Master Thesis Questions

André Oskar Andersen

Teoretiske spørgsmål

1. Hvorfor bruger man heatmaps?

   Heatmaps bruges til at modellere usikkerheden der er ved annoteringen

2. Hvorfor bruger PCK en threshold baseret på torso height?

   Det gør man for at inkorporere størrelsen af subjektet. Havde man fikset thresholden, havde predictionen været meget nemmere når personen er lille i billedet.

3. Hvorfor kigger man på om afstanden ligger under en radius?

   Ligesom ved årsagen til brugen af heatmaps, så modellerer thresholden usikkerheden der er ved annoteringen.

4. Hvorfor gør du brug af MSE?

   DeciWatch gør brug af en modificeret udgave af MSE. Unipose-LSTM gør brug af MSE. For at holde det konstant, valgte jeg bare at bruge MSE for alle modellerne.

5. Hvad er fordelen ved at bruge ReLU?

   Vigtigst af alt, lader det modellerne modellere ikke-lineære funktioner. Yderligere, så er der nogle andre fordelle, såsom (1) ReLU er hurtigt at udregne, (2) den afledte er nemt at udregne og (3) modsat andre activation functions såsom Sigmoid, har den en mindre sandsynlighed for at introducere vanishing gradientsa

6. Hvorfor hjælper layer normalization

   Ved at normalizere lagende af dataen undgår vi at inputtet ikke "ping-ponger" frem og tilbage, så dataen ikke pludseligt forskydes, idet den centrerer dataen. Dette burde sørge for, at færre epochs skal bruges.

   - Batch normalization er svært at udregne for sekvenser med skiftende længder, små batches giver en dårlig representation (mean og std) for hele datasættet og batch normalization er dårlig til parallelisering.
   - Input values in all neurons in the same layer are normalized for each data sample
   - Gør træningen mere "stabil", ved at sørge for, at activation i forward og gradienterne i backpropegation ikke bliver for store
   - Input til et neuron er normalizeret (standardizeret) ved at anvende mean og std af neuronerne for dette lag

7. Hvorfor gøres der brug af 1x1 convolutions

   1x1 convolutions bruges som et fully-connected layer på tværs af filtrerne. Det bruges ofte til at downsample antallet af filtre.

8. Hvorfor initializerer med Glorot?

   Der er flere grunde til det. (1) vægtene hverken for store eller små, hvilket hjælper på vanishing/exploding gradients. (2) vi sørger for at der ikke er nogen symmetri, hvilket ville resultere i neuronerne have samme udregninger.

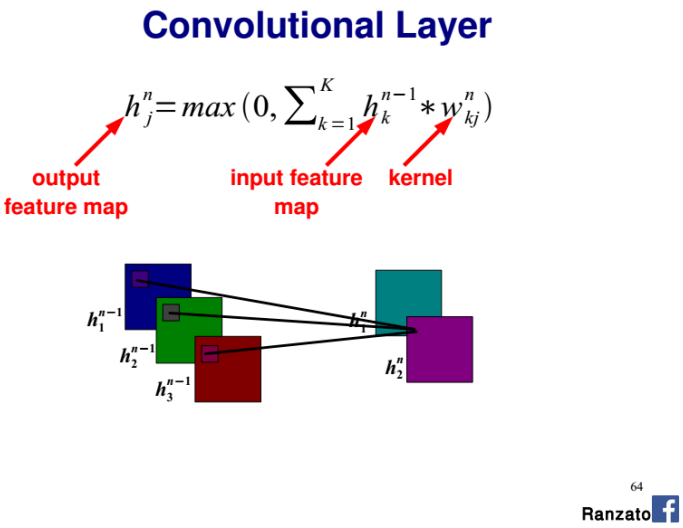
9. Kan du forklare cross-attention i en transformer?

   Forskelligt fra self-attention hvor man kun arbejder med én sætning, så mixer cross-attention to sætninger (én fra encoder og én fra decoderen).

10. Hvordan fungerer en convolution

    Hvert local receptive field i de n input feature maps foldes (convolution) med dette feature maps tilknyttede kernel resultaterne af disse n resultater lægges så sammen og placeres i output.

11. Hvorfor bruger man positional encoding i transformeren?



Transformeren tager alle ordene på én gang som input, som så bliver processed samtidigt. Hvis vi ikke havde positional encoding ville positionen af ordene ikke blive brugt, men da denne information er vigtig, vælger man at bruge positional encoding.

