VLMo

- is a pretrained Vision-Language Model
- first work to introduce the Mixture-of-Modality-Experts (MoME) architecture for Transformer models
- here modality encoders, so image and text encoders, are not seperate
- image and text encoder are build using a Fusion Encoder
  - encoder consists of, as usual, Transformer blocks, but with two MLPs per block instead of one
  - one MLP for image and one for text
  - inspired by the Mixture-of-Experts (MoE) architecture