SAPIENZA
UNIVERSITÀ DI ROMA

Visual Spectral Probing

Faculty of Information Engineering, Informatics, and Statistics
Master's degree in Data Science

Andrei Caraman
ID number 1664744

Advisor
Prof. Fabrizio Silvestri

Academic Year 2024

Thesis defended on 19 January 2024
in front of a Board of Examiners composed by:

prof. Brutti Pierpaolo (Presidente)
prof. Casalicchio Emiliano
prof. Daraio Cinzia
prof. Galasso Fabio
prof. Quattrociocchi Walter
prof. Silvestri Fabrizio
prof. Scardapane Simone (chairman)

Visual Spectral Probing
Master thesis. Sapienza University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Website: https://github.com/AndreiCaraman/Visual-Spectral-Probing
Author's email: caraman.1664744@studenti.uniroma1.it

To my beloved mother, dear family, and friends – your support has been my anchor throughout this extensive chapter of my life.

Abstract

This thesis explores a novel approach to enhance interpretability and aims to improve task-specific classification performance in Visual Transformers for image recognition.

Traditional computer vision relies on feature extraction, but this work employs spectral analysis to dissect and control information at different scales. By applying spectral filters to ViT activations, the study aims to disentangle scale-specific information, learning which frequencies are relevant for image classification tasks.

The experiments use two datasets, COCO Backgrounds and Food 101, with minimal and structured variability, respectively.

Experiments include probing accuracy analysis, spectral profiling, and inspecting filter weights.

Results indicate a marginal improvement in accuracy with a differentiable autofiltering technique, while band-pass modes show diminished performance. Inspecting filter weights does not reveal a clear contribution by distinct frequency bands.

We conclude by suggesting that future research may explore innovative training approaches using bounding boxes for image embeddings.

The accompanying code for implementing these analyses is available on the url: `https://github.com/AndreiCaraman/Visual-Spectral-Probing`. This repository contains the codebase used for our experiments and can be accessed for further details on the methodology employed.

Contents

Chapter 1

Introduction

Images depict subjects that exhibit structured patterns at various scales, spanning from low-level pixel arrangements in textures, to nearly-uniform pixel arrangements in backgrounds. Such structure can be leveraged to improve image classification without drawing on specific priors.

Prior work in computer vision has shown how these kinds of structures can be analysed employing techniques such as feature extraction to unveil known visual levels of structure. One notable example is scene understanding through multi-level analysis. This involves capturing structures at various levels, from low-level features like edges and textures to mid-level components such as objects and their spatial arrangements [8]. Incorporating these different levels of information enables a more comprehensive understanding of complex visual scenes. This approach facilitates not only object recognition but also the extraction of meaningful relationships and context within a given scene.

In the following work we follow a different apporach by employing tools from spectral analysis, widely used in signal processing and other fields, to separate and control information at different scales.

Intuitively, any sequence of values, such as a neuron's activations across input tokens, can be represented as a weighted sum of cosine waves with different frequencies. The weight for a particular frequency indicates the amount of structure in the sequence at that scale: weight on higher frequencies indicates faster changes in the neuron's activation from token to token, while weight on lower frequencies indicates activations that shift more gradually across an input. By removing certain frequencies, called spectral filtering, we can remove information about variation at particular scales.

In this work, we apply spectral filters to the activations of individual neurons in the Visual Transformer [5]. This enables the separation of information in model representations that changes at different rates across the input — for example, subject details such as human faces have a high variability at the pixel level, while image backgrounds are much more gradual.

The ultimate goal is to learn which frequencies are relevant for a given task of image classification by disentangling scale-specific information in existing embeddings and train the model to learn more about particular scales.

Chapter 2

Related Work

Analyzing the contextualized embedding representations of pre-trained language models (LMs) using lightweight probes, as discussed by Hewitt and Liang [7], has unveiled latent features within the untuned encoders. These features are notably relevant to downstream Natural Language Processing (NLP) tasks across various layer depths, as highlighted by Tenney et al. in 2019 [14]. Furthermore, linguistic phenomena are observed to be encoded at different timescales, encompassing rapidly changing (sub-)word-level information and slower-changing sentence or paragraph-level information.