12. Hvad bruges self-attention til?

    Self-attention bruges til at få modellen til at få en forståelse af context af ordene i sætningen.

13. Hvad er målet med multi-head attention?

    Ved hjælp af self-attention laver multi-head attention en meget kontekst-bevidst vector af input vector

14. Hvad er målet med skip-connections (residual connections)?

    De bruges til at overkomme vanishing gradient

15. Hvorfor bruger man look ahead masks in transformeren?

    Hvis man ikke bruger dem, så har decoderen mulighed for at "se" de ord som den prøver at predicte. Look ahead mask bliver brugt ved at tilføje minus infinity til de ord som der ikke må ses

16. Hvorfor bruger man padding mask i transformeren?

    Det gør man for at sørge for, at modellen ikke lærer at "attend" til disse padding tokens.

17. Hvad er ideen bag LSTMs?

    En LSTM cell består af tre gates: (1) input-gate kontrollerer om ny information skal akkumuleres i cellen, (2) forget-gate kontrollerer om cellens hukommelse skal gensættes og (3) output-gate kontrollerer om cellens information skal sendes videre til den sidste state.

Thesis-based spørgsmål

1. Hvilke andre metoder kunne man bruge?

2. Du siger at du har introduceret noget evaluation bias i pretraining, da forskellige frames fra den samme video sequence kan optræde i forskellige subsets og den samme person derved optræder påtværs af subsets. Har du ikke samme problem i finetuning?

Både jo og nej. Jo, fordi den samme person kan optræde påtværs af subsets. Nej, fordi det er forskellige video sequences.

3. Hvorfor er 3DConv den bedst performing model?

   Det er svært at svare på, da de andre modeller ikke ser ud til at overfitte, som ellers kunne være en forklaring

4. Hvordan performer din deciwatch i forholdet til det fra artiklen?

   Vores DeciWatch performer noget dårligere end den fra artiklen, som nok højst sandsynligt kan forklares af forskellen i dataen

5. Hvilke andre metoder findes der til at undgå overfitting?

   Data augmentation og at gøre modellerne mere simple

6. Hvorfor brugte du ikke data augmentation?

   Det kunne jeg godt have gjort, men jeg skal så være lidt påpasselig, da man nemt kan "ødelægge" datasættet, ved eksempelvis at rotere videoerne for meget.

7. Hvorfor bruger DeciWatch ikke alle frames?

   Da nærliggende frames indeholder overflødigt information, argumenteres der for, at modellen ikke behøver at kigge på alle frames.

8. Hvorfor denoiser DeciWatch sit input?

   Det er svært at recover frames, især hvis de samplede poses indeholder noget noise.

9. Hvorfor er fem frames det optimale valg?

   Skaberen af Unipose har undersøgt sammenhængen imellem Unipose's performance og antallet af frames. De kom frem til, at det var udnødvendigt at bruge mere end fem frames, da performancen plateuer efter fem frames. Dette kan selvfølgelig være anderledes i vores tilfælde.

10. Hvorfor hjælper dine modeller ikke på PCK0.2?

    Vi tror det skyldes, at de predicted keypoints af pose-detectoren der klassificeres som forkert, simpelthen er så forkerte, at de ikke kan redes af vores modeller.

11. Hvorfor splitter du data'en forskelligt for pretraining og finetuning?

    Set i bakspejl, så burde jeg nok have fulgt min fremgang i finetuning igennem hele opgaven, men jeg havde ikke tænkt over det tidligere. I pretraining tænkte jeg, at datasættet var så stort og der var så meget variance, at det ikke var nødvendigt at tage de forhold som jeg gjorde i finetuning.

12. I 3DConv hvorfor valgte du at have en temporal-dybde på T-længde med K filtrer og ikke omvendt?

    Ved den nuværende løsning vil de lærte kernels bestå af vægte der lærer at vægte den temporale information. Omvendt ville de lære at vægte de forskellige led, hvilket ikke giver så meget mening

13. I 3DConv, kunne du ikke bare have ét filter som du bruger til alle keypoints?

    Nej, fordi nogle af keypointne er nok mere afhængig af de tidligere frames end andre.

14. Har du andre teorier for hvorfor fingrene er de sværeste led?

    Det kan også skyldes, at disse led er de mindste led og de derfor er svære for pose-detectoren at finde, hvilke så også går videre i vores modeller.