokowej, która podzielona jest na kilka warstw. Informacje przekazywane są z warstwy do warstwy, a każda kolejna warstwa jest bardziej wyspecjalizowana niż jej poprzedniczka. Neurony w danej warstwie reagują wyłącznie na określone bodźce.



1. Wyszczególnione warstwy kory wzrokowej.

2. Architektura sieci konwolucyjnych.



2. Schemat graficznej reprezentacji sieci konwolucyjne[1]

Podstawowa sieć konwolucyjna składa się z warstw:

1. Obrazu wejściowego.
2. Warstwy konwolucyjnej.
3. Warstwa aktywacji, najczęściej z użyciem funkcji ReLU.
4. Warstwy łączenia (pooling).
5. Warstwy połączeń każdy z każdym.

Warstwa konwolucyjna jest głównym budulcem sieci konwolucyjnej. Składa się ona z niezależnych filtrów oraz jądra. Dane wejściowe są przetwarzane dla każdego filtru osobno, przez jego jądro, które tworzy tablicę danych wynikowych przesuwając się po danych wejściowych. Z każda taka operacją, wysokość i szerokość danych zmniejsza się, a ich głębokość rośnie. Głównym zadaniem warstwy konwolucyjnej jest przefiltrowanie danych w taki sposób, aby umożliwić wyselekcjonowanie dla nich cech charakterystycznych, takich jak krawędzie lub kolory z nimi związane.

Warstwa łączenia często używana jest zaraz po warstwie konwolucyjnej, a jej zadaniem jest stopniowe zmniejszenie zużycia pamięci w głębszych warstwach, jak i wyłuskanie najważniejszych cech z danych. Warstwa ta pomaga również kontrolować problem przeuczania sieci, ponieważ zmniejsza liczbę parametrów do zapamiętania. Najbardziej popularne strategie dla warstwy łączenia to „max-pooling" i „average-pooling", które próbkują przestrzeń danych wejściowych względem wysokości i szerokości, zwracając na wyjściu pomniejszoną macierz danych.

Warstwa aktywacji z użyciem funkcji ReLU (rectified linear unit), odpowiada za aktywację neuronów w sposób nieliniowy. Używa on funkcji f(x) = max(0, x) na wszystkich danych wejściowych, dzięki czemu sieć uczy się generalizować na danych.

[1] https://towardsdatascience.com/deep-dive-into-convolutional-networks-48db75969fdf

Ostatnia warstwa połączeń każdy z każdym „spłaszcza” dane, to jest tworzy z nich jednowymiarowy wektor, którego każdy element połączony jest z każdym elementem wejściowym, nadaje im wagi i zajmuje się klasyfikacją zbioru cech które zostały wyłuskane przez poprzednie warstwy. Ostatnia warstwa z użyciem funkcji „softmax” zwraca rozkład prawdopodobieństwa przynależności danych wejściowych dla każdej klasy.

Dodatkowo stosuje się jeszcze warstwy:

a) „dropout” - ma ona na celu dezaktywować część neuronów przy przetwarzaniu przez daną warstwę co może zapobiegać “przeuczaniu” sieci.

b) „batch-normalization” – pomaga przyspieszyć i ustabilizować naukę sieci, normalizuje ona warstwę wejściową nie tylko na początku działania sieci, ale również wewnątrz.

3. Rozwiązanie

3.1. Opis problemu.

Postanowiliśmy stworzyć model sieci konwolucyjnej który byłby w stanie rozpoznawać kilka typów emocji na podstawie mimiki twarzy. Wybrane emocje to:
- złość;
- obrzydzenie;
- strach;
- szczęście;
- smutek;
- zaskoczenie;
- brak emocji/stan neutralny;

