Deep Learning
Kunskapskontroll 2
Ronja Larsson
DS22 Stockholm
2023-10-05

Introduktion

Bakgrund

Deep learning har revolutionerat sättet vi löser komplexa problem på genom att använda neurala nätverk med flera lager av neuroner för att automatisera inlärning från data. Genom att extrahera och förstå komplexa mönster direkt från rådata möjliggör deep learning för oss att lösa utmanande problem. I detta projekt fokuserar vi på att träna CNN-modeller för att klassificera ansiktsuttryck och kön i realtid från kamerabilder. I projektet används två dataset för att träna två olika typer av CNN-modeller, en multiklassmodell för att klassificera ansiktsuttryck och en binär klassificeringsmodell för att klassificera kön.

Syfte och frågeställning

Syftet med detta projekt är att skapa ett computer vision program som klassificerar en persons ansiktsuttryck och kön direkt på kamera i realtid. Då det kan finnas många olika lager i en modell och CNN-modeller kan vara hur komplex som helst, så vill vi besvara följande frågeställningar i detta projekt:

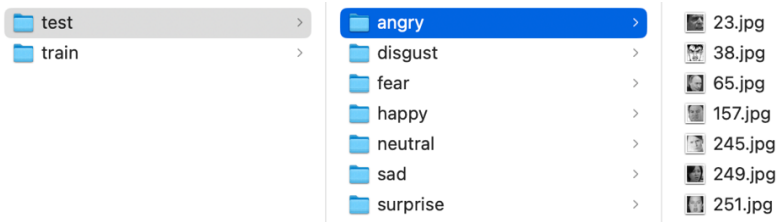
- Är modellens prediktionsförmåga bättre om modellen är mer komplex?
- Är modellens prediktionsförmåga bättre om modellen tränas i flera epoker?

Frågeställningarna undersöker prediktionsförmågan hos CNN-modeller. För multiklassklassificeringsmodellen, som syftar till att klassificera 7 olika ansiktsuttryck, utvärderas prestanda genom mått som validation loss, validation accuracy och F1 score. För binär klassificeringsmodellen, som identifierar kön, används måtten validation loss och validation accuracy.

