A Review: Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input

The intriguing subject of whether or whether a machine can acquire both spoken language and visual perception, with similar constraints as babies, is explored in the paper "Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input." To address this, the authors propose a neural model that can identify the areas of a picture that correspond to the speaker's verbal description. The paper is well-motivated because it recognizes that infants learn to talk by breaking down continuous speech, a task that is more difficult than parsing text due to the idiosyncrasies of each speaker's voice, dialects, emotions, and environmental noise.

The authors utilize a dataset of unaligned picture and audio descriptions (1100 hours of audio with ten-second captions), which are more detailed than MSCOCO captions, to evaluate their method. The authors expanded the original Places Audio Caption dataset by an additional 200K captions for this research. Importantly, this dataset calls for a model to identify in-image objects using only the audio description.

In order to encode both the visual image and the spoken audio caption into a common embedding space, the authors propose using a pair of convolutional neural networks. Instead of mapping whole images and spoken utterances to fixed points in an embedding space, as was done in previous work, this model learns representations that are spatially and temporally distributed, enabling it to co-localize directly inside both modalities. The models are optimized for a ranking criterion, such that related photos and captions share more of an embedding space than unrelated ones. The imaging branch of the architecture is based on the framework established by the VGG16 network. The authors employ a model that provides a feature map over the audio rather than a single embedding vector in order to represent the spoken audio captions during training. Sum image sum audio, Max image sum audio, and Sum image max audio were the three similarity metrics employed by the authors. Throughout the training of all models, the sampled margin ranking objective was used.

Image/caption retrieval, object localisation guided by speech, audio-visual pattern clustering, and concept discovery were the four tasks used to assess the model. When pre-trained, the model outperformed prior state-of-the-art models on the image/caption retrieval task, using the MISA similarity measure. Nevertheless, there was no spatial or temporal localisation in the evaluation, just a broad assessment of representation quality. The authors employed ADE20k's part-level annotations to determine which objects were being described in the captions for the

speech-prompted object localization challenge. Although this is a useful exercise, more insight may have been gained if the authors had relied on human review to determine object localisation instead of thresholding and selecting word-object combinations. Extraction of related components from the matchmap, averaging and concatenating the audio and visual components, and executing hierarchical clustering were all interesting steps in the clustering of audio-visual patterns challenge. The authors annotated each cluster with its most common word as determined by automatic speech recognition (ASR). Using a concept discovery task, the authors found the "concept"-corresponding dimension in both the visual and auditory networks and then displayed the activation maps in the spatial and temporal activation maps. Unfortunately, there was no examination of the overlap between the thousands of unique terms in the dataset and more frequent words, and the concept score did not take into account numerous words in the audio caption.

Many caveats exist in the paper. To begin with, it is challenging to scale since this dataset was not scraped off the internet and was carefully acquired. Second, the held-out set represents less than 1% of the dataset (1K out of 400K image-caption pairings). Third, a pre-trained language component, which would have enhanced performance, was not used by the authors. The fourth argument is that human judgment should be used to evaluate object location. Some of the thresholding and selection of words and objects may have been avoided. Sixth, there was a decline in the quality of the word-object clusters after the initial few were formed. Sixth, there is no examination of the degree of overlap between the unique terms in the dataset and more frequent words, which would be useful for concept discovery. Finally, the findings may have been impacted because the concept score did not take into consideration the presence of numerous words in the audio caption.