Contrastive Language-Image Pre-training & Latent Diffusion

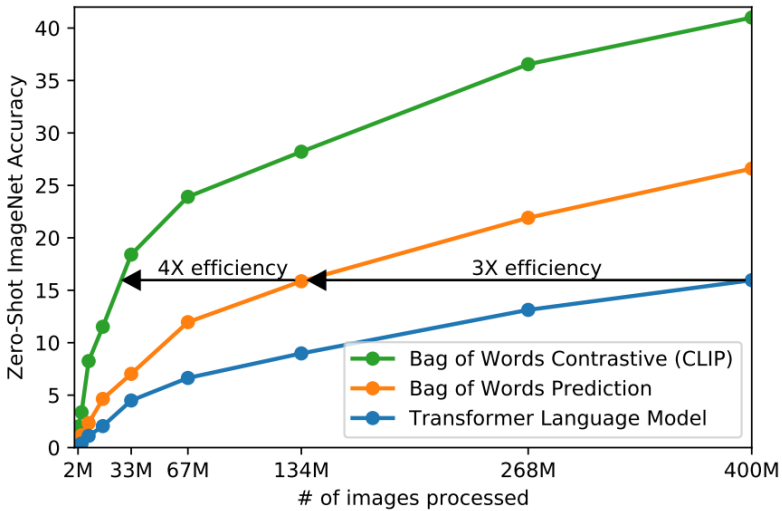
David Valecký

CLIP

- Learning method from natural language supervision
- Dataset of 400 million (image, text) pairs
- Perform a wide set of tasks during pre-training
  - OCR
  - geo-localization
  - action recognition

Comparison with other methods

- Benchmark on 30 existing datasets
- Strong at zero-shot classification
- Better than public ImageNet models



Dual-Input model architecture

- Two types of inputs: an image and a text prompt
- Vision encoder for images
- Language encoder for text