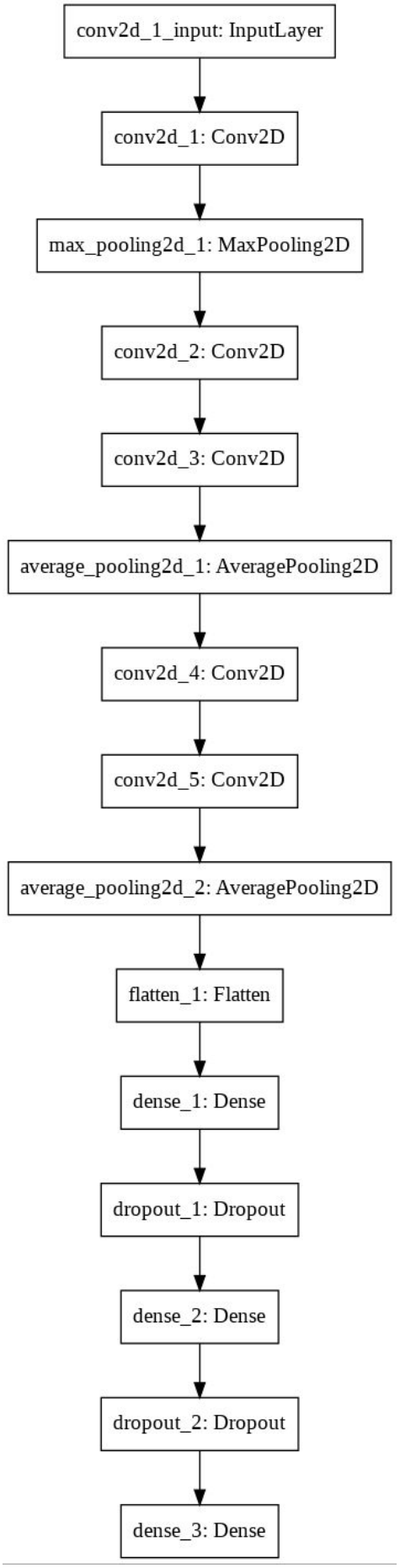
3.2. Nasze dane.

W tym celu uczenia sieci posłużyliśmy się dwoma zbiorami:
- "face-recognition-dataset.zip"[2]
- "fer2013.csv"[3]

Oba zbiory składają się ze zdjęć o wymiarach 48x48x1, i posiadają łącznie około 80 tysięcy elementów które należą do danych typów emocji. Zaletą tych zbiorów jest fakt, iż posiadają obszerną ilość danych, którą można stosunkowo szybko przetworzyć, jako że zdjęcia nie mają wysokiej rozdzielczości. Zdjęcia o lepszej rozdzielczości pozwoliłyby sieci neuronowej nauczyć się więcej szczegółów, jednakże dane o takim rozmiarze pozwolą nauczyć się cech najważniejszych. Fakt że zdjęcia są czarno-białe również pozbawia sieć pewnych cech, których mogłaby się nauczyć, jednakże w przypadku gdy baza danych posiadałaby zdjęcia jedynie określonej etniczności, mogłoby się okazać, że sieć nie nauczyłaby się dobrze generalizować.

[2] https://www.kaggle.com/jonathanoheix/face-expression-recognition-dataset
[3] https://www.kaggle.com/deadskull7/fer2013

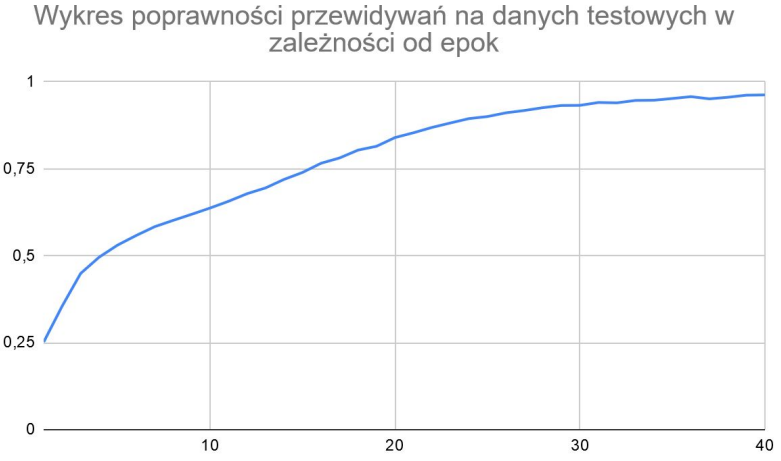
3.2. Nasz model sieci konwolucyjnej.



