1. Introduction

Given a video and a natural language query, the task of Video Moment Retrieval (VMR) involves temporally localizing moments (video segments) within the given video that are relevant to the query.

Query: a girl talking from a bus window on a running bus

2. Our Approach

No newly trained models; Zero-shot transfer from off-the-shelf models (CLIP/UniVL) to VMR.

2.1 Overall Procedure

2.2 Moment-Query Matching

3. Key Results

1. Best Zero-shot Performance

2. Better Performance than some Supervised approaches!

(That do not use labor-intensive frame-wise saliency loss or compute-intensive pretraining)

3. Great Performance for Short Segment Detection

4. But there’s room for improvement

4.1 Our query-moment matchers are still far away from the best possible ‘oracle’ query-moment matchers

4.2 We do not beat pretrained + supervised VMR approaches, despite using a pretrained V+L model (CLIP)

4. Conclusion

1. A simple shot detector reveals CLIP’s query-moment matching power and leads to performance close to some supervised approaches on VMR.

2. The shot detection method is especially good at detecting short segments, outperforming strong supervised and pretrained models.

3. Despite the simplicity of the shot detector, CLIP is not the best query-moment matcher for it; there’s quite some room for improved matchers.