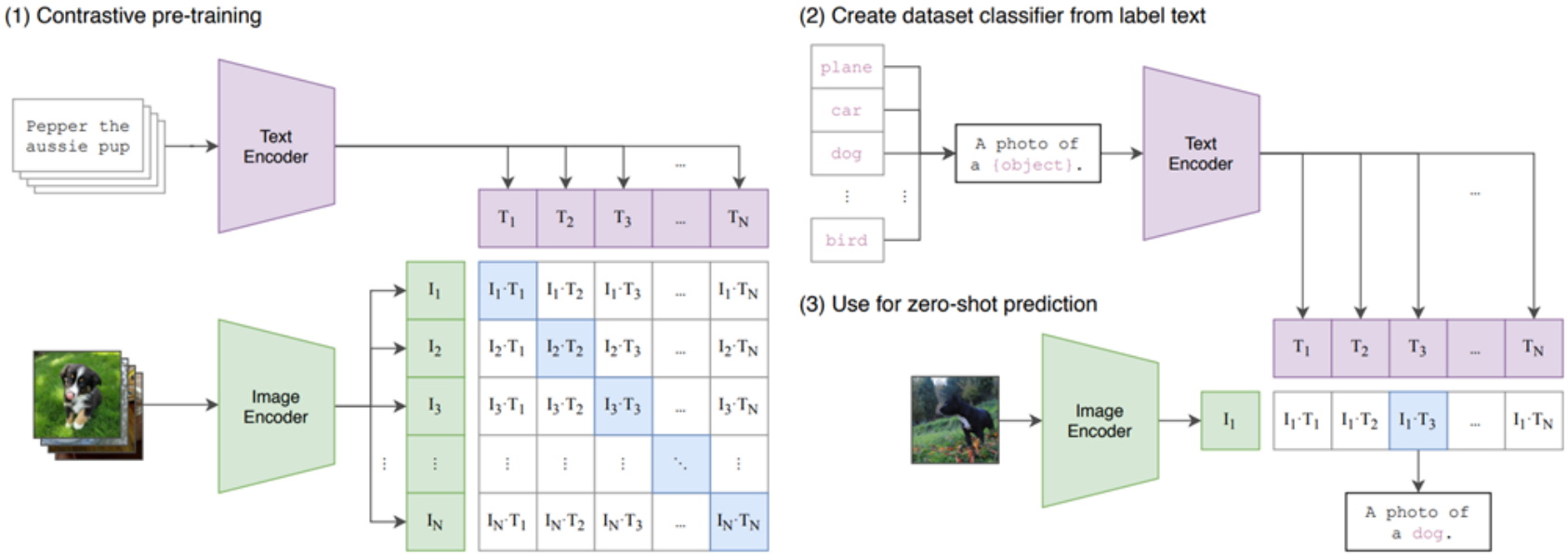


Image Encoder

- Variant of Vision Transformer (ViT)
- Based on CNN
- Originally designed for natural language processing tasks
- Employs self-attention mechanisms
- High-dimensional vector representations - CLIP image embeddings
    - Space where images and text are jointly represented
    - 3 vectors - Query (Q), Key (K), Value (V)
    - Self-attention - can run in parallel

Text encoder

- Transformer-based architecture
- No pre-trained weights
- Linear projection to map from each encoder’s representation to the multi-modal embedding space
- Simplified because many CLIP’s pre-training dataset are only a single sentence

Performance

- The largest ResNet model
- Model name - RN50x64
  - 18 days to train on 592 V100 GPUs
- Largest Vision Transformer
  - 12 days on 256 V100 GPUs
- Model name - ViT-L/14

Zero-Shot Transfer

- Performing unseen tasks
- Standard image classification datasets using a generically pre-trained model
    - Second study zero-shot transfer to existing image classification datasets at the time of publishing paper

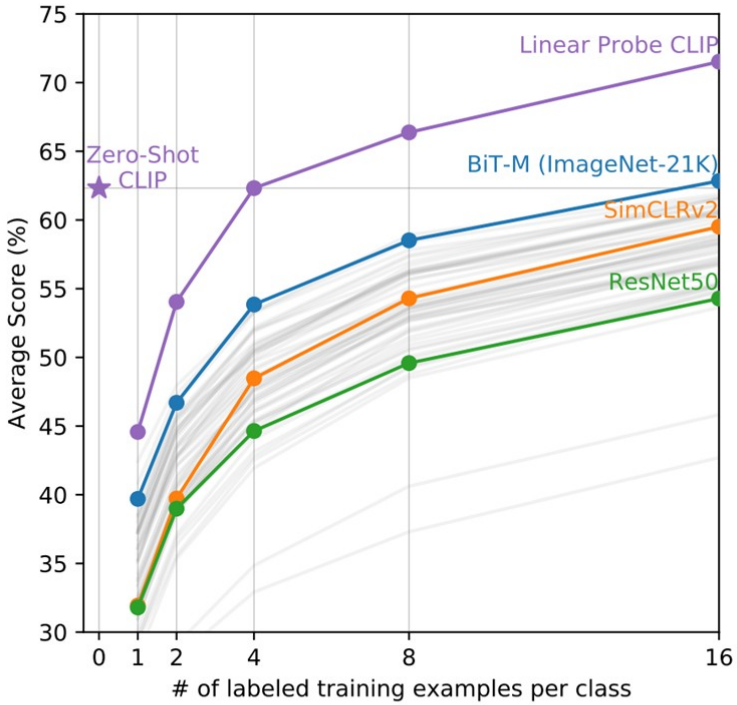


Figure 6. **Zero-shot CLIP outperforms few-shot linear probes.** Zero-shot CLIP matches the average performance of a 4-shot linear classifier trained on the same feature space and nearly matches the best results of a 16-shot linear classifier across publicly available models. For both BiT-M and SimCLRv2, the best performing model is highlighted. Light gray lines are other models in the eval suite. The 20 datasets with at least 16 examples per class were used in this analysis.

