100 Spørgsmål

1 Theory

1.1 I PCA, hvorfor gør man så brug af covariance matricen?

Covairance matricen viser om forskellige dimensioner increaser eller decreaser sammen

1.2 Hvordan fungerer Batch Normalization som regularization?

Det er nok en fejl, at jeg har skrevet Batch Normalization er den regularization som jeg anvender, idet den vel ikke helt fungerer som det, men istedet bare hjælper modellen med at træne. Istedet gør jeg brug af en mini batch størrelse på 16 hvilket fungerer som regularization, idet den ikke husker hele datasættet

1.3 Hvordan kan man gøre brug af classification til pose estimation?

Man kan inddele input billedet i del billeder. Ved at klassificere i hvilket delbillede et led ligger i, kan man finde ud af positionen af ledet.

1.4 Kan du beskrive PCK med ord?

For et givet led kigger man på afstanden imellem hvor ledet rigtigt er og hvor det estimeret led er. Afstanden er så normalizeret og sammenlignet med en radius. Ligger den under radiusen tæller den som rigtigt estimeret.

1.5 Hvorfor normaliserer man i PCK?

Man normaliserer for at tage højde for forskellige størrelser af objektet i billedet. Gjorde man ikke ville det ikke være fair at sammenligne en afstand på 10 px for en person der fylder meget i billedet vs en person der fylder lidt i billedet.

1.6 Er det dog ikke forkert at du så normalizerer med en konstant i stedet for et tal der afhænger af personen?

Jo, det er det egenligt. Men, jeg valgte at følge Camilla Olsen og Newells kilde kode for at få en bedre sammenligning i min diskussion. Havde jeg gjort brug af noget andet ville sammenligningen ikke være helt så fair

1.7 Hvorfor kigger man på om afstanden ligger under en radius?

Det gør man af to årsager: (1) et led kan fylde mere end én pixel i et billede og (2) der kan være noget usikkerhed i labeling af billedet

1.8 Hvorfor gør du brug af MSE i SHG?

Jeg valgte at følge Newell og gøre brug af MSE, idet målet med denne thesis ikke i første omgang var at forbedre SHG men istedet at udforske den. Hvorfor Newell valgte at gøre brug af MSE istedet for eksempelvis RMSE vides ikke, men der er dog en række fordele/ulemper ved MSE, såsom at outliers vægtes højt

1.9 Kan du forklare momentum i SGD?

Momentum sker i linje 4, hvor der gøres brug af ligningen $\boldsymbol{v} = \alpha\boldsymbol{v} - \eta\boldsymbol{g}$. Her bestemmer α effekten af tidligere velocities, som så bruges til at peje i retning af hvor algoritmen skal bevæge sig hen imod.

1.10 Kan du forklare RMSProp?

RMSProp minder meget om SGD. Ideen er at straffe parameterer der får gradienten til at svinge meget. Den første forskel ligger i linje 5, hvor $\boldsymbol{r} = \rho\boldsymbol{r} + (1-\rho)\boldsymbol{g} \odot \boldsymbol{g}$, som vægter hvilke parametrer der får gradient til at svinge meget. Her bruges ρ til at bestemme effekten af tidligere squared gradients, samt sørger for, at $\boldsymbol{r}$ ikke eksploderer og $\boldsymbol{g} \odot \boldsymbol{g}$ bruges til at få parameter der svinger meget, til at svinge endnu mere. I den næste linje udregnes så $-\frac{\eta}{\sqrt{\delta+\boldsymbol{r}}} \odot \boldsymbol{g}$, som straffer parameterne der får gradienten til at svinge meget.

1.11 Hvad er fordelen ved at gøre brug af ReLU?

Fordelene ligger primært i (1) *universal approxmiation theorem* der siger, at et neuralt netværk med mindst ét lag med ReLU og nok knuder, kan approksimere en begrænset funktion på et lukket interval, (2) ReLU er nem/hurtig at udregne, idet den bare skal vælge imellem 0 og x, samt (3) den afledte af ReLU er nem/hurtig at udregne, idet den bare er 0 hvis $x < 0$ og 1 hvis $x > 0$. Derudover undviger den lidt *the vanishing gradient problem*, mosat Sigmoid hvis afledte ligger imellem 0 og 0.25.

1.12 Hvorfor hjælper batch normalization

Batch normalization hjælper idet det sørger for, at inputtet ikke "ping-ponger" frem og tilbage, så dataen ikke pludseligt forskydes, idet den centrerer dataen.

1.13 Hvorfor er de to træningsparametrer i batch normalization der?

De to parametrer skal læres under træning og er bare to parametrer der kan læres for at optimere modellens performance.

