Vision Language Model

- A General Classification of Vision Language Models (VLMs)

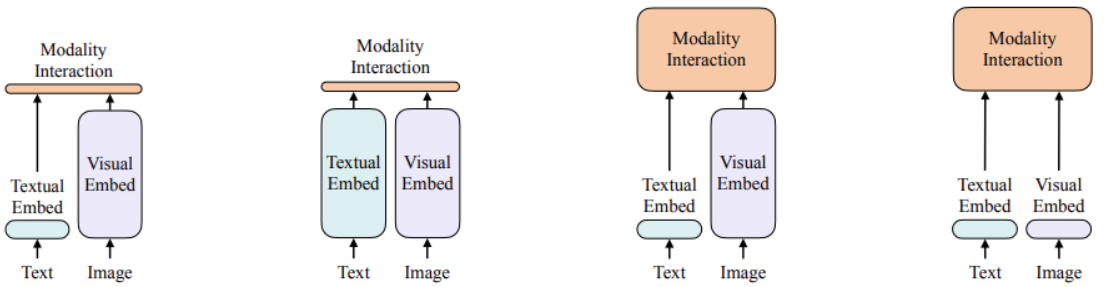

Figure 2. Four categories of vision-and-language models. The height of each rectangle denotes its relative computational size. VE, TE, and MI are short for visual embedder, textual embedder, and modality interaction, respectively.

  - Class 1: Small text encoder + Large vision encoder + Simple interaction
    - VSE++, SCAN
  - Class 2: Large text encoder + Large vision encoder + Simple interaction
    - CLIP
  - Class 3: Small text encoder + Large vision encoder + Large interaction
  - Class 4: Small text encoder + Small vision encoder + Large interaction

- **Learning Transferable Visual Models From Natural Language Supervision [arXiv 2021]** Alec Radford (arXiv) (pdf) (Citation: 7930)

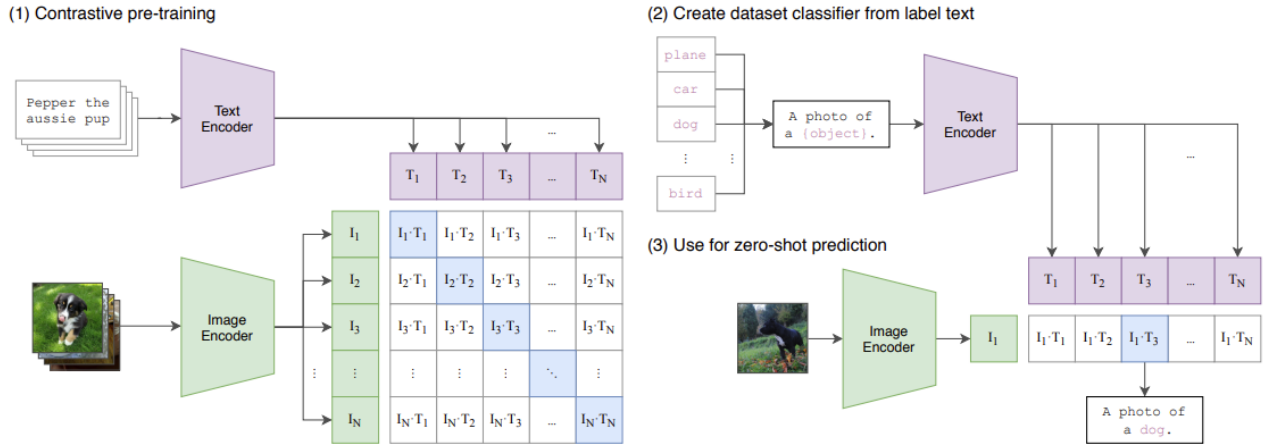

Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

  - Core: use natural language to supervise image training
    - Map visual semantic and language semantic
  - Constrastive pre-training:
    - Training data - Image & text pairs
    - Image encoder (ViT, ResNet) - Extract image features
    - Text encoder (Transformer) - Extract text features
    - Training goal - Minimize the distance (cosine similarity) between text and image features if they are a pair (i.e., I_1T_I, I_2T_2 in the image).

        - Both image and text encoders are trained from scratch.
    - Downstream task example: Object detection
        - Prompt template: `A photo of {objects}`, this `{objects}` are objects candidates, i.e., dog, cat, etc.
        - Calculate the consine similarity between the prompt features and image features and find the one with the highest similarity.

- **ViLT Vision-and-Language Transformer Without Convolution or Region Supervision [`ICML 2021`]**
  Wonjae Kim, Bokyung Son, Ildoo Kim (arXiv) (pdf) (Citation: 808)

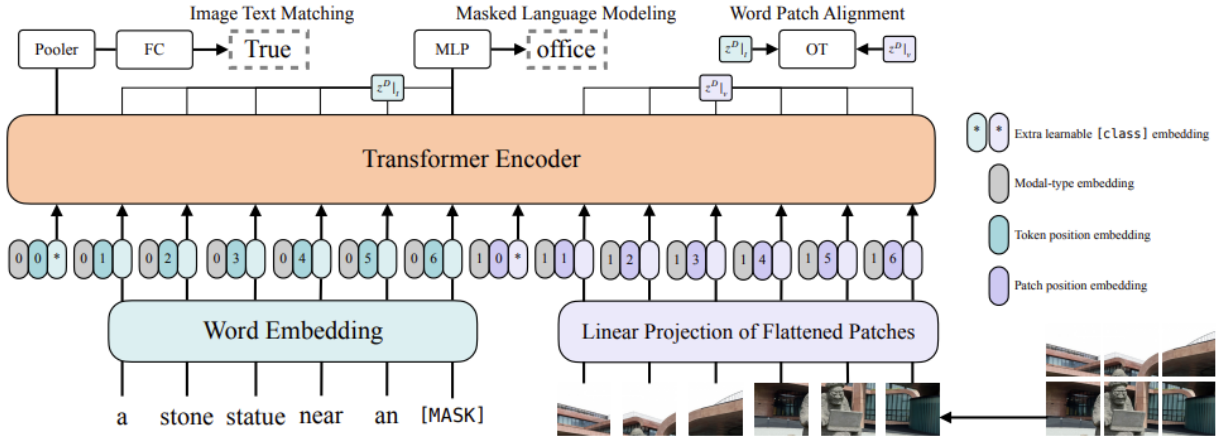

Figure 3. Model overview. Illustration inspired by Dosovitskiy et al. (2020).

    - Remove feature extraction model in traditional Transformer-based vision detector
    - The Performance of ViLT is similar to the feature extration method, but the running time is much shorter.