Nasz model sieci dobraliśmy metodą prób i błędów, zaczynając od przykładowych modeli[4] i rozbudowując je, próbując znaleźć taki, który dla naszych danych zacząłby wykazywać tendencję do nauki generalizacji na danych, a nie zwykłego zapamiętywania. Proces ten można było kontrolować poprzez obserwację wyników w trakcie uczenia. Jeśli sieć coraz trafniej klasyfikowała zdjęcia ze zbioru uczącego, a przeciwnie ze zbioru testowego, był to znak, że model uczy się zapamiętywać dane. Dzieje się tak na przykład, gdy model sieci ma za dużo parametrów, lub sieć jest zbyt trywialna dla danego problemu i nie potrafi wyizolować cech charakterystycznych dla danych klas.

3.4. Proces uczenia.

Do procesu uczenia sieci użyliśmy zbioru danych “fer2013”, a do opracowywania wyników zbioru “face-recognition-dataset”.



4 https://towardsdatascience.com/the-4-convolutional-neural-network-models-that-can-classify-your-fashion-images-9fe7f3e5399d

4. Eksperymenty

4.1. Wstęp

Eksperymenty przeprowadzono na danych pochodzących ze zbioru z w/w źródeł. Zdjęcia wykorzystane do uczenia były przygotowane w postaci pliku .csv zawierającego:
numer_emocji, macierz_pikseli, grupę.

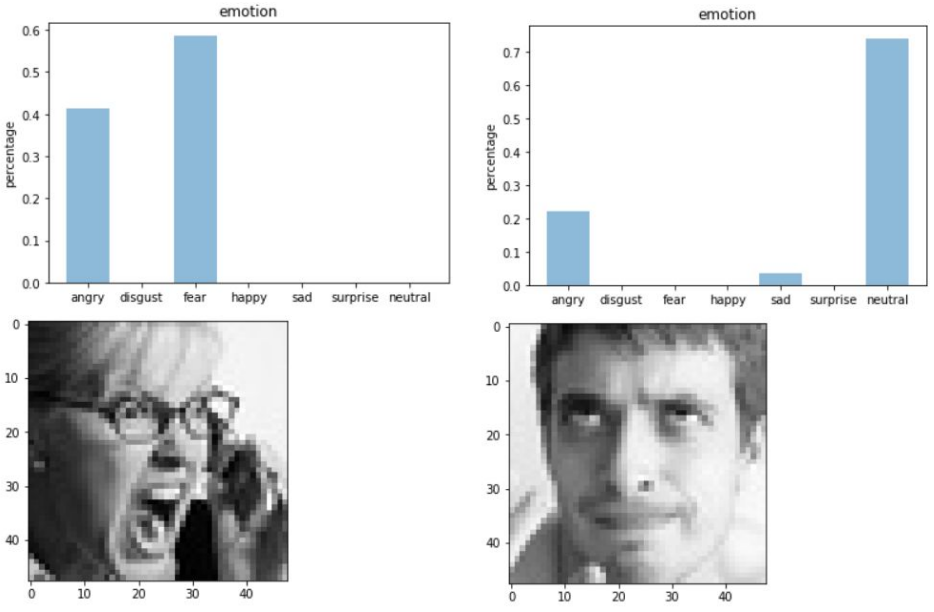
Wszystkie zdjęcia wykorzystane zarówno do trenowania sieci, jak i testowania jej były w formacie 48x48, czarno-białe. Do testów została wykorzystana specjalnie do tego przygotowana seria danych testowych. Dla każdego testu został przygotowany wykres, który określał procent z jakim wytrenowany model wskazuje na daną emocję.