In a related vein, the decomposition of contextualized embeddings from BERT [4] into frequencies has provided valuable insights into the pertinent frequency bands applicable to a diverse array of NLP tasks, as demonstrated by Tamkin [13].

Drawing inspiration from these seminal works in the realm of NLP, we extend and apply these analytical methods to the domain of images. This approach seeks to unravel latent patterns and relevant features in contextualized embeddings within the image domain, akin to their counterparts in the NLP domain.

2.1 The discrete cosine transform and spectral filters

To analyze sequences in the frequency domain, a spectral transform is essential. In this study, we employ the discrete cosine transform (DCT2) [12], a widely utilized tool in various domains such as audio coding, texture analysis, image classification, and compression.

The DCT, akin to the discrete Fourier transform (DFT), measures the similarity between a signal and cosine waves of varying frequencies. Notably, the DCT is chosen for its real-to-real nature, practical utility, and potential for fewer artifacts compared to the complex-to-complex DFT when filtering.

Manipulating structure at different scales is facilitated by the DCT, allowing actions like low-pass filtering to smooth and retain the overall trend of the input, high-pass filtering to normalize each term with respect to its neighbors, neutralizing longer-term trends. Combining these operations yields a band-pass filter, permitting only a specific band of frequencies to pass through.

In essence the DCT is a reversible technique for breaking down a sequence of real values $x_0, ..., x_{N-1}$ (e.g. the values of an embedding dimension) into a weighted

sum of cosine waves with different frequencies. This representation is obtained by computing dot products between the signal and cosine waves, yielding coefficients in the frequency domain. The number of frequencies corresponds to the sequence length N, where the lowest frequency wave is a constant ($k = 0$), and the highest frequency wave completes one cycle every timestep $k = N - 1$.

Inverting the DCT (IDCT) using all X_n values will reconstruct the original sequence. However, setting weighting coefficients for certain k to 0 will result in a filtered version. For instance, zeroing out k values above a threshold retains lower frequencies, causing values to oscillate slowly. Conversely, eliminating k values below a threshold preserves higher frequencies, amplifying short-term changes.

2.2 Spectral approach for language representations

Studies have demonstrated that contextual word representations encapsulate not only token meanings but also a broad spectrum of linguistic aspects, such as semantic roles, entity types, constituent labels, relations between entities, and coreference. This implies that the encoded information spans scales from the subword level to the document level.

In examining deep language representations with the DCT Tamkin et al. [13] focus on sequences of contextual word representations in BERT [4]. They examine three English-language tasks, involving classification of word-, utterance-, and document-level phenomena, providing a natural testbed for investigating the content of these representations.

Their focus is on discerning whether these linguistic phenomena can be disentangled at the level of individual neurons by employing spectral filters. The filters aim to separate structural information at different scales within a neuron's activations across the input. The selection of spectral filters plays a crucial role, as it impacts the classifier's performance on tasks at different scales using the filtered representations.

Their goal is to strategically choose bands that align with different scales, where the scale of a frequency is determined by its period, indicating the number of tokens required to complete a full cycle.

For a given sequence of contextual word representations their approach involves applying the DCT to a specific neuron's slice. A visual represenation can be seen in lower part of Figure 2.1. To implement filters they zero out relevant values in the spectrum and apply the IDCT to restore the sequence to its original domain.

They segment the frequency spectrum into five distinct bands, a selection made to align with the inductive bias that linguistic units at a higher scale consist of multiple units from the scale below (e.g., phrases composed of several words).

Different spectral filters yield specialized representations tailored to the intended task. The highest probing accuracy for part of speech tagging is achieved when extracting the HIGH band, aligning with the word-level nature of the task. However, the highest frequency spectral band performs less effectively than the original representations, indicating that lower frequency information is sometimes crucial (e.g., for parts of speech correlated over several tokens, like strings of numbers or lists of nouns).

In contrast, topic-classification excels with information from the LOW band, consistent with its document-level nature. Intriguingly, the accuracy for the LOW band surpasses that of the original representations, suggesting that higher frequency variation in the original representations may be detrimental for this task.

