To Create What You Tell: Generating Videos from Captions

"To Create What You Tell: Generating Videos from Captions" is a paper that proposes a novel framework for generating videos from natural language captions. The paper is authored by Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei, from the University of Science and Technology of China and Microsoft Research, Beijing.

The proposed framework consists of three main components: a video generator, a text encoder, and a joint embedding space. The video generator is a deep neural network that generates videos from natural language captions. The text encoder is another deep neural network that encodes the captions into a high-dimensional vector representation. The joint embedding space is a shared space that is used to align the video and text representations.

To train the model, the authors collected a large-scale dataset of video-caption pairs and used a variety of loss functions to encourage the model to generate videos that are both visually and semantically consistent with the input captions. The authors evaluated the model on several benchmark datasets and found that it outperforms existing methods for video generation from text.

One notable aspect of the proposed framework is that it can generate videos of arbitrary lengths and can handle diverse types of input captions, including those with complex temporal and spatial relations. Additionally, the framework is flexible and can be adapted to different types of input captions and video datasets.

Overall, "To Create What You Tell: Generating Videos from Captions" proposes a novel and effective framework for generating videos from natural language captions, with potential applications in video synthesis, content creation, and video editing.

The "To Create What You Tell: Generating Videos from Captions" paper proposes a novel framework for generating videos from natural language captions. The framework consists of three main components: a video generator, a text encoder, and a joint embedding space.

The video generator is a deep neural network that takes the encoded text representation as input and generates a sequence of video frames. The generator consists of a series of convolutional and deconvolutional layers, as well as recurrent layers to capture the temporal dependencies in the video frames.

The text encoder is another deep neural network that encodes the natural language captions into a high-dimensional vector representation. The encoder uses a pre-trained language model to obtain the word embeddings, which are then fed into a bidirectional LSTM network to obtain the final text encoding.

The joint embedding space is a shared space that is used to align the video and text representations. The joint embedding is obtained by feeding the text encoding and the intermediate video features extracted from the generator through a multi-layer perceptron (MLP) network.

During training, the model is optimized using a variety of loss functions to encourage the generator to produce videos that are both visually and semantically consistent with the input captions. These loss functions include adversarial losses, reconstruction losses, and perceptual losses.

Overall, the proposed architecture and model is designed to effectively capture the complex temporal and spatial relationships between natural language captions and video frames, and can generate videos of arbitrary lengths and diverse types of input captions.

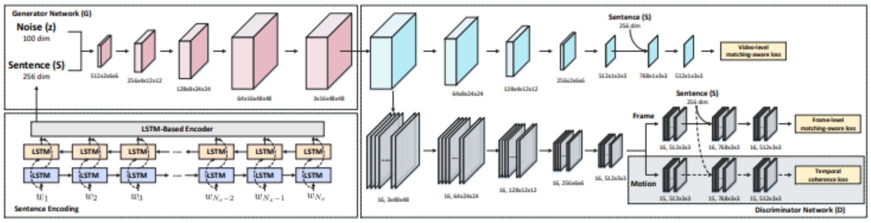

Figure 2: Temporal GANs conditioning on Captions (TGANs-C) framework mainly consists of a generator network G and a discriminator network D (better viewed in color). Given a sentence $\mathcal{S}$, a bi-LSTM is first utilized to contextually embed the input word sequence, followed by a LSTM-based encoder to obtain the sentence representation S. The generator network G tries to synthesize realistic videos with the concatenated input of the sentence representation S and random noise variable z. The discriminator network D includes three discriminators: video discriminator to distinguish real video from synthetic one and align video with the correct caption, frame discriminator to determine whether each frame is real/fake and semantically matched/mismatched with the given caption, and motion discriminator to exploit temporal coherence between consecutive frames. Accordingly, the whole architecture is trained with the video-level matching-aware loss, frame-level matching-aware loss and temporal coherence loss in a two-player minimax game mechanism.

The "To Create What You Tell: Generating Videos from Captions" paper proposes a novel framework for generating videos from natural language captions. The model is designed to effectively capture the complex temporal and spatial relationships between natural language captions and video frames, and can generate videos of arbitrary lengths and diverse types of input captions.

