E2E-VLP: End-to-End Visual-Language Pre-training Enhanced by Visual Learning
Alibaba Group ACL 2021

Until now pre-training for cross-modal downstream tasks included only image based features and not text based features. Image based features(region based visual features) were added through an object detection model whose image representation embedding was concatenated with text embedding which was then fed into the transformer. As these models just included image representation in the pre-training and not text semantics, these have a high chance of the object not being identified correctly and therefore not optimized for cross-modal tasks. So the authors in the papers propose E2E-VLP, the first end-to-end pixel-level vision-language pre-trained model in which encoder-decoder transformers jointly learn both visual(object detection) and semantic features(image captioning) before doing task specific fine-tuning to improve visual language understanding by mapping region extracted to the image caption.

The E2E-VLP model is pre-trained on Masked Language Modeling and Image-text matching along with object detection and Image-text generation. To learn cross-modal features, an encoder-decoder transformer is used. The input to the encoder is sentence embedding from the caption text along with the image embedding which are extracted by training Faster-RCNN on Visual Genome dataset which learns cross-modal interactions between image-grid and language tokens. Transformer decoder is then used to help capture fine-grained text and image features where the object detected is mapped with the image caption.

The E2E-VLP model was evaluated on various vision-language tasks which included visual question answering, natural language visual reasoning, cross modal retrieval and image captioning.

I found it very interesting that researchers have finally started to exploit image and text features together to better understand and perform complex downstream tasks but it's a long way to go as they can be exploited more for our advantages. The only thing I found a little strange is they are exploiting pixel-level features which are too fine-grained and may exploit unwanted features and will be time-consuming and expensive. I found the paper easy to read, understand and well-structured.

To conclude, authors proposed the E2E-VLP model to incorporate text semantics along with visual features in cross-modal downstream tasks as opposed to just object detection.