Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning

What is the problem?

A multi-modal single-stage dense event captioning model pretrained on narrated videos which are readily-available at scale. The Vid2Seq architecture augments a language model with special time tokens, allowing it to seamlessly predict event boundaries and textual descriptions in the same output sequence. Such a unified model requires large-scale training data, which is not available in current annotated datasets. We show that it is possible to leverage unlabeled narrated videos for dense video captioning, by reformulating sentence boundaries of transcribed speech as pseudo event boundaries, and using the transcribed speech sentences as pseudo event captions. The resulting Vid2Seq model pretrained on the YT-Temporal-1B dataset improves the state of the art on a variety of dense video captioning benchmarks including YouCook2, ViTT and ActivityNet Captions. Vid2Seq also generalizes well to the tasks of video paragraph captioning and video clip captioning, and to few-shot settings. Our code is publicly available at this https URL.

What has been done earlier?

1.Traditional Dense Video Captioning:

- Earlier models for dense video captioning typically involved a two-stage process: detecting events in videos first and then generating captions for those detected events.
- These models relied heavily on annotated datasets with human-labeled event boundaries and captions, which limited their scalability due to the labor-intensive nature of creating such datasets.

Use of Pretrained Language Models:

- Pretrained language models, such as BERT and GPT, have been used for generating captions by leveraging large amounts of textual data.
- However, these models were primarily trained on static images or text datasets rather than narrated videos, making them less effective for capturing the temporal dynamics of videos.

Harshit Goyal,B421024

Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning

What are the remaining challenges? What novel solution proposed by the authors to solve the problem?

Lack of Large-Scale Annotated Data:

- Dense video captioning models typically require large annotated datasets with event boundaries and corresponding captions. However, creating such datasets resource-intensive, making it difficult to scale.

Generalization Across Diverse Video Types:

- Models often struggle to generalize well across different types of videos and tasks (e.g., video paragraph captioning, clip captioning) due to the variability in video content and styles of narration.

Pretraining on Large-Scale Narrated Videos:

- To overcome the scarcity of annotated datasets, Vid2Seq is pretrained on a massive dataset of narrated videos (YT-Temporal-1B). The model leverages transcribed speech from these videos, treating sentence boundaries as pseudo event boundaries and the transcribed sentences as pseudo event captions.

e Tokens for Temporal Awareness:

- Vid2Seq introduces special time tokens into the language model, enabling it to better understand and process temporal information. This innovation helps the model accurately identify event boundaries and maintain temporal coherence in the generated captions.

Harshit Goyal,B421024