BEiT

Research Problem/Objective

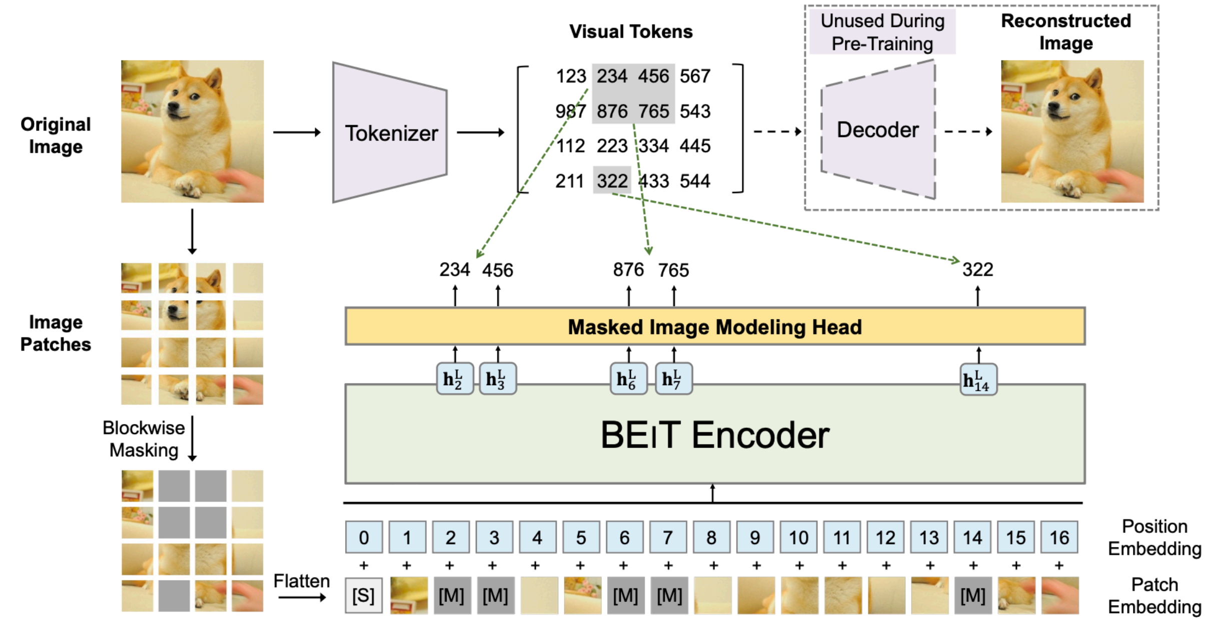
- Vision Transformers (ViTs) achieve strong performance but typically require large amounts of labeled data for training (data-hungry).
- Self-supervised pre-training is a promising solution, but directly applying BERT's Masked Language Modeling (MLM) concept to images is challenging. Specifically:
  - Images lack a predefined discrete vocabulary like text (words/subwords).
  - Predicting raw pixel values for masked patches (pixel regression) is suboptimal, often forcing the model to focus on low-level details and short-range dependencies rather than higher-level semantics.
- The objective is to develop an effective self-supervised pre-training method for ViTs, inspired by BERT, that overcomes these challenges and learns robust visual representations from unlabeled images.

Core Idea/Hypothesis

- The central hypothesis is that a ViT can be effectively pre-trained using a Masked Image Modeling (MIM) task where it learns to predict discrete **visual tokens** corresponding to masked **image patches**.
- The core idea is to use two different views of an image:
  1. **Image Patches:** The standard input for ViTs (raw pixel values).
  2. **Visual Tokens:** Discrete latent codes representing the image content, obtained from a separate, pre-trained image "tokenizer" (specifically, a discrete Variational Autoencoder - dVAE).
- By predicting these abstract visual tokens instead of raw pixels for the masked parts, the model is guided to learn higher-level semantic representations rather than getting stuck on surface-level reconstruction.

Model Architecture (Conceptual Overview)

- The backbone network is a standard Vision Transformer (ViT) architecture (e.g., ViT-Base, ViT-Large).
- Input consists of a sequence of linearly projected image patches, prepended with a [S] token, and combined with learnable positional embeddings.



