Exercise Session 7

MAEs, iBOT and DINO

Prepared by Lukas Klein and Benedikt von Querfurth

Overview

Background Information. This exercise session is about Masked Autoencoders (MAEs, Paper), Image BERT Pre-Training with Online Tokenizer (iBOT, Paper) and knowledge DIstillation with NO labels (DINO, Paper). We will discuss how these image pretraining methods differ and compare to the already reviewed methods from natural language, how they relate and build upon each other, and how we can use them to train powerful foundation models. After this exercise sheet, you should be able to understand the pros and cons of each method, their failure modes, and how to apply them at scale.

It is not required to read the respective papers, and it is encouraged to first write down your thoughts for the "understanding the concepts" questions before doing so.

Task 1. MAE: Understanding the Concepts.

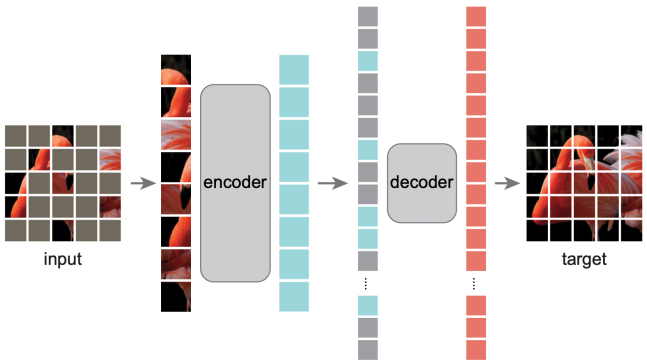

Figure 1: Architecture overview of MAE.

During pre-training via MAE, a large random subset of image patches (e.g., 75%) is masked out (see Figure 3). The encoder is applied to the small subset of visible patches. Mask tokens are

introduced after the encoder, and the full set of encoded patches and mask tokens is processed by a small decoder that reconstructs the original image in pixels. After pre-training, the decoder is discarded and the encoder is applied to uncorrupted images (full sets of patches) for recognition tasks (from He et. al. (2021)).

(a) In Exercise Session 1 (Task 2), we talked about BERT style masked language modeling. How must masked pretraining be adapted when transitioning from language models like BERT to image data, and what new challenges arise from the inherent characteristics of visual data?

(b) Given that alternative image pretext tasks such as jigsaw puzzles (predicting the right position of a patch, Paper) or in-coloring (inpainting the colored version of a black and white image, Paper) have been proposed, why do you think MAEs in the end prevailed for self-supervised visual representation learning?

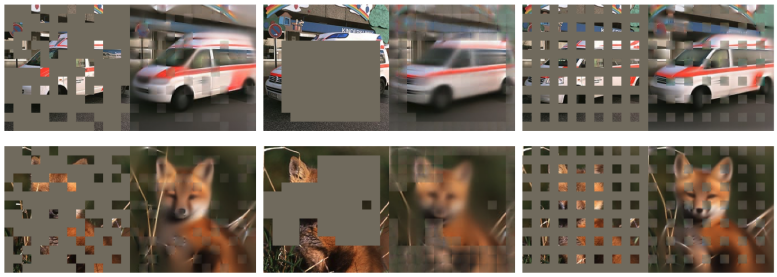

Figure 2: The three different masking strategies examined in MAE.

(c) How do you think are the three proposed masking strategies (random sampling, block-wise masking, and grid-based masking, see Figure 2) differentially influence the pretraining dynamics and learned representations?

Task 2. MAE: Optimal Masking Ratios in MAE vs. BERT.

In this exercise, we analyze why BERT-style models (discrete 1D sequences) and MAE (continuous 2D images) have different optimal masking ratios. We'll use information-theoretic arguments to discuss why BERT performs best around 15% masking while MAE benefits from 75% masking.

Setup

Text sequences (BERT):

- Input: Token sequence $\mathbf{x} = (x_1, x_2, \ldots, x_L)$ where each $x_i \in \{1, \ldots, V\}$ (vocabulary size $V \approx 30000$)
- Masking: Randomly mask a fraction r of tokens

- Let $\mathbf{x}_v$ denote visible tokens and $\mathbf{x}_m$ denote masked tokens (where $|\mathbf{x}_m| = m := rL$)

Image patches (MAE):

- Input: Image divided into N patches $\mathbf{p} = (p_1, p_2, \ldots, p_N)$ (e.g., $N = 196$ for 14×14 grid)
- Masking: Randomly mask a fraction r of patches
- Let $\mathbf{p}_v$ denote visible patches and $\mathbf{p}_m$ denote masked patches (where $|\mathbf{p}_m| = m := rN$)

Tasks

(a) Mutual Information for Text Sequences

Consider text sequences with mutual information $I(\mathbf{x}_m; \mathbf{x}_v) = H(\mathbf{x}_m) - H(\mathbf{x}_m|\mathbf{x}_v)$. Derive an expression for $I(\mathbf{x}_m; \mathbf{x}_v)$ using vocabulary size V and an effective vocabulary size $V_{\text{eff}}(r)$ (i.e. the possible tokens based on the context from the none masked tokens) that captures context constraints, then explain how $I(\mathbf{x}_m; \mathbf{x}_v)$ changes with masking ratio r considering text's 1D sequential structure and low redundancy.