Probing for dialog speech acts, a classification task over utterances, achieves optimal results at the MID band, with performance comparable to that of the original representations.

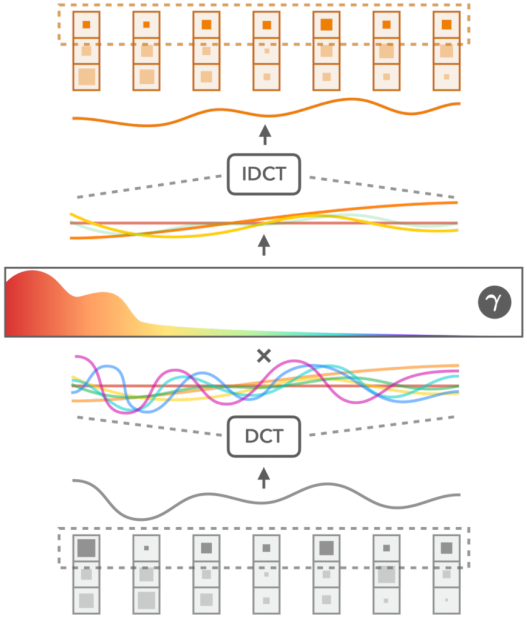

Figure 2.1. Visualization of Spectral Probing. The sequence of embedding values is first decomposed into composite frequency waves using DCT 2. Then the learned filter is applied retaining a subset of waves, for which IDCT returns the filtered sequence of values.

2.2.1 Differentiable spectral filters

In Müller-Eberstein's spectral probing framework [10], a differentiable approach is introduced, building upon Tamkin et al.'s work. This method autonomously discerns the pertinent spectral profiles within contextualized embeddings for various NLP tasks across diverse languages. The spectral filter, represented as a vector $\gamma \in R_N$, dynamically adjusts its weights for each frequency to capture the structural information at different scales. Higher weights on frequencies denote faster changes in neuron activation, while lower weights indicate more gradual shifts.

The filtering process involves sigmoid-scaled weights before inverting the Discrete Cosine Transform (DCT), with $\gamma(k) \in [0, 1]$ determining the retention or filtration of frequencies at index k. To adapt to varying sequence lengths, the spectral probe employs adaptive mean pooling. The lightweight parameter γ is integrated between the frozen encoder and probing head, leveraging the existing

training objective to collectively learn which frequencies to amplify or filter out.

The visual depiction of the filtering process is illustrated in Figure 2.1.

The experimental setup involves comparing spectral probing to fixed-band filters, replicating Tamkin et al.'s highest and lowest frequency experiments. Results demonstrate equivalent or higher accuracy for the spectral filter across tasks and languages, particularly benefiting sequence-level information-dependent tasks. The learned filter mirrors fixed-band results, emphasizing the importance of mid-low frequencies in topic information.

Müller-Eberstein's spectral probing not only confirms and refines frequency ranges from prior work but also unveils more detailed insights. It outperforms fixed-band filters, requiring only a single probing run without manual engineering, showcasing its effectiveness in identifying communicative levels across languages and tasks.

2.3 The Vision Transformer

The Vision Transformer (ViT), is based on the Transformer architecture [15], originally designed for sequential data. To adapt it for 2D images, the input image $x \in \mathbb{R}^{H \times W \times C}$ is reshaped into a sequence of flattened 2D patches $xp \in \mathbb{R}^{N \times (P \cdot C)}$, where H, W are the image dimensions, C is the number of channels, P is the patch resolution, and $N = \frac{HW}{P^2}$ is the number of patches. These patches serve as the effective input sequence length for the Transformer.

A constant latent vector size D is maintained throughout all layers of the Transformer. The patches are flattened and mapped to D dimensions using a trainable linear projection. A learnable embedding is then prepended to the sequence of embedded patches, serving as the image representation.

The visual depiction of this stages is illustrated in Figure 2.2.

Similar to BERT's [class] token, a classification head is attached to the output of the Transformer encoder during both pre-training and fine-tuning. Position embeddings are added to retain positional information, using standard learnable 1D position embeddings.

