0.1. Implementation Details

For our experiments, we adopted the TempAgg [1] model from [2] as the baseline and made use of the RGB features provided by the authors. TempAgg model uses non-local-blocks [3] to encode the observed video snippets at different temporal scales creating recent and spanning features used for anticipation of future action, verb and nouns. In addition to the already existing branches, we introduced a separate goal prediction branch that operates on the concatenation of spanning features. For training, instead of batch size 32, we used batch size 64. It improves the baseline performance, so we report the improved results, as compared to [2], in Section ???. We left other implementation details unchanged. For further details please refer to [2].

0.2. Datasets and Evaluation

Since in our approach we rely on the predefined action-goal hierarchy, for training and evaluation we made use of two procedural activity datasets that contain hierarchical action annotations: Assembly101 [2] and COIN [4].

Assembly101 is a large-scale dataset that contains 362 recordings of 15 toy vehicle assembly and disassembly sequences shot from 12 different viewpoints. This in total amounts to 4321 videos spanning 513 hours. The videos are annotated with 1M fine-grained and 100K coarse action segments, that we use as fine-grained actions and goals respectively. Fine-grained segments span 1380 action classes composed of 90 objects and 24 verbs, while coarse actions span 202 action classes formed by 11 verbs and 69 objects. Assembly101 is divided into training, validation, and test splits using 60%, 15%, and 25% of the original video recordings respectively. At the time of writing, the test set was not available, thus in our work we report results on the validation split. Following [2], we additionally provide results on two subsets of validation examples - Tail and Unseen - that contain video segments with tail action classes and toys unseen during training time respectively.

COIN dataset consists of 11827 videos that were collected from Youtube and amount to 476 hours in total. Videos depict 180 high-level tasks and are annotated with 46354 action segments from 778 lower-level action classes. In our experiments, we regard video-level task annotations as goal actions, while segment-level action annotations as fine-grained actions. Training and testing splits contain subsets of 9030 and 2797 videos respectively sampled from the original set of recordings.
1. REFERENCES