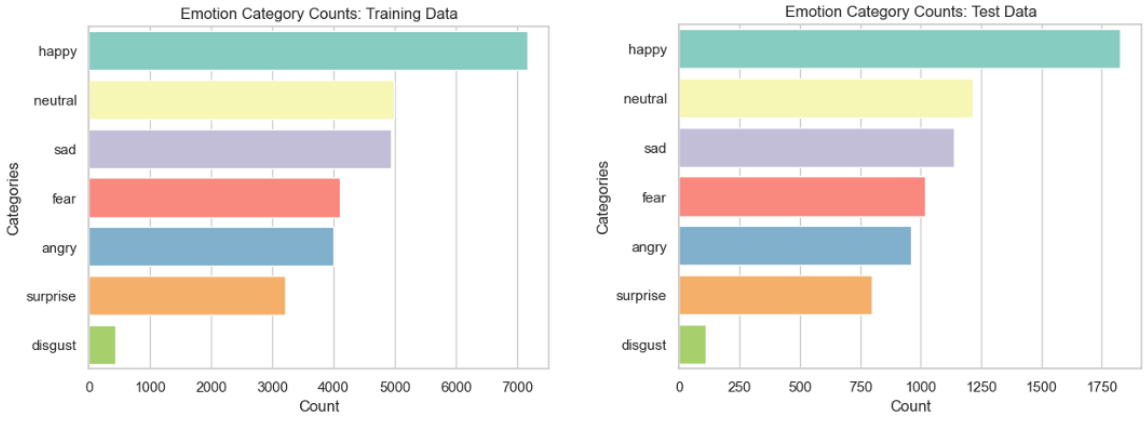
EDA

Dataset för multiklassklassificeringsmodellen

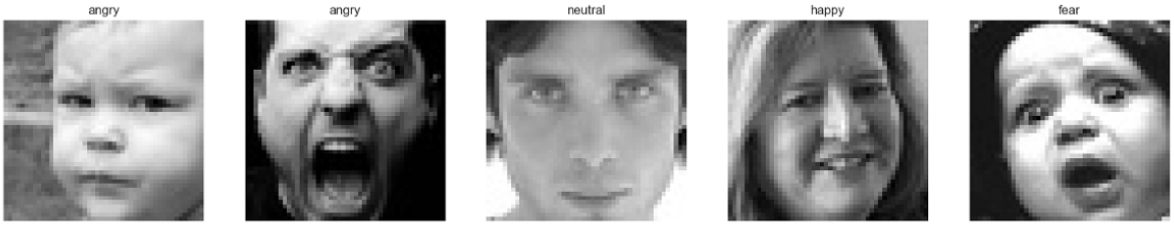
Dataset för ansiktsuttrycksmodellen är delad i två delar: en mapp med testdata och en mapp med träningsdata. Under respektive mapp finns det ytterligare 7 mappar som innehåller de ansiktsuttrycken som klassificeras. Bilderna i detta dataset är gråskalade på 48x48 pixlar.



Det finns 28821 bilder delade i 7 klass i träningsdata och 7066 bilder delade i 7 klasser i testdata. Det är värt att notera att bilderna inte är jämnt fördelade i både träningsdata och testdata – datasetet innehåller mycket flera bilder i klassen 'happy' men mycket färre i klassen 'disgust'.



Nedan är några bilder tagna ur träningsdata:



För att få en effektiv tränings- och testprocess har vi skalat om bilderna genom att dela varje pixel genom 255 och bildstorleken har ändrats till 48x48, utifall det finns bilder i olika storlekar. Att bilderna har samma storlek är viktigt när man tränar och testar CNN modeller.

Dataset för binär klassificeringsmodellen

Dataset för könsmodellen är delad i två delar: en mapp med bilder på män och en mapp med bilder på kvinnor. Bilderna i detta dataset är färgade på varierande storlekar.

Det finns 1173 bilder på män och 1134 bilder på kvinnor och nedan är några bilder tagna ur datasetet:

Bilderna i detta dataset har också skalats om genom att dela varje pixel genom 255. Bildstorleken är ändrad till en standard av 96x96. Detta är väldigt viktigt då alla bilder i detta dataset är av olika storlekar.

Modell och Metod

Modellerna är byggda med Convolutional Neural Network på grund av dess överlägsna förmåga att hantera bilddata. I en CNN-modell identifieras först ”low level features” såsom enklarare former och färger. Sedan kombineras dessa enklare egenskaper för att skapa ”high level features”, såsom delar av ögon, mun och ansikte. Beroende på vilka egenskaper CNN-modellen hittar, görs prediktioner om ifall ett objekt finns på bilden eller inte.

CNN modeller använder ”convolutional layers” i kombination med ”pooling layers”. Ett ”convolutional layer” består av flera filter där varje filter söker eller betonar vissa lokala attribut. I praktiken används flera filter för att hitta flera olika attribut och modellen lär sig själv vilka filter som ska användas, dvs vilka vikter som används. Ett ”pooling layer” fokuserar på de viktigaste delarna i en bild, till exempel ett maxvärde.

I våra modeller tillåter ”batch normalization” högre inlärningshastigheter och förbättrar generaliseringsprestanda, vilket leder till bättre och mer effektiv träning av neurala nätverk. "Callbacks" används för att utföra åtgärder vid specifika händelser eller villkor i programkoden. Inom maskininlärning används callbacks ofta för att övervaka och kontrollera träningsprocessen för neurala nätverk eller andra maskininlärningsmodeller.

Dropout är en regulariseringsteknik som ”droppar” utvalda delar av neuroner under en träningsomgång. Detta leder till att neuronerna tvingas till att ”lära sig själva” och inte ”samarbeta”. Om en modell överanpassar sig så får man höja ”dropout rate”. Early stopping är en annan regulariseringsteknik som används i våra modeller för att motverka överanpassning. Denna teknik kollar på en modells valideringsfel för att tillämpa regularisering. När valideringsfelet slutar minskar i ett visst antal epoker så kan vi stoppa träningen av modellen. Vi specificerar antalet epoker genom att ge parameter ”patience” ett värde.

