There are two popular methods in the text-video field. One is the embedding-based method. It utilizes the embeddings of features extracted from the text and video separately but suffers from the low-accuracy because of the loss of correspondence details.

The other is the model-based method. It iterates all the text-video pairs to evaluate the distance without extracting embeddings. So it achieves better accuracy but suffers from low efficiency.

The authors of this paper proposed a novel EDB method named Cross-modal retrieval transformer, which solves the conflict between efficiency and accuracy.

The overview of the CRET model is shown in the slide. We can see from the picture that the proposed CRET model encodes the video and text separately. The text encoder applies the BERT as the basement to encode features into global and local features. The video encoder consists of spatial and temporal encoders which extract the spatial and temporal features separately(encode sampled multiple frames and frame-level encoding). The spatial encoder encodes all the features into global and local features. The CLS represents the global features. We can see that we feed the global features into the temporal encoder to get the temporal embeddings. On the other side, the local features are projected to the same dimension as the text embeddings.

At the top left of this picture, we estimate the parameters for the distribution of the features extracted from the video frames using the global temporal features. Then we calculate the GEES loss according to the estimated parameters and the global text features.

As for the right part of this picture, we can see that we align text and video features using the CCM module.

Next, we will discuss more details about the two important parts of this model—CCM module and GEES loss.

Actually, the CCM module utilises the self-attention mechanism. We align these features using the transformer as we can see from this equation. We first calculate the distance between the token features and the query centre and regard the distance as the weight of this feature. Afterwards, we calculate the similarity score of these aligned features from text and video modalities.

Let us move to the GEES loss. The traditional loss function NCE requires a trade-off between the accuracy and computational burden. The author improved this loss function by first making an assumption about the frame distribution. Then simplifying and approximating the NCE function which combines with the assumption. This improvement enables the model to be optimized by the SGD algorithm which enhances the efficiency of training.