The Transformer encoder consists of alternating layers of multiheaded self-attention (MSA) and MLP blocks. Layernorm (LN) is applied before every block, and residual connections after every block. The MLP contains two layers with a GELU non-linearity.

The Vision Transformer has less image-specific inductive bias compared to CNNs. In CNNs, locality, two-dimensional neighborhood structure, and translation equivariance are embedded in each layer. In ViT, only MLP layers are local and translationally equivariant, while self-attention layers are global.

The two-dimensional neighborhood structure is used sparingly, mainly at the beginning of the model by cutting the image into patches and adjusting position embeddings for different resolutions during fine-tuning. Position embeddings at initialization carry no information about 2D positions, and spatial relations between patches are learned from scratch.

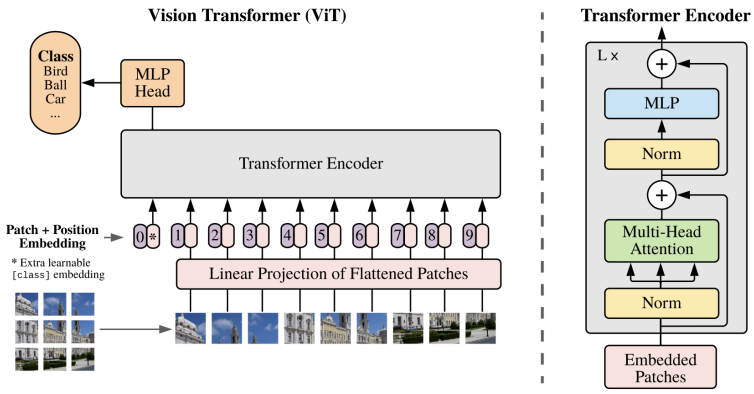

Figure 2.2. ViT model overview. The process involves partitioning an image into patches of fixed sizes. Each patch undergoes linear embedding, followed by the addition of position embeddings. The resulting sequence of vectors is then fed into a conventional Transformer encoder. For the purpose of classification, a standard technique is employed, which entails introducing an additional trainable "classification token" to the sequence.

2.3.1 Self-Supervised Learning in Vision Transformers

Deep learning has experienced a surge in architecture capabilities, with models demanding vast amounts of labeled images. Self-supervised pre-training, successfully employed in natural language processing (NLP), serves as a solution to address the data hunger.

The approach, exemplified by autoregressive language modeling in GPT [11] and masked autoencoding in BERT, involves removing a portion of data and learning to predict the removed content. These methods have enabled the training of generalizable NLP models with over one hundred billion parameters.

For visual representation learning a scalable form of a masked autoencoder (MAE) was intruduced by He et al. [6]. The MAE masks random patches from input images, reconstructing the missing patches in the pixel space. With an asymmetric encoder-decoder design, the encoder operates on the visible subset of patches, optimizing accuracy and reducing computation.

Following the paradigm set by ViT, the authors divide an image into regular non-overlapping patches. A subset of patches is sampled using a random sampling strategy without replacement. The MAE encoder, based on ViT, operates exclusively on visible, unmasked patches. The encoder embeds patches through a linear projection with added positional embeddings, followed by processing through a series of Transformer blocks. Notably, the encoder handles only a small fraction (e.g., 25%) of the full set, allowing for the training of large encoders with reduced compute and memory requirements.

In contrast to the encoder, the MAE decoder processes the full set of tokens, including encoded visible patches and mask tokens. Mask tokens, shared learned vectors indicating missing patches, play a crucial role. The decoder, designed in-

dependently of the encoder, features Transformer blocks and operates solely during pre-training for the image reconstruction task. Its asymmetrical design, narrower and shallower than the encoder, significantly reduces pre-training time.

The MAE reconstructs the input by predicting pixel values for each masked patch. The last layer of the decoder is a linear projection, reshaping the output to form a reconstructed image. The loss function computes the mean squared error (MSE) between the reconstructed and original images in the pixel space, focusing solely on masked patches. Additionally, a variant normalizes pixel values, improving representation quality in experiments.

A visual depiction of the model architecture is represented in Figure 2.3.