(b) Mutual Information for Image Patches

Consider images with differential entropy $I(\mathbf{p}_m; \mathbf{p}_v) = h(\mathbf{p}_m) - h(\mathbf{p}_m|\mathbf{p}_v)$. Explain why images have high spatial redundancy (2D connectivity, local correlation) and why $I(\mathbf{p}_m; \mathbf{p}_v)$ decays slowly with increasing r, distinguishing between texture interpolation (low-level) and semantic understanding (high-level) prediction strategies. For simplicity, assume that each patch is only affected by its 8 direct neighbors and that they have properties of a Markov Random Field (MRF, not exam relevant), i.e. each masked patch is independent from all other patches when conditioned on its direct neighbors (local Markov property).

(c) Relationship between Image and Text Masking Ratios

From previous exercises, we know that mutual information $I(\mathbf{x}_m; \mathbf{x}_v)$ decays rapidly for text (1D structure, 2 direct neighbors) but slowly for images (2D structure, 8 direct neighbors). Model this as exponential decay $I(r) = I_0 e^{-\lambda r}$ where λ is the decay rate, and prove that the optimal masking ratio $r^* = \arg\max_r [I(r) \cdot r]$ satisfies $r^*_{\text{text}} < r^*_{\text{images}}$ under the assumption $\lambda_{\text{text}} > \lambda_{\text{images}}$.

Task 3. iBOT: Understanding the Concepts.

iBOT combines several mechanisms that we saw before in MAE but also in BERT in Exercise Session 1 and in BYOL in Exercise Session 2. Specifically, it combines self-distillation with masked image modeling by using a student-teacher framework where both networks process two differently augmented views (u and v) of the same input image (see Figure 3). The student network (f_s) processes a view with randomly masked patches (indicated by the hatched pattern), while the teacher network (f_t) processes an unmasked view and uses exponential moving average updates without gradient backpropagation (stop grad). Each encoder produces two types of outputs: a global `[CLS]` token representation and patch-level tokens, both passed through separate prediction heads ($h^{\texttt{[CLS]}}$ and h^{patch}). The student learns by matching its predictions

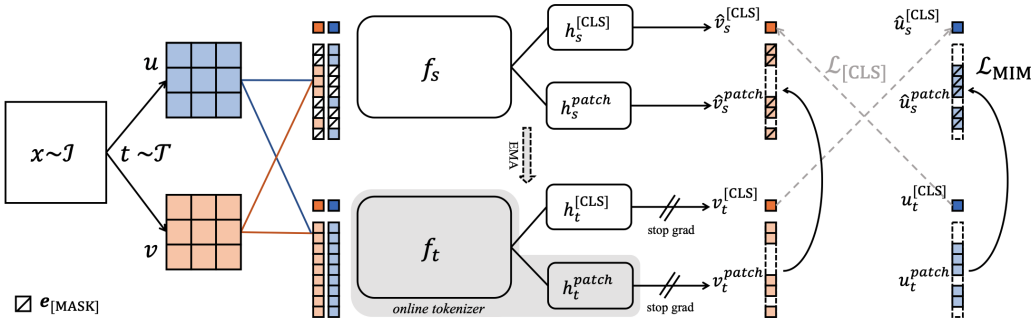

Figure 3: Architecture overview of iBOT.

to the teacher's outputs through two complementary objectives: a [CLS]-level contrastive loss ($\mathcal{L}_{\texttt{[CLS]}}$) that aligns global image representations, and a masked image modeling loss ($\mathcal{L}_{\text{MIM}}$) that predicts the teacher's patch tokens for masked regions. This dual approach allows iBOT to learn both semantic global-level understanding through self-distillation and fine-grained local features through the masked prediction task.

(a) Given that iBOT employs BERT-style masked token prediction in the feature space, what are the comparative advantages and disadvantages of this approach relative to direct pixel-level image reconstruction in the original input space as in MAE?

(b) How does iBOT's architecture and training methodology differ from the original BERT framework, and what motivates these architectural adaptations for image data?

(c) How does iBOT's student-teacher framework differ from BYOL's approach, and what practical advantages emerge from maintaining separate patch-level and [CLS] token representations?

Task 4. iBOT: Centering and Mode Collapse Prevention in iBOT.

To prevent mode collapse, iBOT applies a centering operation to the teacher outputs.

Setup

The output of the teacher is defined as:

$$\mathbf{z}_t = \text{softmax}\left(\frac{h_t(f_t(\mathbf{x})) - \mathbf{c}}{\tau}\right),$$

where τ is a temperature parameter and $\mathbf{c}$ is a running center updated as:

$$\mathbf{c} \leftarrow \lambda \cdot \mathbf{c} + (1-\lambda) \cdot \mathbb{E}_{\text{batch}}[h_t(f_t(\mathbf{x}))].$$

Further, consider the cross-entropy loss used in self-distillation:

$$\mathcal{L} = -\sum_k p_k \log q_k$$

where p_k are teacher outputs and q_k are student outputs.

Tasks

(a) Mode Collapse Without Centering

Explain why without centering, the model might collapse to producing constant outputs. Use the cross-entropy loss between student and teacher outputs and how it creates a reinforcing feedback loop (i.e. how teacher update effects student update which again effects the teacher update).

(b) Zero-Mean Property

Show that centering encourages the mean output to be zero: $\mathbb{E}[f_t(\mathbf{x}) - \mathbf{c}] \to \mathbf{0}$. Derive the equilibrium value of $\mathbf{c}$ when the training reaches a stationary distribution.

Task 5. DINOv2: Coding Exercise.

See ipython notebook.