Deep Learning
Kunskapskontroll 2
Ronja Larsson
DS22 Stockholm
2023-10-05

Tre multiklassklassificeringsmodeller för ansiktsuttryck skapades med olika antal av lager och callbacks. Utvärderingsmått är validation accuracy, validation loss samt confusion metrix och F1 score. Två binär klassificeringsmodeller för kön byggdes men den enda skillnaden mellan dessa två modeller är ifall early stopping tillämpas eller inte. Då det bara är 2 klasser (man eller kvinna) tittar vi värderingsmåtten validation accuracy och validation loss.

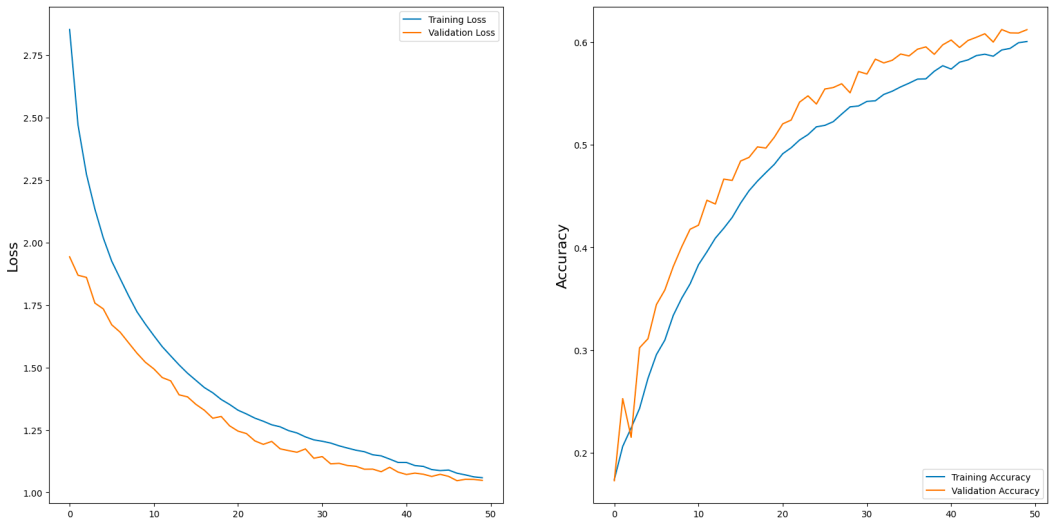
Resultat och Analys

Resultat

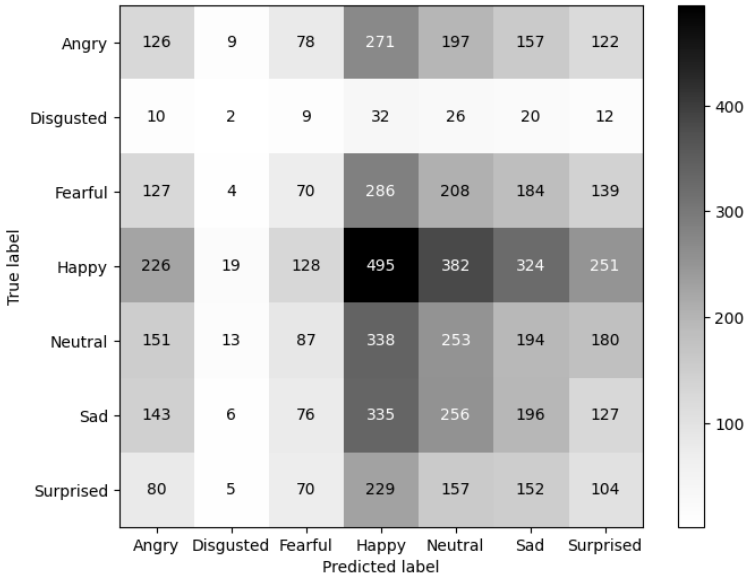
Resultaten för 3 multiklassklassificeringsmodeller

emotion_model_1

Denna modell har flest lager (4 Conv2D, 2 Dense) men endast har early stopping (med 10 patience) som callback. Den tränades på alla 50 epoker vilket betyder att early stopping aldrig triggades. Grafen nedan visar ett ganska bra resultat när man tittar på validation loss (1.0481) och validation accuracy (0.6122). Träningstiden var 103 minuter och 19 sekunder.



