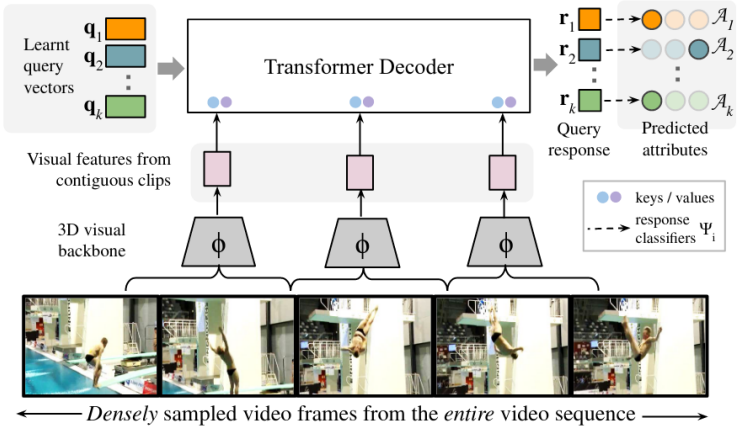
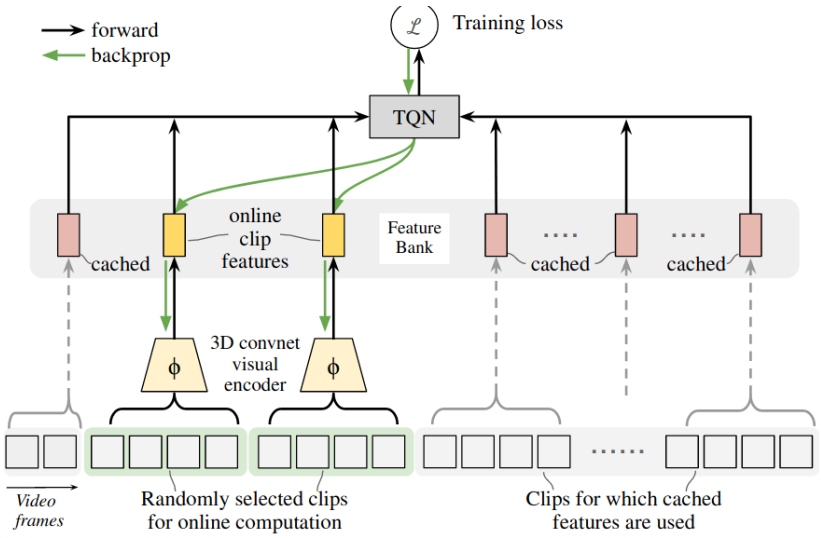
Temporal Query Networks for Fine-grained Video Understanding

Chuhan Zhang, Ankush Gupta, Andrew Zisserman

Temporal Query Networks or TQNs propose a novel module for long-range fine-grained video understanding. TQN takes as input a set of predefined queries that depend on the classes we want to detect and their hierarchies, where each query corresponds to some concept or a node in the hierarchy, and the attributes of each query to the subnodes of the elements of the concept. The queries are then updated sequentially with transformer layers taking at each time stamp the features of the current step, and the objective is to have responses in the queries that match the current actions taking place.



Given a video and a pretrained 3D backbone, the video features are extracted by sliding the backbone over the video and taking as input a given clip of N and a stride of S between to frames (for example 8 x 1, where we sample 8 successive frames as one clip). The features of the whole video are then passed to the transformer which updates the queries via cross attention where the keys and values are generated from the current input features. The transformer and the queries are then trained with a cross-entropy loss, where the responses of the attributes of each query are used as logits of the corresponding class.



Another important aspect of training TQNs is updating the backbone when training. First, the backbone is pre-trained on small snippets and then used to cache the features. Then at the second stage, the cached features are used as input to the TQN, but where some of the cached features are replaced with features produced by the backbone being fine-tuned.