4.2. Seria zdjęć testowych ze zbioru face-expression-recognition-dataset.zip

Zdjęcia przygotowane były podzielone na katalogi emocji. Dla każdej emocji w zbiorze testowym została utworzona macierz pomyłek, która wylicza ile razy dla danego zdjęcia został popełniony błąd. Jak przedstawia poniższa tablica, rozkład w danym zbiorze testowym jest stosunkowo proporcjonalny, co świadczy o stosunkowo małej granicy błędu. Ciężko jest także wskazać która emocja była mylona najczęściej z którą.

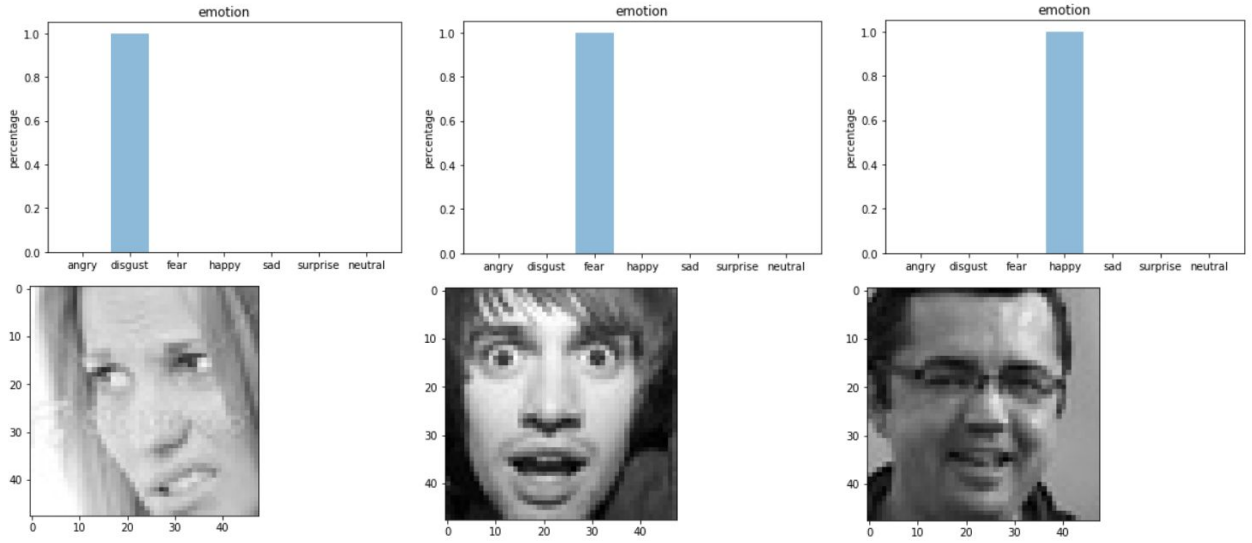
	angry	disgust	fear	happy	sad	surprise	neutral
angry	843	2	30	19	31	7	28
disgust	5	96	5	0	3	0	2
fear	27	0	883	18	35	26	29
happy	14	1	12	1750	12	12	24
sad	31	2	37	27	1003	7	32
surprise	3	0	12	11	9	754	8
neutral	22	0	16	33	36	10	1099

Tablica 1: Confusion matrix.



Rys. 5: Błędne wskazania dla emocji wskazującej na gniew.

Każda z emocji ma swoje charakterystyczne grymasy twarzy, po której można odróżnić ją na tle innych. Jak to mówi polskie porzekadło “strach ma wielkie oczy”, tak i w rzeczywistości przedstawiona jest większość osób przestraszonych: z szeroko otwartymi oczami oraz otwartymi ustami. Z kolei radość, będącą najczęściej spotykanym grymasem, wyróżnia się przynajmniej lekko podniesionymi kącikami ust oraz podniesionymi brwiami na kształt podkówki.
Na podstawie takich charakterystycznych cech twarzy jesteśmy w stanie odróżnić poszczególne emocje od siebie, na tej zasadzie się także uczy nasz model. Czy dostrzegalne są takie cechy na zdjęciach poniżej?