Pre-training Task: Masked Image Modeling (MIM)

1. **Image Tokenization:** An input image x is processed in two ways:
   - Split into a grid of N image patches {xp_i} (e.g., 16x16 pixels each). This is the input to the Transformer.
   - Tokenized into a grid of N discrete visual tokens {z_i} using a pre-trained dVAE image tokenizer. The vocabulary size |V| is 8192. These tokens serve as the prediction target.
2. **Masking:** A significant portion (~40%) of the input image patches are randomly masked.
   - **Blockwise Masking:** Instead of masking individual random patches, contiguous rectangular blocks of patches are masked to encourage learning beyond local redundancies.
   - Masked patches in the input sequence are replaced with a shared, learnable [M] embedding.
3. **Prediction:** The corrupted sequence of patches (unmasked patches + [M] embeddings) is fed through the ViT encoder.
4. **Objective:** For each masked position i, a softmax classifier placed on top of the corresponding final Transformer output vector hL_i predicts the original visual token z_i for that patch. The model is trained to maximize the log-likelihood of predicting the correct visual tokens for all masked positions.
   - Loss = - Σ (over masked i) log p(z_i | corrupted_image)

Key Innovations & Contributions

- **Masked Image Modeling (MIM):** Proposing a novel BERT-like pre-training task for vision, specifically predicting discrete visual tokens for masked image patches.

- **Two-View Image Representation:** Utilizing image patches as input and discrete visual tokens (from dVAE) as the prediction target, effectively bridging pixel space and a more semantic discrete space.
- **Demonstrating Effectiveness of BERT-style Vision Pre-training:** Showcasing that masked auto-encoding can work very well for ViTs if the prediction target is appropriately chosen (visual tokens vs. raw pixels).
- **Blockwise Masking for Images:** Adapting and showing the benefit of blockwise masking strategy in the visual domain for the MIM task.
- **Strong Empirical Results:** Achieving state-of-the-art or competitive performance compared to previous self-supervised and supervised pre-training methods on downstream tasks.

Research Methodology & Evaluation (High-Level)

- **Methodology:** Pre-trained ViT-Base and ViT-Large models using the MIM task on the unlabeled ImageNet-1K dataset (1.2M images) for 800 epochs. Used a publicly available dVAE tokenizer (from DALL-E work).
- **Downstream Tasks:** Evaluated by fine-tuning the entire pre-trained BEIT model on:
  - Image Classification: ImageNet-1K, CIFAR-100.
  - Semantic Segmentation: ADE20K.
- **Evaluation Metrics:** Top-1 Accuracy (Classification), mean Intersection over Union (mIoU) (Segmentation).
- **Comparisons:** Compared against training ViTs from scratch, supervised pre-training (on ImageNet-1K/22K), and other self-supervised methods (DINO, MoCo v3, iGPT). Also evaluated intermediate fine-tuning on ImageNet before the final task.

Significant Findings & Results

- BEIT significantly outperforms training ViT from scratch on all evaluated tasks.
- BEIT achieves state-of-the-art results among self-supervised methods on ImageNet fine-tuning (e.g., BEIT-B 83.2% Top-1).
- BEIT pre-training surpasses standard supervised ImageNet-1K pre-training for semantic segmentation on ADE20K.
- BEIT scales effectively: BEIT-L shows significant gains over BEIT-B. Large BEIT models pre-trained only on ImageNet-1K can outperform supervised models pre-trained on larger labeled datasets (e.g., BEIT384-L outperforms ImageNet-22K supervised ViT384-L).
- Ablation studies confirmed that:
  - Predicting visual tokens is crucial and significantly better than predicting raw pixels. Pixel prediction performed poorly.
  - Blockwise masking provides additional benefits, especially for segmentation.
- Fine-tuning BEIT converges significantly faster than training from scratch.

- Analysis of self-attention maps shows BEIT learns meaningful object boundaries and semantic regions without explicit supervision.

Conclusions & Implications

- BEIT demonstrates that BERT-style masked auto-encoding pre-training is a highly effective approach for Vision Transformers when using discrete visual tokens as the reconstruction target.
- This approach successfully learns robust and transferable visual representations from unlabeled data, reducing the dependency on large labeled datasets for training high-performing ViTs.
- Predicting semantic tokens instead of raw pixels is a key factor for the success of masked image modeling.
- BEIT provides strong performance on downstream tasks, often matching or exceeding supervised pre-training, and scales well to larger models. It also offers faster convergence during fine-tuning.