RECONSTRUCTION NETWORK FOR VIDEO CAPTIONING

- RecNet
- Dataset: MSR-VTT, MSVD
- Goal: to minimize the differences between original and reproduced video features
- Forward flow (video to sentence)
- Backward flow (sentence to video)
RECONSTRUCTION NETWORK FOR VIDEO CAPTIONING

- Encoder (CNN): extracts semantic representations of video frames
- Decoder (LSTM): generates natural language for visual content description
- Reconstructor: exploits the backward flow to reproduce frame representations

Figure 2. The proposed RecNet consists of three parts: the CNN-based encoder which extracts the semantic representations of the video frames, the LSTM-based decoder which generates natural language for visual content description, and the reconstructor which exploits the backward flow from caption to visual contents to reproduce the frame representations.
C3D Feature Extraction

- Pytorch
- Keras
THANK YOU