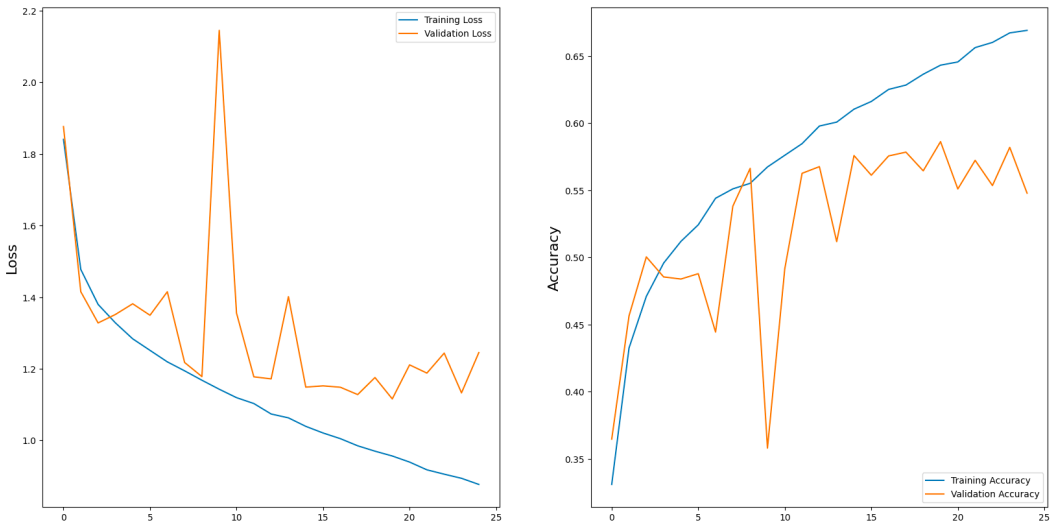
Tittar man på nedan confusion metrix så ser man att modellen inte predikterar uttrycket 'disgusted' så väl/ofta medan uttrycket 'happy' predikteras mest frekvent. F1 score här är 0.18.



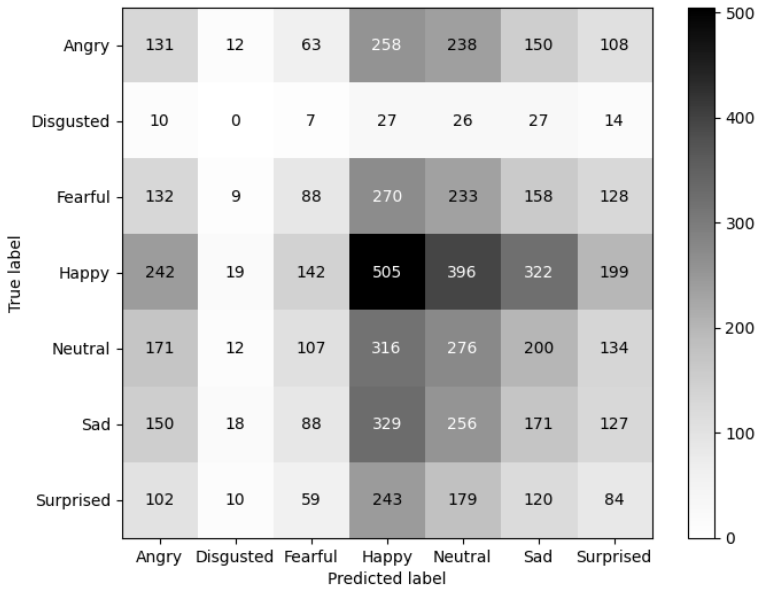
Deep Learning
Kunskapskontroll 2
Ronja Larsson
DS22 Stockholm
2023-10-05

emotion_model_2

Denna modell har lite färre lager (3 Conv2D, 1 Dense) och också endast har early stopping (med 5 patience denna gång) som callback. Den tränades på 25 utav alla 50 epoker vilket betyder att early stopping triggades efter 25 epoker. Grafen nedan visar validation loss (1.2452) och validation accuracy (0.5479). Träningstiden var 62 minuter och 12 sekunder.



