# A Review: MERLOT: Multimodal Neural Script Knowledge Models

The paper "MERLOT: Multimodal Neural Script Knowledge Models" addresses the problem of teaching a computer to interpret temporal information from simply a picture at inference time. The authors describe MERLOT, a model that self-supervised pre-trains over 6M unlabeled YouTube videos to acquire common sense representations of multimodal events. Individual video frames are matched with contextualized representations of the accompanying transcripts, and these frame-level representations are contextualized over time by unmasking distant word-level corruptions and reordering scrambled video frames. The proposed dataset YT-Temporal-180M is derived from 6 million public YouTube videos encompassing a wide range of domains, datasets, and subjects, allowing the model to learn about a diverse assortment of objects, activities, and situations.

The authors explain their motivation by pointing out that our capacity for common-sense thinking is changed by our understanding and experiences with causes and consequences through time. Machines struggle with script knowledge and the capacity to construct temporal reasoning by looking at such visuals. They claim that we can't teach robots to gain this type of information by listing every single fact, conclusion, and counterfactual about a lot of these images.

The authors suggest a resnet-based image encoder, which is followed by a vision-language transformer encoder that is modelled after the RoBERTa basic architecture. To pretrain MERLOT, three objectives are utilized, including 'full-stack' visual reasoning, from recognition subtasks (such as object detection) at the frame level to more 'thinking' tasks at the video level. Contrastive frame-transcript matching, Masked Language Modeling, and Temporal Reordering, which directly teaches temporal thinking, are among these goals.

The authors conduct tests and assessments, transferring MERLOT to 14 downstream jobs and testing on three VCR settings, as well as demonstrating SOTA on zero-shot transfer and fine-tuning. MERLOT outperforms models with heavy supervision, as well as models with bounding boxes, according to the results. As a result, the authors conclude that MERLOT can gain multimodal script knowledge from unlabeled data, allowing the model to reason about events across time and generalize to a broad range of downstream tasks.

The paper has several strengths that contribute to its contribution to the field of multimodal learning. To begin, the dataset used to train MERLOT is a significant strength since it spans numerous domains and covers a wide variety of themes, allowing the model to learn about a diverse selection of objects, actions, and scenarios. Furthermore, the scientists did not restrict the dataset to instructional videos alone, making it more diversified and useful to a variety of downstream applications. The authors also sought to

preserve users' privacy by allowing them to opt out of the dataset.Another advantage of the MERLOT Image Encoder is its speed, which is 50X quicker than Faster-RCNN. Moreover, initializing the combined vision-language encoder using RoBERTa weights provides the encoder with a broad variety of natural language processing capabilities while potentially reducing total training time. The pre-training objectives assure full-stack visual reasoning, ranging from identification subtasks at the frame level to more 'thinking' challenges at the video level. The combination of contrastive loss and temporal reordering produces high outcomes for downstream tasks, demonstrating the usefulness of the pre-training strategies.

The paper does, however, have certain limitations. For starters, the collection only contains movies in English, restricting its applicability to other languages. Second, video descriptions are repetitious and may contain inaccurate captions, lowering the dataset's quality. Apart from efficiency factors, there is no explanation of why the grid-based hybrid ResNet/Vision Transformer was the best choice. Also, the model does not learn from audio, which may be an issue for some applications. Another limitation is the potential loss of intra-video coherence when all other frame-caption pairings in the batch are treated as negative samples, even if they come from the same movie. There is also a lack of human appraisal and experimentation with architectural alterations. Further, the model in this case has been trained on millions of YouTube videos. Thus, the model learns a poor alignment between videos and languages as spoken words do not always correlate to visual data.