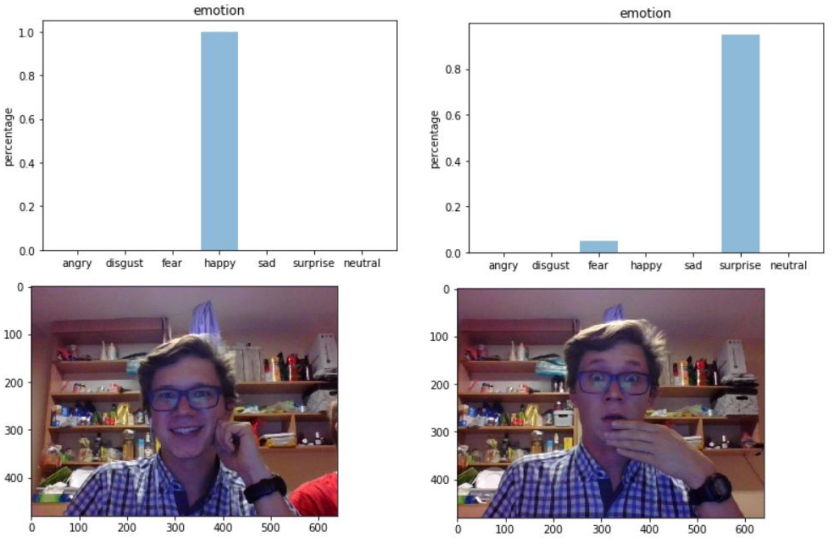


Rys. 6: Przykładowe poprawne wskazania dla różnych emocji.

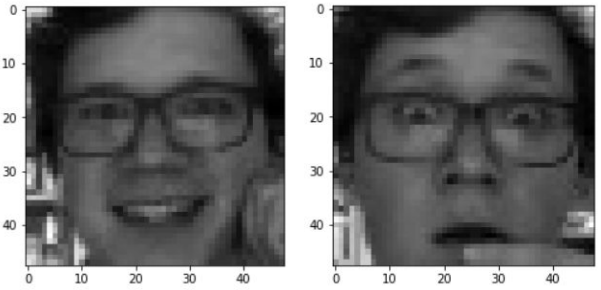
4.3. Seria zdjęć przygotowana własnoręcznie

Zdjęcia wykonane własnoręcznie mają wadę, którą są duże spadki na jakości w stosunku do oryginalnego obrazka. Zdjęcie konwertujemy do formatu 48x48 w odcieniach szarości, a następnie podajemy je jako parametr dla funkcji określającej procent emocji.

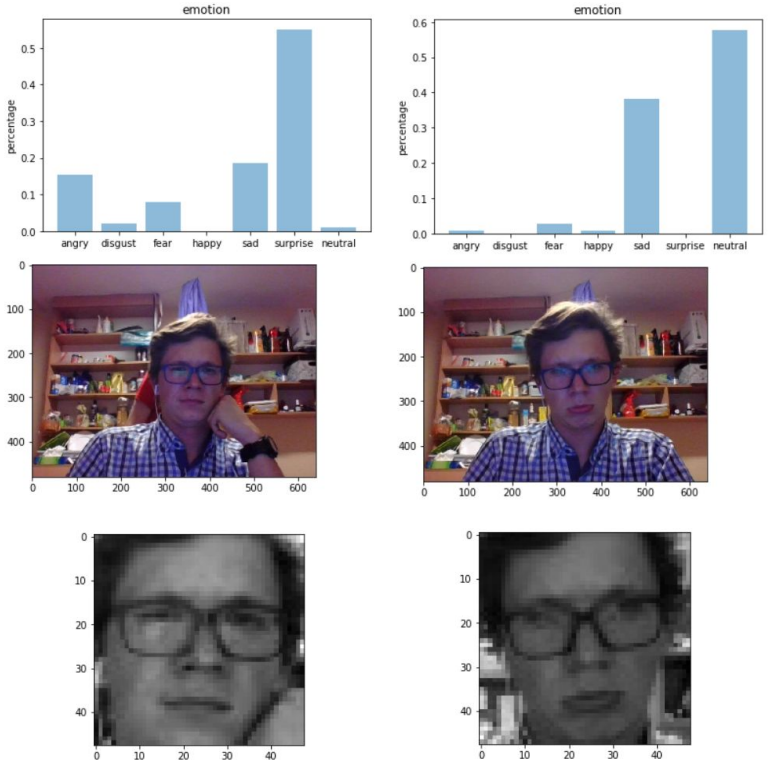
Zdjęcie konwertowane jest na zasadzie wykrywania twarzy z pomocą biblioteki OpenCV i wytrenowanego już klasyfikatora XML (haaracascade_frontalface_default.xml), który jest dostępny w repozytorium OpenCv. Oznaczamy twarz kwadratem i następnie według tego tniemy, a wynikowi zmieniamy rozmiar do odpowiadającego 48x48.