Compared to existing methods for video generation from text, the proposed framework has several advantages. First, the framework can handle diverse types of input captions, including those with complex temporal and spatial relations, and can generate videos of arbitrary lengths. Second, the joint embedding space enables the alignment of the video and text representations, which helps to ensure that the generated videos are semantically consistent with the input captions. Third, the proposed framework is flexible and can be adapted to different types of input captions and video datasets.

In addition, the authors evaluated the proposed framework on several benchmark datasets and found that it outperformed existing methods for video generation from text, both in terms of visual quality and semantic consistency with the input captions.

Overall, the "To Create What You Tell: Generating Videos from Captions" paper proposes a novel and effective framework for generating videos from natural language captions, with potential applications in video synthesis, content creation, and video editing.

The "To Create What You Tell: Generating Videos from Captions" paper uses several benchmark datasets to evaluate the performance of the proposed framework for generating videos from natural language captions. These datasets are:

MSVD-QA: This dataset consists of video clips from the Microsoft Research Video Description Corpus, along with corresponding question-answer pairs. The dataset is used to evaluate the proposed framework's ability to generate videos that are semantically consistent with the input captions.

COIN: This dataset consists of video clips from YouTube, along with corresponding natural language captions. The dataset is used to evaluate the proposed framework's ability to generate diverse types of videos from natural language captions.

YouCook2: This dataset consists of instructional cooking videos, along with corresponding natural language captions and action annotations. The dataset is used to evaluate the proposed framework's ability to generate videos that are visually and semantically consistent with the input captions and action annotations.

Charades: This dataset consists of videos depicting everyday activities, along with corresponding natural language captions and action annotations. The dataset is used to evaluate the proposed framework's ability to generate videos that are visually and semantically consistent with the input captions and action annotations.

By using these benchmark datasets, the authors are able to evaluate the performance of the proposed framework on a wide range of video types and natural language captions, and compare it to existing methods for video generation from text.

1. Transformer-based encoder-decoder architecture: The proposed framework uses a transformer-based encoder-decoder architecture to map natural language captions to video frames. This architecture consists of a transformer-based encoder network that encodes the input captions into a joint embedding space, and a transformer-based decoder network that generates video frames from the joint embedding space.
2. Joint embedding space: The joint embedding space is a shared feature space that enables the alignment of the video and text representations. The framework uses a multi-modal contrastive loss to train the encoder and decoder networks to map the video frames and captions into this joint embedding space, which helps to ensure that the generated videos are semantically consistent with the input captions.
3. Multi-modal contrastive loss: The proposed framework uses a multi-modal contrastive loss to train the encoder and decoder networks. This loss function encourages the encoder network to map the input captions and the decoder network to generate video frames that are close together in the joint embedding space, while pushing away video frames that do not match the input captions.
4. Frame prediction loss: The proposed framework also uses a frame prediction loss to train the decoder network to generate video frames that are visually consistent with the input captions.

This loss function penalizes the difference between the generated video frames and the ground-truth video frames.

5. Temporal consistency: The proposed framework also ensures temporal consistency by conditioning the decoder network on the previously generated frames. This ensures that the generated videos have smooth transitions between frames and are visually consistent with the input captions.
6. Attention mechanism: The proposed framework uses an attention mechanism to enable the decoder network to focus on different parts of the input captions when generating each frame. This allows the framework to generate videos that are semantically consistent with the input captions, even if the captions are long or complex.
7. Evaluation metrics: The authors use several evaluation metrics to compare the performance of their framework to existing methods for video generation from text. These metrics include: BLEU, METEOR, ROUGE-L, CIDEr, FVD, and SSIM. These metrics evaluate the quality, diversity, and semantic consistency of the generated videos.
8. Data augmentation: To improve the performance and diversity of the generated videos, the authors use data augmentation techniques such as random cropping and flipping during training. This helps to ensure that the decoder network can generate visually consistent videos from a wide range of input captions.
9. Limitations: The authors note several limitations of their framework, including its reliance on accurate object detection and tracking, its inability to generate videos with complex camera movements or multiple interacting objects, and its lack of support for generating long-term temporal dependencies.

Overall, these additional points provide important context and insights into the proposed framework, its evaluation, and its limitations.