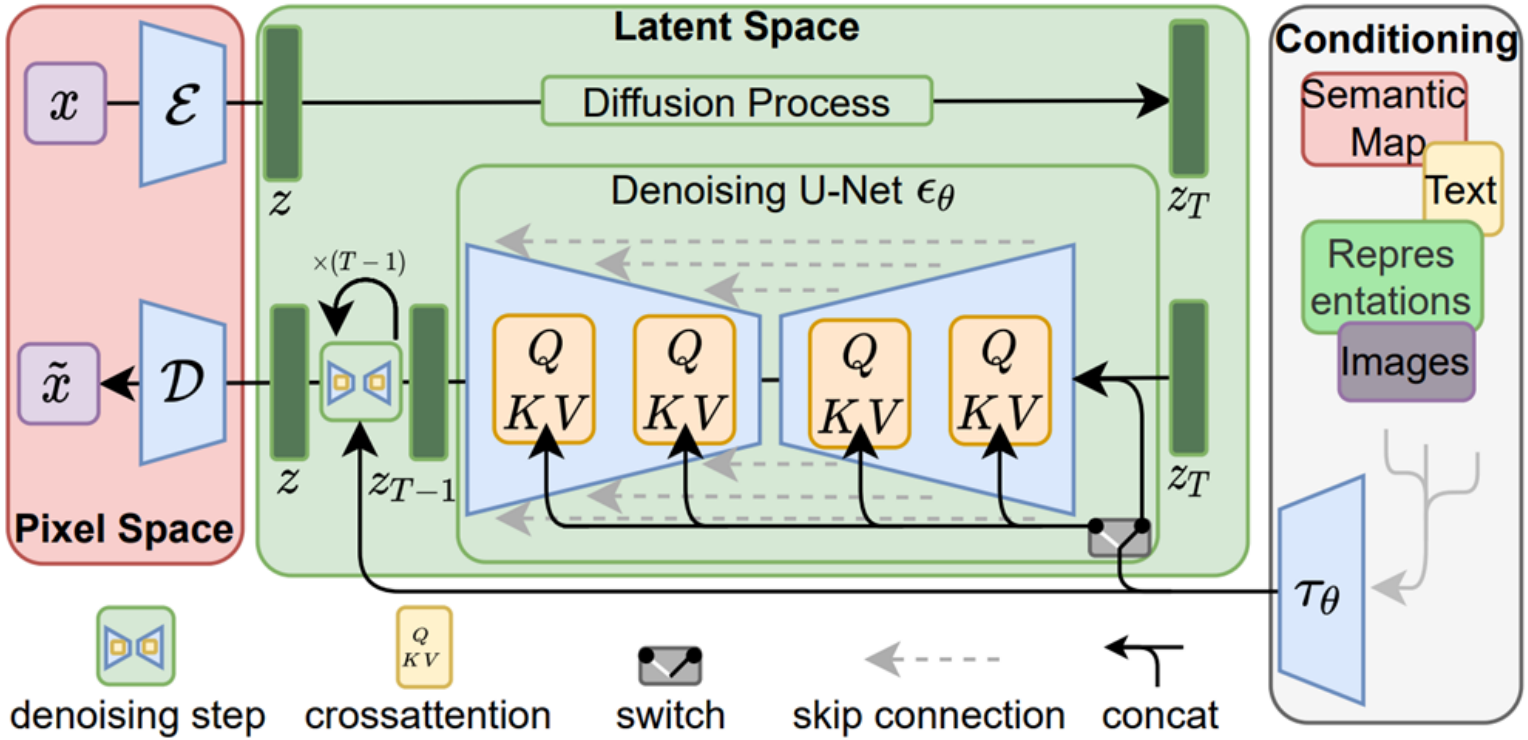
Applications and Use Cases

- Image Classification
- Text-to-Image Retrieval
- An ability to generalize to a wide array of task

Latent Diffusion

- Autoencoder (VAE)
- U-Net
- Text Encoder - CLIP, pre-trained

Latent diffusion model



Autoencoder (VAE)

- Transforms a high-dimensional image (512x512x3) into a compact latent representation (64x64x4)
- "Latents" serve as input to the U-Net
- Conversion significantly reduces memory requirements (48 times less) compared to pixel-space diffusion models
- The decoder reconstructs the original image from the latent representation
- Only the decoder is needed to convert denoised latents into actual images
- Isotropic Gaussian distribution

U-Net

- Predicts denoised image representations from noisy latents
- By subtracting this noise from the noisy latents, actual latents are obtained
- Contains:
  - Encoder (12 blocks)
  - Middle block
  - Skipconnected Decoder (12 blocks)
  - Totally 25 blocks
- 8 blocks handle down-sampling or up-sampling convolution layers
- 17 blocks containing four ResNet layers and two Vision Transformers (ViTs)

Text Encoder

- Stable Diffusion uses a pre-trained text encoder CLIP
- Latent space can be used to train multiple generative models
- Downstream application single-image CLIP-guided synthesis

Odkazy

- https://arxiv.org/pdf/2112.10752.pdf
- https://cdn.openai.com/papers/dall-e-2.pdf
- https://arxiv.org/pdf/2103.00020.pdf