2 Autoencoder, Stacked Hourglass, PCA and K-Means

2.1 Hvorfor gøres der brug af skip-connections i residual modules?

Skip-connections bruges af to årsager: (1) dybbe netværk har ofte problemer med at lære hvilke værdier parametrerne skal have. Ved at gøre brug af skip-connections kan en model lære at "hoppe" lag over og derved fungere som et mindre dybt netværk. (2) Der kan ofte forekomme *the vanishing gradient problem* hvor en gradient bliver så lille, at den ikke har en effekt på netværket. Dette skyldes ofte, at der ganges en lille partial derivative med en anden lille partial derivative, hvilket giver noget endnu mindre. Ved istedet at tilføje kan man undgå dette.

2.2 Der gøres ofte brug af 1×1 convolutions. Hvorfor gøres der det?

1×1 convolutions er gode når det eneste man skal er at ændre på antallet af featuremaps som der produceres.

2.3 Hvad kommer der efter hvert hourglass?

Efter hvert hourglass er der en residual og to 1×1 convolutions

2.4 Hvorfor stacker man flere hourglasses?

Man stacker flere hourglasses, idet hvert hourglass så reestimerer det tidligere hourglass's estimat.

2.5 Hvad starter SHG med?

SHG starter med en $256 \times 7 \times 7$ convolution, et residual module og max pooling for at produce data af størrelse $256 \times 64 \times 64$. Dette gøres for at reducere hukommelses brugen

2.6 Kan du beskrive silhouette score med ord?

Silhouette score fungerer ved at man for ethvert punkt sammenligner afstanden til de andre punkter i sit eget cluster, med afstanden til de andre punkter i det næst tætteste cluster. Her er målet så, at afstanden til sit eget cluster skal være meget mindre end afstanden til det andet cluster.

3 The Dataset

3.1 Hvorfor valgte du at gøre brug af COCO-datasættet?

Der er to grunde til det: (1) det passer til vores problem, og (2) det er state-of-the-art inden for pose estimation

3.2 Hvordan tager du højde for billeders aspect ratio?

Alle billeder skal have de samme dimensioner ved input til SHG. Derfor vælger vi, at de alle skal være kvadratiske ved cropping, som så senere ændres til 256×256.

3.3 Hvilke trin foretager du når du preprocesser billederne?

1. Anvend COCOs bounding box
2. Centrer bounding box omkring personen og ret bounding box til
3. Udvid bounding box med 10%
4. Flyt bounding box op eller ned hvis der er brug for det
5. Crop billede til bounding box
6. Resize til 256×256
7. Træk den gennemsnitlige RGB fra alle billeder

3.4 Hvorfor valgte du gemme dine billeder som PNG?

Jeg valgte at gemme dem som PNG istedet for eksempelvis JPEG, idet PNG er lossless.

3.5 Hvad mener du ved, at du indsatte et 1 it et 64×64 istedet for i et 256×256 some så blev ændret til 64×64?

Hvis man indsætter 1 i et 256×256 istedet for i et 64×64 er der stor sandsynlighed, at dette 1 ikke bliver repræsenteret rigtigt i det transformeret billede idet den bliver tabt under resizing. Ved at gøre det på den anden måde sørger man for, at den ikke tabes.

4 Experiment

4.1 Hvorfor laver du ikke flere eksperimenter for at optimere SHG?

Målet var ikke at finde den optimale setting for SHG, men istedet at udforske og forstå SHG.

4.2 Hvorfor initializerer du ved hjælp af den Glorot normal fordeling?

Der er flere grunde til det: (1) Det er hvad Camilla Olsen gør. Ved at jeg også gør det, får jeg en mere fair sammenligning. (2) Vi sørger for, at vægtene ikke er lig 0, hvilket ville ødelægge backpropegation. (3) Det ødelægger symmetri, hvilket ellers ville gøre det svært for modellen at træne. (4) Det resulterer i en mere balanceret læring af de forskellige lag.

4.3 Hvorfor foretager du mange af de samme valg som Camilla og Newell?

Målet var ikke at optimere SHG i starten, men istedet at udforske SHG. Denne viden skal så senere bruges til at optimere modellen. Ved at tage de samme valg som Camilla og Newell bliver det mere tydeligt hvor stor en forbedring vores ændringer, kun baseret på vores optjent viden, har haft.

4.4 Hvorfor gør du ikke brug af early stopping?

Early stopping har en parameter *patience*, som kan være svær at bestemme inden modellen sætte stil at træne. Vi valgte istedet bare selv at følge modellens udvikling og stoppe den når det passede.

