1 Discussion

The following section discusses our approach, including the our choice of configuration of the models, the data and the training parameters, as well as suggests some future work.

1.1 Data Configuration

- Vanishing gradient for unipose - den fungerer bedre på range [0, 255] end på range [0, 1]
- Handling af data uden for pred bbox ved CA-data
- Valg af shifting-scalar:
  - Jeg kunne have approximeret den fra mask rcnn
  - Jeg kunne have gjort den afhængig af det enkelte led
- Pretrain: jeg shifter gt med op til 6 std, men jeg udglatter højst med 3, så jeg dækker egentlig ikke den rigtige placering af gt

For the data we used three datasets. For the pretraining stage we used the BRACE datase and the Penn Action dataset, whereas for the finetuning stage we used the ClimbAlong dataset. For our choice of pretraining datasets we see a couple of suboptimal choices.

BRACE and Penn Action can be very semantic different from each other. For instance, in the BRACE dataset the person in the video sequence is very often upside down and performs some very quick movements, whereas the people in the Penn Action dataset is usually in some more usual poses an do not make movements that are as quick as the ones from BRACE. We tried to fix this by only including video sequences from the Penn Action dataset that had more poses where the person was not just standing straight up, however, even by doing so, the two datasets are still sematically different. Further, the video sequences from the two datasets are often filmed from two very different camera angles, resulting in the same keypoints being placed at different locations. If the two datasets are very different this will have an effect on the fitting of the models.

We also see some problems in the pretraining datasets individually. For the BRACE dataset, it is clearly documented, that the video sequences is filmed using 30 frames per second. For Penn Action, on the other hand, a similar information is not stated anywhere and the video sequences comes as individual frames without the duration of the video sequences noted, hence why cannot either compute the frame rate of these video sequences. This could cause some problems, as a two windows of frames, one from each dataset, could span two different durations and thus capture two different duration of context. For the BRACE we also see an issue, as the dataset has not been fully manually annotated, but instead annotated using a pose estimator and only certain incorrectly predicted poses were manually annotated, where incorrectly predicted poses were detected by a machine learning model. However, if this machine learning model for detecting incorrectly predicted poses delivers suboptimal results, then the annotations would also be suboptimal and thus contain some noise, which is very difficult for our models to predict.

1.2 Model Configuration

- Model parametrer

1.3 Training Configuration

- Skulle have brugt en længere patience til early-stopping
-
- Kør eksperimenter flere gange for at fjerne randomness i resultater.

1.4 Future Work

- Transformer uden skipping af frames
- Dataaugmentation til CA
- Træn uden PA
- Vision transformer