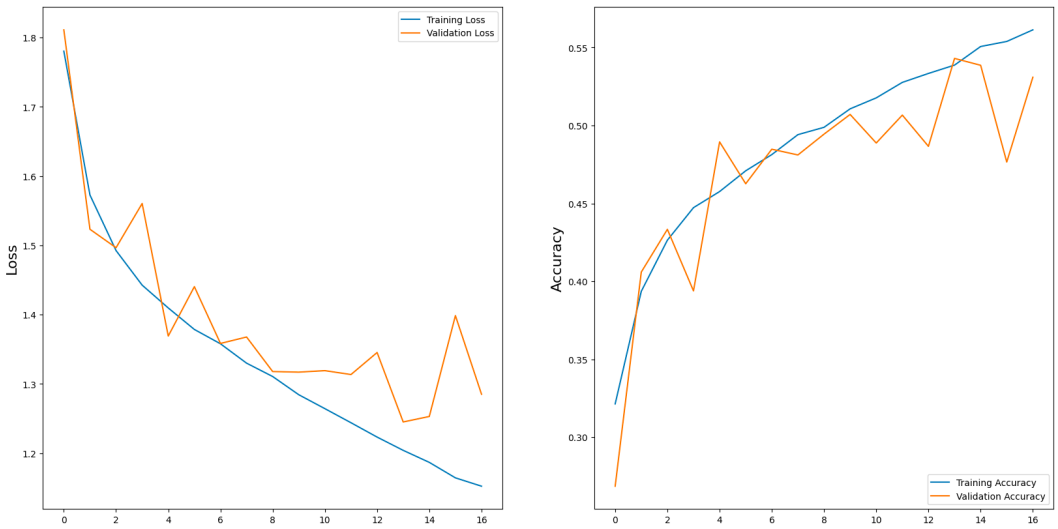
Tittar man på nedan confusion metrix så ser man att modellen inte predikterar uttrycket 'disgusted' så väl/ofta medan uttrycket 'happy' predikteras mest frekvent såsom i fallet med emotion_model_1. F1 score här är 0.18, såsom emotion_model_1.



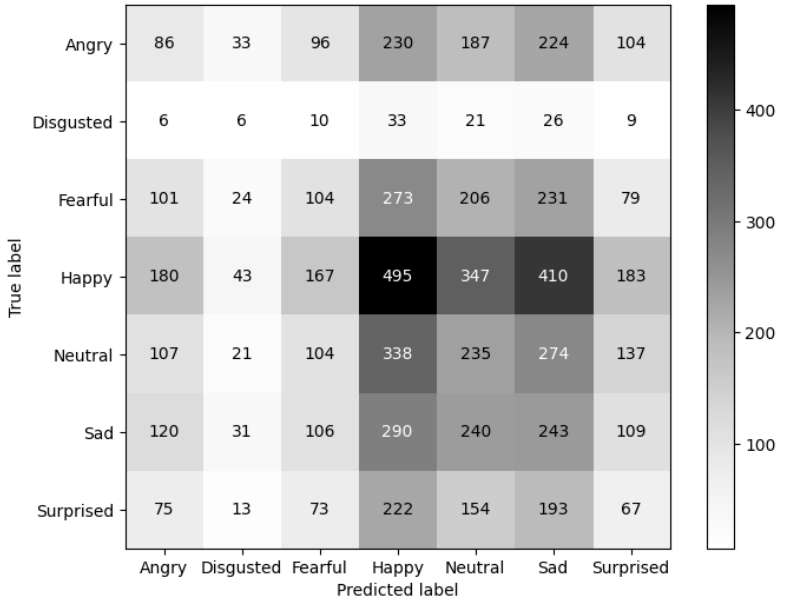
Deep Learning
Kunskapskontroll 2
Ronja Larsson
DS22 Stockholm
2023-10-05

emotion_model_3

Denna modell har ännu färre lager (2 Conv2D, 1 Dense). Det som skiljer denna modell från tidigare 2 är att flera callbacks använts: early stopping (med 3 patience), model check point samt reduced learning rate. Den tränades på 17 utav alla 50 epoker vilket betyder att early stopping triggades efter 17 epoker. Grafen nedan visar validation loss (1.2851) och validation accuracy (0.531). Träningstiden var 32 minuter.