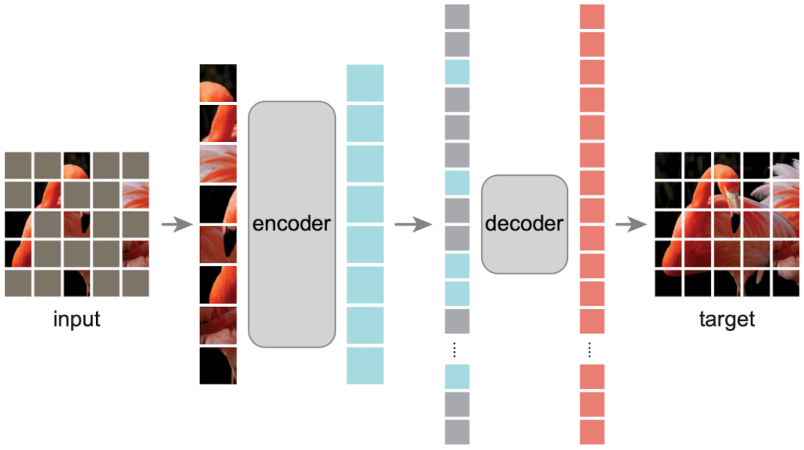

Figure 2.3. ViT-MAE model overview.

Chapter 3

Spectral Approach for Image Representations

3.1 Motivation and goals

Understanding the intricate patterns embedded in images is crucial for enhancing interpretability in image classification models. While traditional computer vision methods rely on feature extraction to discern structured patterns, this work takes the approach of using spectral analysis. The rationale behind employing spectral analysis is intuitive – it allows us to dissect and control information at different scales. By representing neuron activations as a weighted sum of cosine waves with varying frequencies as discussed in 2.1, we can discern the structure at each scale. Spectral filtering, achieved through the DCT, enables the removal of information about variation at specific scales. This process not only enhances interpretability but also facilitates the identification of scale-specific features crucial for image classification tasks.

As outlined in section 2.3, in contrast to traditional models that may carry specific priors, the Visual Transformer operates without explicit inductive biases tailored for image understanding. The absence of predefined biases allows for a more data-driven and adaptable approach, with the model learning relevant features and patterns directly from the input.

The ultimate objective of this study is to improve task-specific classification performance in image recognition. By disentangling scale-specific information in existing embeddings using spectral filters, we aim to guide the model in learning more about scales relevant to the task at hand. Understanding which frequencies are pertinent to a given classification task enables the model to be trained more effectively, enhancing its ability to discern and utilize these scale-specific features and, consequently, improving overall performance.

3.2 Methodology

The project method consists in the application of spectral analysis to image representations, guided by Tamkin [13] and Müller-Eberstein [10] approach. Building

upon their work, we extended the application of spectral filters to the activations of individual neurons in the Visual Transformer.

We utilize a pre-trained ViT-MAE model [6] to produce contextual representations. Spectral filters are then applied to these representations along each dimension, paving the way for a comprehensive evaluation through probing experiments. The evaluation involves encoding each training example, applying spectral filtering, and utilizing a linear classifier for the image classification tasks.

The key innovation in our research is the prism layer. The prism layer, as introduced by Tamkin et al. allows the application of spectral filters to the activations of individual neurons in the Visual Transformer. This layer specializes neurons in the model for particular scales of structure. The prism layer serves as a the tool in disentangling scale-specific information in existing embeddings, enabling a more nuanced exploration of the frequencies relevant to image classification tasks.

3.3 Model design

3.3.1 The base model

ViT-MAE choice is motivated by some of it's key similarities with BERT. Both follow the original transformer implementation as closely as possible, and both pre-trained the encoder by masking random patches of the input and predicting the missing content. There are some key differences however.

Differences and Analogies with reference work

Masking Ratio. The choice of masking ratio plays a pivotal role in both ViT-MAE and BERT, influencing the self-supervised learning objective during pre-training. While BERT adopts a masking ratio of 15%, ViT-MAE takes a divergent approach with a substantially higher masking ratio of 75%. This difference implies that ViT-MAE faces a more challenging self-supervisory task, as a larger proportion of input patches are randomly masked, demanding a nuanced understanding of the relationships between visible and masked patches.

