2-11-2020

Bryan campagne

Opdracht van AAI Docent marius

Is er een connectie tussen kaften van boeken en hun recensie

# Input neuraal netwerk

Er is getest met verschillende input als afbeelding. De variabele hierin zijn de grote van de afbeelding, kleur(RGB, HSV of grijs).

## Afbeeldingsgrote

Hoe groter de afbeelding is hoe meer data het netwerkt nodig heeft of specifieke features te vinden in een afbeelding maar de afbeelding moet ook niet te klein worden dat het neuraal netwerk geen features meer kan vinden. Na verschillende grote afbeeldingen te testen van 75X75 tot 25X25 was de beste afbeeldingsgrote 40X40.

## Afbeeldingskleur

Net als de grote van een afbeelding kan de afbeelding ook te veel informatie bevatten op kleur basis daarom heb ik uitgebreid getest met het kleurenspectrum van de afbeelding. Als eerste heb ik de afbeeldingen zwart wit gemaakt. Daarin merkte dat het netwerk erg laag scoorde en kwam niet hoger dan 52 % accuraatheid. Dit betekent dat hij net iets beter is dan gewoon maar gokken. Ten Opzichte van RGB wat een gemiddeld een accuraatheid had van 70 % had maar was niet consistent de accuraatheid in het trainen was erg laag maar de verificatie test was dan wel hoog. Na een idee van mijn docent Marius heb Ik HSV (hue, saturatie en brightness) en HS(hue en saturatie) geprobeerd. HSV was al consistenter dit zag ik uit de train accuraatheid die ook na elke epoch aan het groeien was. HS was nog accurater en daarom heb ik voor Hue en saturatie als afbeeldingsspectrum gekozen.

## Weging van de data

Omdat de date erg ongebalanceerd is kan je al makkelijk een accuraatheid van 95% bereiken. Dit komt doordat in de lijst met boeken die ik heb gekregen van de leraar meer goede (95%) recensies hebben dan slechte(5%). Dus als voor elk boek goed gokt heb je een accuraatheid van 95%. Daarom heb ik een weging toegepast waarin slecht boek 0.95 mee telt en een goed boek maar voor 0.05 meetelt.

# Opbouw neuraal netwerk

Ik ben begonnen met alleen maar dense layers maar hier kreeg ik geen goede uitkomt uit. Door een tip van de leraar ben in naar CONV2D layers gaan kijken. Ik heb deze layers getest met de mnist dataset om er een gevoel voor te krijgen (wat ook een tip van de leraar was).

Op basis van het testen met CONV2D heb ik voor mijn eigen neuraal netwerk de eerste laag een CONV2D gekozen met een masker van 3,3 met 128 neuronen en voor de tweede en derde laag ook CONV2D 3,3 ,masker maar dit met 64 neuronen. Na deze lagen moet er een Flatten laag komen als je na CONV2D een dense laag wilt gebruiken.

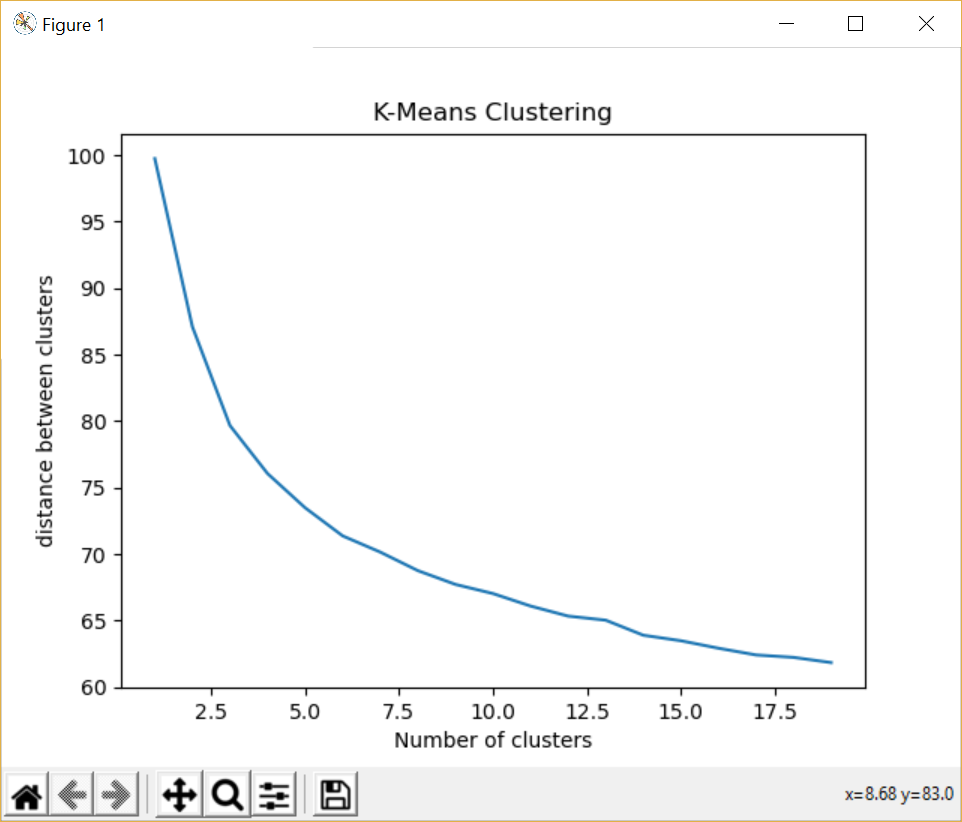
De Middellagen heb ik veel getest. Dit omdat ik met de output van de een na laatste laag een KMeans wilde doen. Hiervoor wil je niet te veel output neuronen hebben. Voor de vierde laag heb ik 64 gekozen omdat ik slechte resultaten kreeg als ik deze verhoogde of verlaagde. Ik kon deze meteen koppelen aan de output neuron als geen Kmeans ging doen. Maar Ik heb gekozen om Kmeans toe te passen. En daarom heb ik als vijfde laag een 16 neuron dense laag toegevoegd.