Tittar man på nedan confusion metrix så ser man att modellen inte predikterar uttrycket 'disgusted' så väl/ofta medan uttrycket 'happy' predikteras mest frekvent såsom i fallet med emotion_model_1 och emotion_model_2. F1 score för denna modell är 0.17 vilket är något lägre än de 2 tidigare modellerna.

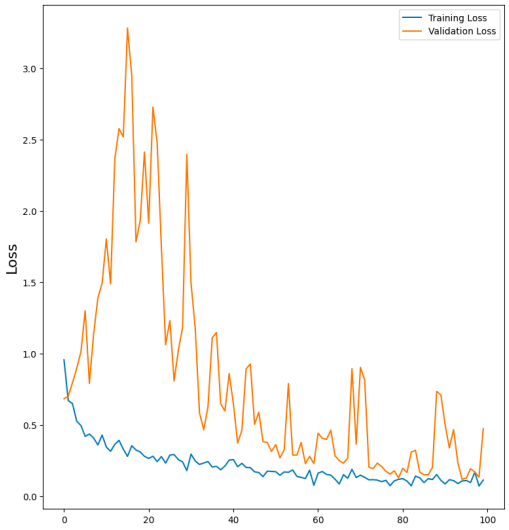


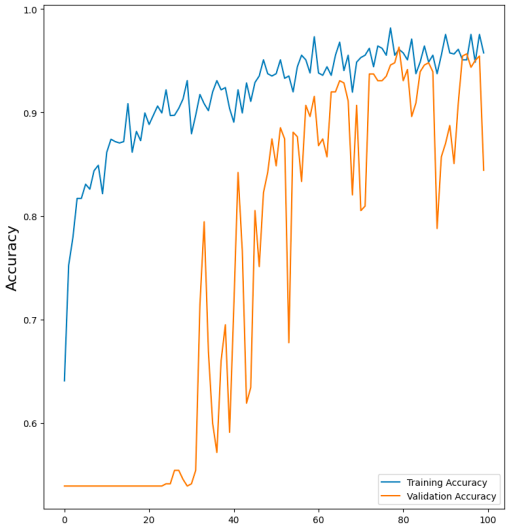
Deep Learning
Kunskapskontroll 2
Ronja Larsson
DS22 Stockholm
2023-10-05

Resultaten för 2 binär klassificeringsmodeller

gender_model_100e

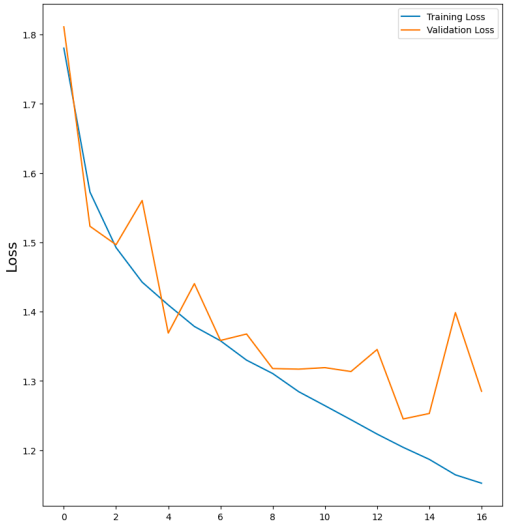
Den första modellen har 6 Conv2D lager och det enda Dense lager också är output lager. Modellen tränades på alla 100 epoker och utan någon callback. Vi kan se att resultaten från varje epok skiljer mycket i både validation loss (0.4734) och validation accuracy (0.8442). Dock ser vi att trenden går åt rätt håll för båda måtten. Träningen tog 5 minuter.