While the principles align, the application differs due to the nature of the input data. BERT focuses on masking individual tokens in a sequence, maintaining coherence within the linguistic context.

Encoder-Decoder Design. The ViT-MAE and BERT models share a common ancestry in the original Transformer architecture, leveraging self-attention mechanisms for capturing contextual dependencies. However, a notable departure lies in their respective encoder-decoder designs. BERT adheres to the traditional transformer design, with a single encoder handling masked input patches. In contrast, ViT-MAE opts for an asymmetric encoder-decoder setup as discussed in Section 2.3.1. The encoder processes only visible, unmasked patches, contributing to efficiency in training large encoders. Meanwhile, the decoder, designed independently and narrower than the encoder, operates exclusively during pre-training for image reconstruction. This design choice introduces asymmetry but reduces pre-training time, aligning with the goal of ViT-MAE.

Embedding Structure The structure of token embeddings in ViT-MAE and BERT exhibits disparities tied to their respective pre-training objectives and input modalities.

BERT constructs token embeddings by summing the token, a segmentation embedding, and a position embedding. See Figure 3.1 for a visual depiction.

Notably, ViT-MAE deviates from this approach by omitting a segment embedding. This distinction arises from the inherent differences in contextual information preservation objectives. While BERT, being language-focused, may benefit from segment embeddings to distinguish different linguistic segments, ViT-MAE's omission aligns with its image-centric focus, where the emphasis lies on understanding relationships within a single modality.

More importantly, both models construct contextualized embeddings, however, there is a key difference in their "quality". BERT embeds documents composed of words, which are semantically distinct elements, with the tokenizer. ViT, on the other hand, embeds images composed of patches, which are arbitrary distinct elements, through a linear projection. This key difference is visually represented in the two respective Figures 3.1 and 2.2. We believe that this distinction played a crucial role in the subsequent tasks.

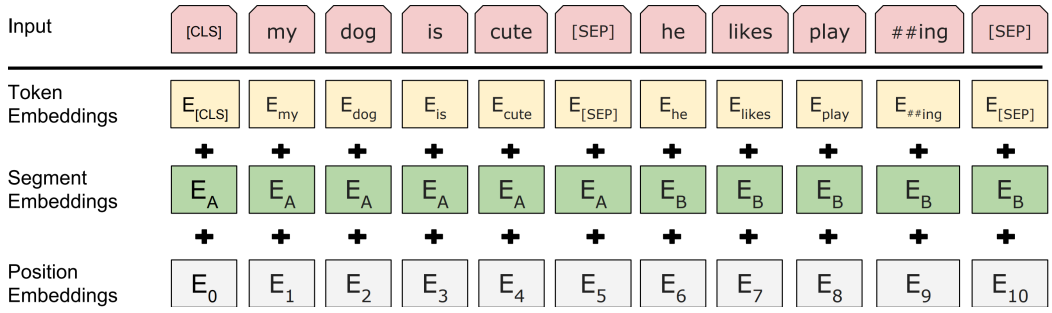

Figure 3.1. BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

3.3.2 The Prism Layer

The final hidden state of the ViT-MAE model is fed to the Prism Layer introduced by Tamkin et al. A conceptual represenation of the prism layer is depicted in Figure 3.2.

The spectral filter offers three distinct modes:

1. **Filter Bypass Mode:** No filtering is performed, allowing an assessment of the original contextual representations.

2. **Auto-Filter Mode:** The filter dynamically adjusts its weights based on the existing training objective, jointly learning which frequencies to amplify or filter out. This mode aims to optimize the model's performance on the classification task. We use an external PyTorch library[1] for computing and backpropagating through the DCT and IDCT.

[1] https://github.com/zh217/torch-dct

3. **Band-Filter Mode:** Indicated frequency bands are selectively removed from the contextual representations. This mode provides insights into how specific frequency ranges contribute to the overall understanding of the image.