4.5 Hvorfor valgte du at gå videre med modellen med det bedste PCK accuracy istedet for validation loss?

Man kan bedst sammenligne modeller baseret på deres PCK accuracy. Vi prøvede dog også at gå videre med modellen med den bedste validation loss, men det gav meget dårligere resultater.

5 Interpretation

5.1 Når du skal finde effekten af skip-cons, hvorfor kigger du så kun på billeder som SHG er helt perfekt på?

Vi vil gerne afkræfte eller bekræft Newells påstand om, at skip-cons sørger for, at tabt information fra max-pooling bliver beholdt. Dette kan vi gøre på denne måde

5.2 Ved testing af skip-cons, hvorfor træner du så ikke en ny model uden skip-cons?

Det er helt rigtigt, at hvis man skulle have bedre resultater skulle man træne en ny SHG uden skip cons. Vores resultater giver dog udmærket resultater, samt grundet tidspress var det ikke muligt at træne en ny model.

5.3 Hvorfor anvender du kun træningsdataen når du laver dit XAI?

Jeg anvender kun træningsdataen, idet jeg vil se hvad modellen har lært.

5.4 Hvorfor anvender du PCA til at udforske latent space?

Jeg gør dette af to årsager: (1) Vi finder ud af hvilke componenter der de vigtigste, ved at kigge på deres variance. (2) Jeg har en step size som ændres alt efter hvilken komponent jeg kigger på. På den måde sørger jeg for, at jeg ikke kommer til at gå alt for langt

5.5 Hvordan udforsker du latent spacet?

1. Find principal componenterne
2. Gennemsnits koordinatet af principal componenterne er fundet
3. Det tætteste rigtige datapunkt er gemt
4. Vi udforsker en givet principal komponent ved at gå med en stepsize langs principal komponenten
5. Det tætteste rigtige datapunkt er gemt

6 Improving the model

6.1 Hvorfor tester du ikke forskellige setups for at optimere forbedringen af SHG?

Målet var ikke at optimere SHG så meget som muligt, men istedet at anvende den optjente viden til at optimere SHG, hvilket skete med det setup jeg valgte

6.2 Hvad mener du med, at ReLU resulterer i, at outputtet ligger i samme range som SHG?

Jeg mener, at SHG kun gør brug af ReLU som sin activationfunction. Havde jeg i AE valgt at gøre brug af eksempelvis sigmoid istedet, ville dette stadig fungere, men den anden fase hvor jeg træner SHG med AE, ville måske tage lidt længere tid, idet den skal lære at konvertere imellem forskellige ranges.

6.3 Hvorfor gør du brug af den normal distribution som du gør, når du tilføjer støj til dataen for AE?

Det kommer af

$$\mathcal{N}\left(0, x^2 e - 2\right) = 0.1 \cdot x \cdot N \sim \mathcal{N}\left(0, 1\right)$$

Så vi lader variancen afhænge af inputtet. Ved nærmere eftertanke er dette måske ikke helt rigtigt, idet x^2 måske er lidt i overkanten.

6.4 Hvorfor reducerer du ikke bare bottleneck istedet for at anvende autoencoder?

Det kunne jeg sagtens have gjort. Den største udfordring med dette er dog en tidsbegrænsning. Havde jeg dog gjort dette skulle jeg også ændre på decoderen, idet den forventer et bestemt input størrelse. Dertil skulle jeg altså også ændre på encoder (pga. skip cons) og vi er snart ude i en helt omskrivning af SHG, hvilket potentielt ville forværre dens performance.

6.5 Hvordan ved du, at SHG ikke bare bliver forbedret fordi den trænes længere og ikke fordi der gøres brug af en autoencoder?

Jeg ved dette, idet ved SHG plateued accuracies, samt validation loss steg, hvilket ikke var tilfældet ved brugen af autoencoderen

6.6 Hvorfor sammenligner du ikke din model med Newells performance på albue og håndled?

Newells performance på albue og håndled er for en model med 8 hourglass, istedet kun for det ene som jeg anvender. Denne sammenligning ville simpelthen være alt for unfair.

7 Discussion

7.1 Er den tredje cluster ved clustering på alle skeleter ikke bare seperationen som du beskriver ved clustering på fulde skeleter?

Det kunne godt være, men jeg tvivler på det. Ved clustering på fulde skeleter ser vi, at der gøres forskel på skeleter i bevægelse og skeleter der er stilstående. Det samme ses ikke i de to clusters i Cluster 0.

8 Andet

Hvilke andre teorier kunne man bruge?