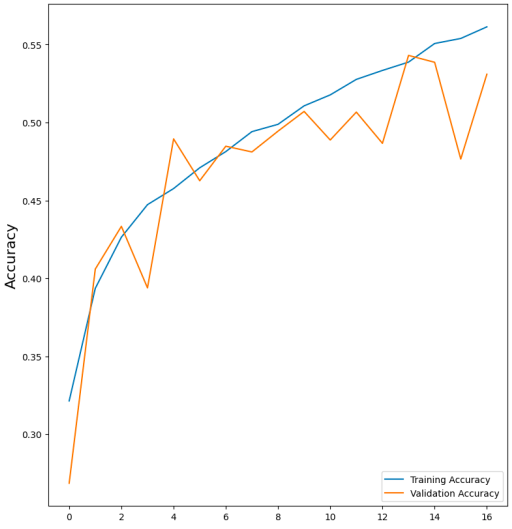




gender_model_es

Den andra modellen har exakt samma konstruktion som gender_model_100e, men callback early stopping (patience är 10) tillämpades här. Modellen tränades på 38 epoker. Validation loss (0.3355) och validation accuracy (0.8593) går också mot rätt håll här. Träningen tog 2 minuter 18 sekunder.





Analys

Dessa modeller tränades på ganska små mängder av data, så man får ta hänsyn till det när man titta på resultaten för både emotion_model och gender_model. Dessutom är datan för uttrycket 'disgust' underrepresenterad jämför med de andra 6 ansiktsuttrycken. Så det är inte så konstigt att ingen av emotion_model klassificerar just 'disgust' så väl.

Deep Learning
Kunskapskontroll 2
Ronja Larsson
DS22 Stockholm
2023-10-05

Nedan är en sammanställning av värderingsmått vi har valt att titta på. Model 1 tog mycket mer tid att träna upp vilket berodde på att den tränades på alla 50 epoker. Den har också högsta validation accuracy samt lägsta validation loss. F1 score ligger i samma nivå som de andra 2 modellerna. Även om tiden för att träna upp Model 1 är längre så anser jag fortfarande att det är en rimlig tid. Jag tog tid att diskutera med några klasskamrater och bestämde att jag inte ska försöka förbättra validation accuracy, validation loss och F1 score genom att justera hyperparametrarna. Efter diskussioner med andra anser jag att 61% accuracy är bra nog med tanken på dataset. Ett sätt att förbättra resultaten kan vara att ta in nya data för att träna upp modellerna. Om data kan vara mer jämt fördelad kan resultaten också förbättras.

EMOTION MODELS	**Model 1**	**Model 2**	**Model 3**
Training time	103 min 19 sec	62 min 12 sec	32 min
val_accuracy	0,6122	0,5479	0,531
val_loss	1,0481	1,2452	1,2851
F1	0,18	0,18	0,17
Epoch	50	25	17

Nedan är en sammanställning av värderingsmåtten för de två binär klassificeringsmodellerna. Båda modellerna tog väldigt lite tid att träna upp. Detta beror på att dataset är ganska liten, endast 2307 bilder för båda könen. Den enda skillnaden mellan nedan 2 modeller var ifall early stopping användes i träningen. Validation loss är ganska lågt i Mode es, medan validation accuracy är ungefär samma. I detta fall så verkar antalet epoker inte spelade så stor roll i modellens klassificeringsförmåga, då Mode es tränades på endast 38 epoker när early stopping triggades i gång. Patience här var 10, vilket inte var så kort.

GENDER MODELS	**Model 100e**	**Model es**
Training time	5 min	2 min 18 sec
val_accuracy	0,8442	0,8593
val_loss	0,4734	0,3355
Epoch	100	38

Slutsats

När vi tittar på resultaten på alla modeller så kan vi dra slutsatsen att modellerna presterar ganska lika. Jag har valt att ändå välja den modellen med bästa resultat på mätmåtten: Model 1 som ansiktsuttrycksmodellen och Mode es som könsmodellen som jag använder i produktion. Förbättringen kan göras med ytterligare dataset, som vi tidigare nämnt.

De frågeställningarna som vi vill besvara är följande:

- Är modellens prediktionsförmåga bättre om modellen är mer komplex?
- Är modellens prediktionsförmåga bättre om modellen tränas i flera epoker?

Efter detta lilla projekt har jag fått en känsla på hur det fungerar när man jobbar med att träna en deep learning-modell. I mitt lilla projekt så verkar det stämma att modellens prediktionsförmåga är bättre om den är mer komplex om man tittar på resultaten för Model 1 (ansiktsuttryck). Denna modell har flest lager och har en callback som inte triggades. Dock tror jag att det inte alltid är fallet – man får titta på vilken sorts problem man vill lösa. Även om det är så i mitt fall, så tror jag inte att mer komplexa modeller presterar alltid bättre. Prestandan hos en CNN-modell beror på flera faktorer, inklusive komplexiteten hos problemet, mängden tillgängliga träningsdata och den korrekta balansen mellan modellens komplexitet och datamängden.

