Video description is one of the most challenging problems in vision and language understanding due to the large variability both on the video and language side