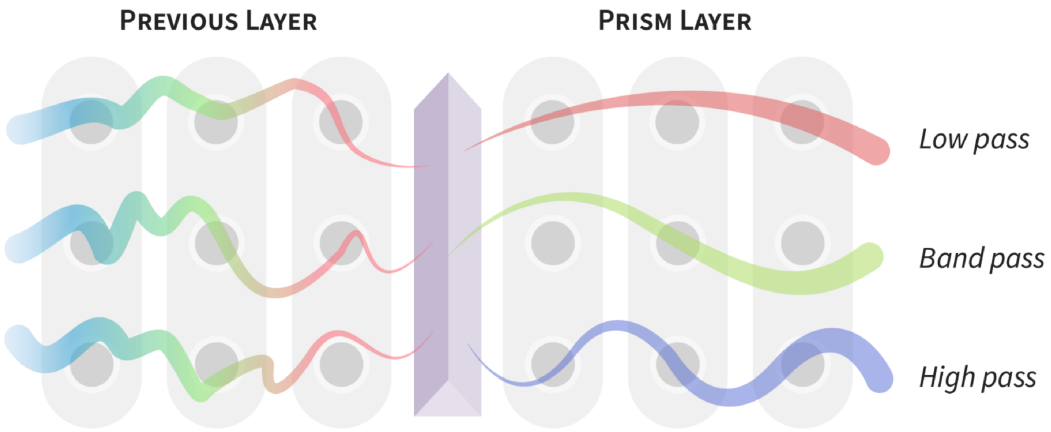

Figure 3.2. Conceptual depiction of the prism layer. First, the representations for an input are computed (left; in this case, the input is of length three). Next, a spectral filter (a low-, high-, or band-pass) is applied along the activations of each individual neuron (right). This produces neurons that are only able to represent structure at particular scales. Curved lines illustrate the scales at which neurons can change over an input.

3.3.3 The classification head

The classification head in the PrismViT model is responsible for transforming the filtered output embeddings of the Visual Transformer into predictions for image classification. This is achieved through a linear layer that takes as input the filtered embeddings and applies a transformation to produce logits for each class. The number of output units in this layer is determined by the number of classes in the classification task.

Chapter 4

Experiments

The experimental setup involves training a linear probe on top of a frozen ViT-MAE encoder, similar to [1]. We perform probing experiments in a set of different configurations to understand the impact of spectral filtering on the model's ability to classify images.

This study is based on insights derived from foundational research highlighting the importance of low-frequency components in embeddings. In the realm of image classification, our attention is directed towards subjects with minimal variability in input images. Essentially, we aim to leverage the structured patterns present in subjects with low variability by focusing on the model's low-frequency components. The underlying concept is that by emphasizing low-frequency information in embeddings, especially for less variable subjects, we can enhance the model's proficiency in accurately discerning and classifying images.

4.1 Data

Image classification experiments involve two distinct datasets, each featuring different types of subjects. One dataset displays minimal variability in input images, while the other showcases structured patterns within the input images.

4.1.1 COCO Backgrounds

To investigate the pivotal role of low-frequency components in contextual representations for image classification, especially in subjects with minimal variability, the choice of data is paramount.

In the absence of a public dataset meeting the specific requirements of this study, a tailored dataset is constructed. This dataset is an adaptation of the COCO-Stuff datset [3], which augments all 164K images of the COCO 2017 [9] dataset with pixel-wise annotations for 91 stuff classes. Stuff classes are divided into outdoor and indoor, each further divided into 15 super-categories (e.g. floor, plant), and finally into leaflevel classes (e.g. marble floor, grass).

The COCO Backgrounds dataset preparation involves a 3 step approach to ensure meaningful labeling:

1. For each image patch, compute the normalized patch area by dividing it by the image area.
2. The image is retained in the dataset if:
    - The patch label corresponds to a "stuff" object.
    - The normalized patch area surpasses a predefined threshold.
    - There is only one qualifying patch meeting the above conditions.
3. Label the image with the "stuff" identifier based on its super-category.

A series of experiments are conducted, where the threshold is set heuristically to the following values: 33%, 50%, and 75%. Respectively, this removed from the COCO Stuff dataset: 92%, 71%, 48% of images. Altough this reduction in size is significant, the amount of data is sufficint to perform probing experiments.