Omdat ik een binaire output heb(slecht of goed boek) was het eerst een idee om te eindigen met enkele dense neuron. Heb dit getest met een sigmoïd. Dit bleek niet zo goede test resulaten te geven. Daarom ben ik overgegaan naar een softmax. Maar een softmax is eigenlijk bedoelt voor een classificatie. Om van mijn netwerk ook een classificatie netwerk te bouwen moest ik met 2 dense neuronen eindigen. Dit heb ik gedaan. Hierdoor is de nauwkeurigheid van 30% naar 65% gegaan.

Door de verschillende factoren te testen en te combineren ben ik uiteindelijk op een vrij consequent algoritme gemaakt die tussen de 58% en 75%.

# Kmeans

Ik heb ook Kmeans op de opdracht toegevoegd om te kijken hoeveel clusters er zijn er zijn. De input van de Kmeans is de output van de vijfde laag van het neuraal netwerk. Door Kmeans te draaien met een max van 20 clusters komt er de volgende figuur eruit.

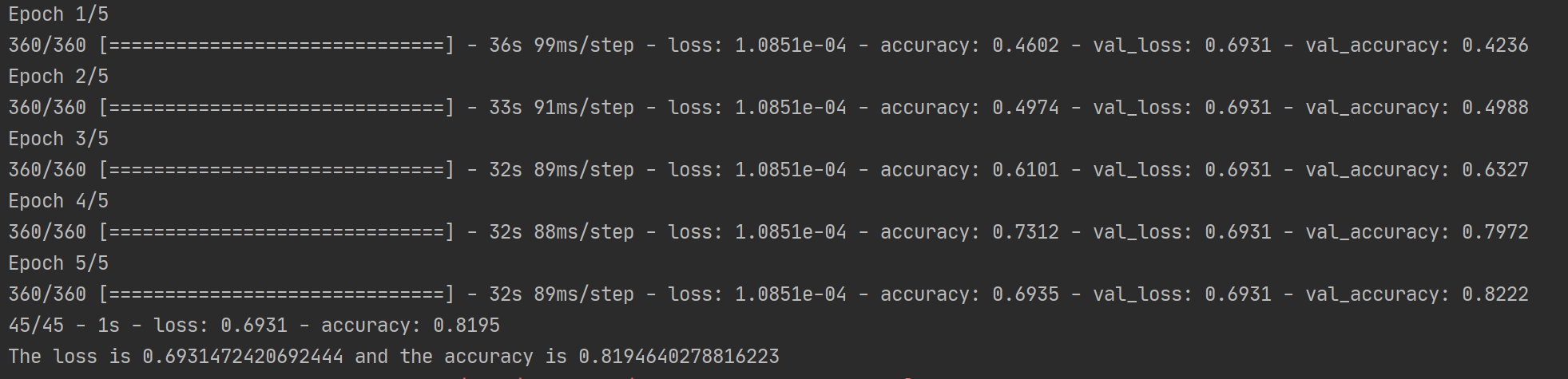


Uit deze figuur haal ik dat er 5 clusters zijn. Die bepaal ik door de e-blow techniek.

Door de functie givelabel kan ik het label per cluster opvragen. Deze geeft bijna altijd [1, 1, 1, 1, 1] terug dit betekent dat de meeste items per cluster een 1 bevatten. Dit klopt omdat 95% van de data goede boeken zijn deze zijn vermeld als 1 in de data.

# Conclusie

Er is een klein verband tussen de kaften van de boeken en de ratings ik kom zelf op een gemiddelde van 65% met een min van 52 en max van 82%. Het dit komt omdat de meeste reviews die geschreven worden gaan over de inhoud van het boek. \



Voor de rest heeft deze opdracht ontzettend veel moeite gekost. Ik heb nog nooit zo veel tijd aan 1 opdracht uitbesteed. Maar omdat ik er al zo veel moeite en tijd had in geïnvesteerd wilde ik het ook afmaken.

Het was al een ontzettend groot gedoe om de covers te downloaden. Ik ben de eerste 2 weken mee bezig geweest om de data in een juist format te downloaden en fouten weg te werken in de files met de informatie. Niet alle download links werkte helaas. Wanneer een download niet werkte moest ook het gehele boek uit de lijst verwijderd worden om de data correct te houden. Dit was mijn eerste oplossing maar dat werkte niet goed. De tweede oplossing was dan een afbeelding met de waarde null aan te maken. Wanneer de afbeeldingen worden geïmporteerd om er een np arrray van te maken wordt er gekeken of een afbeelding een null waarde heeft wanneer dit is wordt deze afbeelding overgeslagen en de info ook.

Omdat ik een Nvidia grafische kaart heb wilde ik deze ook graag gebruiken tijdens deze opdracht. Het heeft me veel moeite gekost om dit werkend te krijgen. Toen ik het eindelijk werkend had gekregen. Wilde ik CONV2D layers gaan gebruiken en die werden niet ondersteund voor de versie van cuda die ik had geïnstalleerd en het is mij uiteindelijk niet gelukt om het werkend te krijgen.