Report: ViLMA

Jesus Miguel Adrian Matos

July 12, 2024

Abstract

Nothing for the moment

1 Previous concepts:

1.1 Visio-linguistic:

Visio-linguistic refers to the intersection and integration of visual and linguistic (language-based) information. It involves understanding and interpreting meaning that is conveyed through a combination of visual elements (like images, diagrams, and visual symbols) and linguistic elements (like text, spoken words, and written language). This term is particularly relevant in fields such as:

- **Cognitive Science and Psychology:** Studying how the human brain processes and integrates visual and linguistic information.
- **Computer Science and Artificial Intelligence:** Developing algorithms and systems that can understand and generate multimodal content, such as captioning images, generating descriptive text for videos, or creating systems that can engage in visual question answering (VQA).
- **Education:** Using visual aids alongside text to enhance learning and comprehension.
- **Communication and Media:** Designing effective ways to convey information using a combination of visual and textual elements, such as infographics, advertisements, and multimedia content.

In essence, visio-linguistic approaches consider how visual and linguistic inputs complement each other to enhance understanding and communication.

1.2 Image-Language models(IMLs):

An image-language model is an artificial intelligence system designed to process and understand both visual and textual data, integrating these two modalities to perform various tasks. These models can generate text descriptions from images, produce images from text descriptions, and understand the relationships between visual and textual elements. Key capabilities and applications of image-language models include:

- **Image Captioning:** Generating descriptive text for a given image.
- **Visual Question Answering (VQA):** Answering questions related to the content of an image.
- **Image Generation from Text:** Creating images based on textual descriptions.
- **Cross-modal Retrieval:** Finding images that match a given text or vice versa.
- **Image-Text Matching:** Evaluating the relevance or similarity between an image and a text description.

These models typically combine techniques from computer vision (for processing videos) and natural language processing (for handling text). They often use deep learning architectures such as:

- **Convolutional Neural Networks (CNNs):** For extracting features from images.
- **Transformers:** For handling and generating text, and sometimes for processing image features.
- **Multimodal Models:** Like CLIP (Contrastive Language–Image Pretraining) and DALL-E, developed by OpenAI, which are specifically designed to understand and generate both images and text.

1.3 Video-Language models(VidLMs):

Video-language models are advanced machine learning systems designed to understand and generate both video content and associated natural language descriptions. These models can interpret and generate content involving the complex interplay between visual data (videos) and textual data (language). Key capabilities and applications of video-language models:

- **Video Captioning:** Automatically generating descriptive text for video content.
- **Video Question Answering (Video QA):** Answering questions based on the content of a video.
- **Text-to-Video Generation:** Creating video sequences from textual descriptions.
- **Video Retrieval:** Finding relevant videos based on textual queries or descriptions.
- **Action Recognition:** Identifying and classifying actions depicted in video clips.

These models typically combine techniques from computer vision (for processing videos) and natural language processing (for handling text). Here, there are some Architecture and Models:

- **Encoder-Decoder Frameworks:** Commonly used where the encoder processes video frames to create a rich representation and the decoder generates the corresponding textual description.
- **Transformer-based Models:** These models leverage attention mechanisms to handle the complexity of video and language data, such as the Vision Transformer (ViT) and variants like VideoBERT.
- **Fusion Techniques:** Methods to effectively combine and align visual and textual data, such as cross-modal attention and joint embedding spaces.

1.4 Video-language datasets:

Video-language datasets are collections of video clips paired with corresponding textual descriptions, annotations, or other language-based data. These datasets are crucial for training and evaluating models in various tasks, including video understanding, captioning, question answering, and more. Here are some notable video-language datasets:

MSR-VTT (Microsoft Research Video to Text):

- **Description:** A large-scale video dataset with 10,000 video clips and 200,000 sentences. Each video has 20 sentences describing its content.
- **Use Cases:** Video captioning, retrieval, and understanding.

YouCook2:

- **Description:** Contains 2,000 long, unedited videos of cooking activities from YouTube. Each video is annotated with temporal segmentations and textual descriptions of each cooking step.

- **Use Cases:** Video segmentation, action recognition, and video captioning.

Charades:

- **Description:** Features 9,848 videos of daily indoor activities, each annotated with multiple textual descriptions and action labels.
- **Use Cases:** Activity recognition, video understanding, and human action detection.

ActivityNet Captions:

- **Description:** A subset of the ActivityNet dataset, with 20,000 videos from diverse activities. Each video is paired with temporally localized sentences describing the video content.
- **Use Cases:** Dense video captioning, temporal localization, and video summarization.

AVA (Atomic Visual Actions):

- **Description:** Contains 430 15-minute video clips with 80 atomic visual actions annotated per frame.
- **Use Cases:** Action detection, video understanding, and temporal action localization.

1.5 Foil:

descriptions of the image that are highly similar to the original ones, but contain one single mistake ('foil word').

2 Video Language Model Assessment(ViLMA):

Es un benchmark independiente de la tarea que detalla evaluacion de las capacidades de los modelos(VidLMs) sobre una base firme.

Atravez de countrafactuales cuidadosamente selectionados, VILMA ofrece un conjunto de evaluciones controladas que arrojan luz sobre el verdadero potencial de estos modelos(VidLMs).

mientras tales evaluaciones arrojan luz sobre tareas de rendimiento y soporte para analisis coperativo, estos estan limitados a sus habilidades para revelar las capacidades visiolinguisticas que los modelos exiben a lo largo de las tareas.

Comun estructura para cada test:

1. In step 1, be harvested high-quality examples from existing **video-language datasets**.
2. In step 2,be created counterfactual examples or '**foils**', so that a test requires distinguishing correct from counterfactual video+text pairs.