På frågan om epoker och prediktionsförmåga, i fallet med ansiktsuttrycksmodellen så stämmer det att modellen som tränades på flest epoker presterade bäst (alla 3 modeller hade early stopping med olika patience), men i fallet med könsmodellen så stämmer det inte. Vi får dock komma ihåg att de 2 binära klassificeringsmodellerna är

identiska med undantaget av early stopping som callback. Generellt skulle jag säga att i många fall förbättras CNN-modellens prediktionsförmåga när den tränas under flera epoker. Träning av en modell över flera epoker ger modellen möjlighet att se och lära sig från hela träningsdatamängden flera gånger, vilket kan leda till bättre generalisering och förbättrad prediktionsförmåga på nya, oberoende data.

Klassificering med realtidkamera

För att starta programmet och testar mina modeller själv, öppna och kör filen med namnet 'USE.ipznb' i mappen för kunskapskontroll 2. Jag har kunnat få programmet att klassificera 6 uttryck av 7. Ansiktsuttrycket 'neutral' kunde mitt program inte hitta på mitt ansikte.

Potentiell vidareutveckling

Man kan utveckla detta projekt på flera sätt, nedan är mina förslag.

- *Predikterar emotioner med flera modaliteter*: Man kan kombinera analys av ansiktsuttryck med andra modaliteter såsom tal och text och analyserar känslor utifrån talmönster, ansiktsuttryck och skriven text. Genom att man integrerar information från flera källor kan modellens prediktionsförmåga vara mer exakt.
- *Transfer Learning*: Man kan använda förtränade modeller som ResNet eller MobileNet som extraherar features och lägger på anpassade lager på dessa klassificeringsmodeller.
- *Experimenterar med dataset*: Man applicerar olika transformationer (rotation, scaling, flipping) för att skapa nya träningsdata. Man kan också balansera dataset över olika emotioner/ansiktsuttryck. I detta projekt så är uttrycket 'disgust' underrepresenterat. För att förbättra modellens förmåga att klassificera detta uttryck så bör man öka antalet bilder i denna klass.

Deep Learning
Kunskapskontroll 2
Ronja Larsson
DS22 Stockholm
2023-10-05

Kort Redogörelse

1. *Utmaningar du haft under arbetet samt hur du hanterat dem.*

   Utmaningen låg i att inte fastnar i hur jag kunde få en bättre prediktionsförmåga, utan att hitta ett mer kreativt sätt att utveckla mitt program. Jag pratade med några klasskamrater och jämförde våra resultat. Då insåg jag att jag hade fastnat i detaljer för mycket. Så jag började tänka på vilket sätt jag kan göra mitt program roligare (utan att bli för mycket fokus på att få 90% korrekta klassificeringar).

2. *Vilket betyg du anser att du skall ha och varför.*

   Jag har alltid försökt att göra mitt bästa för att uppnå högre standard. I detta arbete gjorde jag mitt bästa med den kunskapsnivån jag har i nuläget. Jag anser att jag kan på ett grundläggande sätt att tillämpa deep learning inom bild/videohantering. Kanske hade jag lyckats med att utveckla detta arbete på ett fördjupat sätt genom att bygga en separad modell för könsklassificering. Även om mycket mer kan förbättras, så hoppas att jag har visat att jag kan självständigt genomföra arbetet och utveckla samt analysera inom optimering på ett fördjupat sätt. Jag anser att jag kan få en G+ (men hoppas på en VG-).

3. *Tips du hade ”gett till dig själv” i början av kursen nu när du slutfört den.*

   Under denna kurs hade jag en heltidssommarpraktik i Grekland där jag jobbade 6 dagar i veckan som säkerhetsdykare. Det var svårt att hänga med kursen i kombination med detta jobb och om jag kunde planera om så hade jag sett till att sluta jobba när kursen startades.