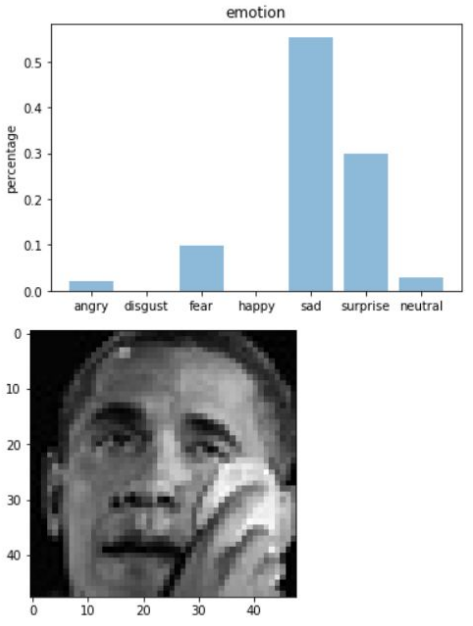
Rys. 7: Jak widać model odgaduje oczywiste emocje i grymasy bez większego problemu.



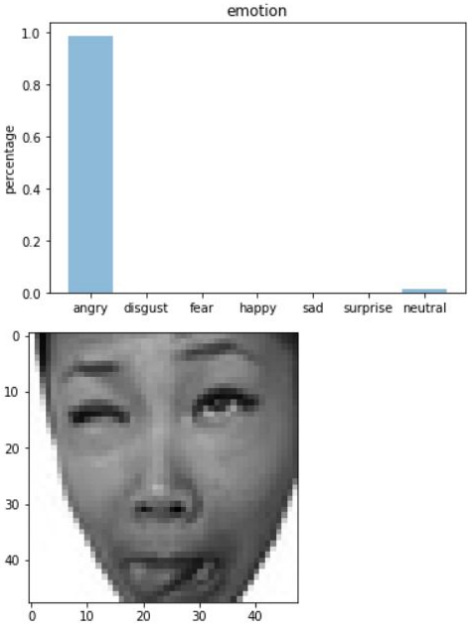
Rys. 8: Na tych obrazkach dostrzegalne jest działanie algorytmu konwertowania oraz można zobaczyć jaki obraz przetwarza sieć podejmując werdykt.



Rys. 9: Oczywiście wciąż można spotkać zdjęcia, które ciężko jest zweryfikować nawet ludziom. Sieć stara się jak może. Dużo też zależy od modela, jak dobrze potrafi okazywać emocje.



Rys. 10: Don't cry Mr Obama. :(



Rys. 11: Obrzydzenie, ale czy na pewno? 🤔

5. Literatura:

[1]
https://medium.com/@gopalkalpande/biological-inspiration-of-convolutional-neural-network-cnn-9419668898ac

[2]
chrome-extension://oemmndcbldboiebfnladdacbdfmadadm/http://home.agh.edu.pl/~horzyk/lectures/ai/SztucznaInteligencja-UczenieG%C5%82%C4%99bokichSieciNeuronowych.pdf

Klasyfikator wykrywania twarzy:
https://github.com/opencv/opencv/tree/master/data/haarcascades

Zbiory danych:
https://www.kaggle.com/deadskull7/fer2013
https://www.kaggle.com/jonathanoheix/face-expression-recognition-dataset