For image examples please refer to 4.1.1 section.

This curation process results in the creation of the COCO Backgrounds dataset, specifically tailored to the study's requirements. It provides a unique and suitable foundation for investigating the role of low-frequency components in contextual representations for image classification tasks, in the context of subjects with minimal variability.

COCO Backgrounds samples

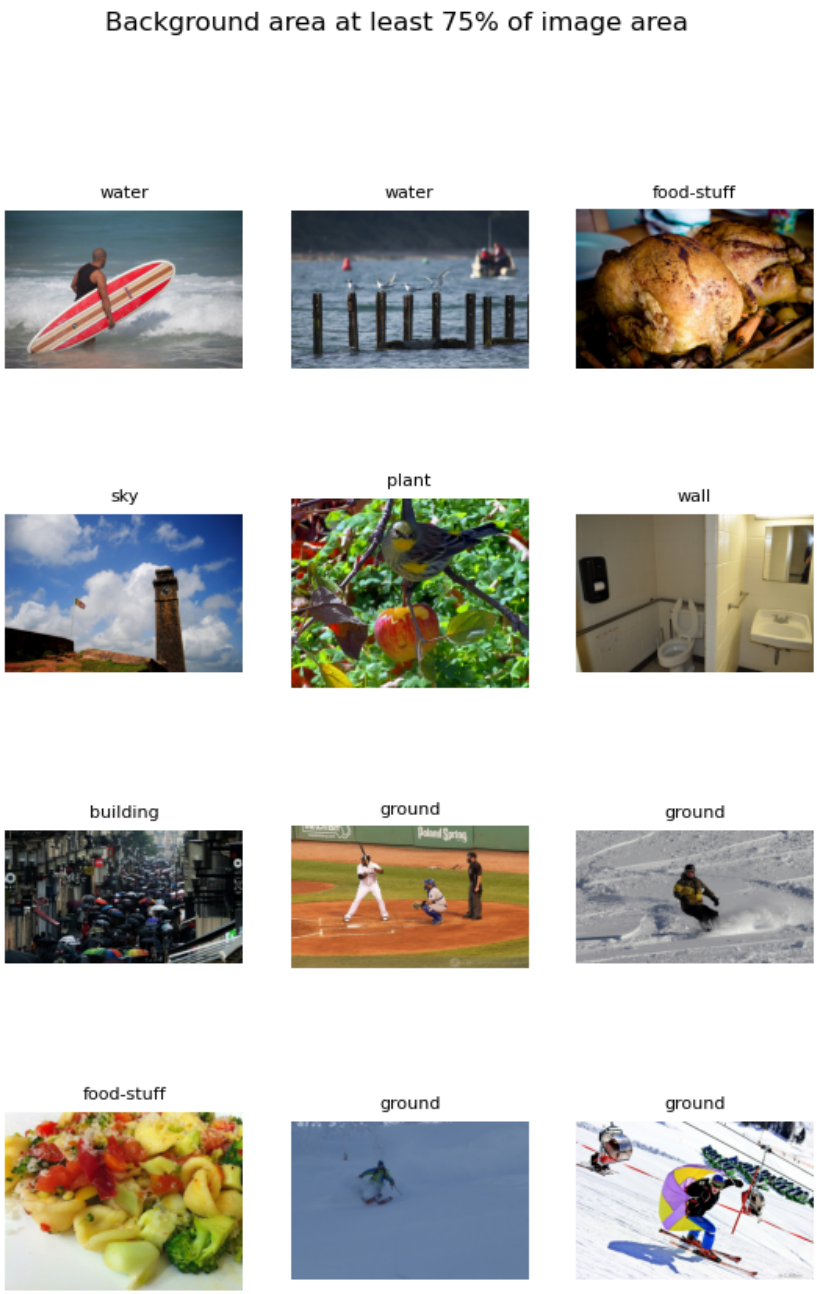

Figure 4.1. More samples from the COCO Background dataset training set with the threshold set to 75%. The image title indicates the corresponding label.

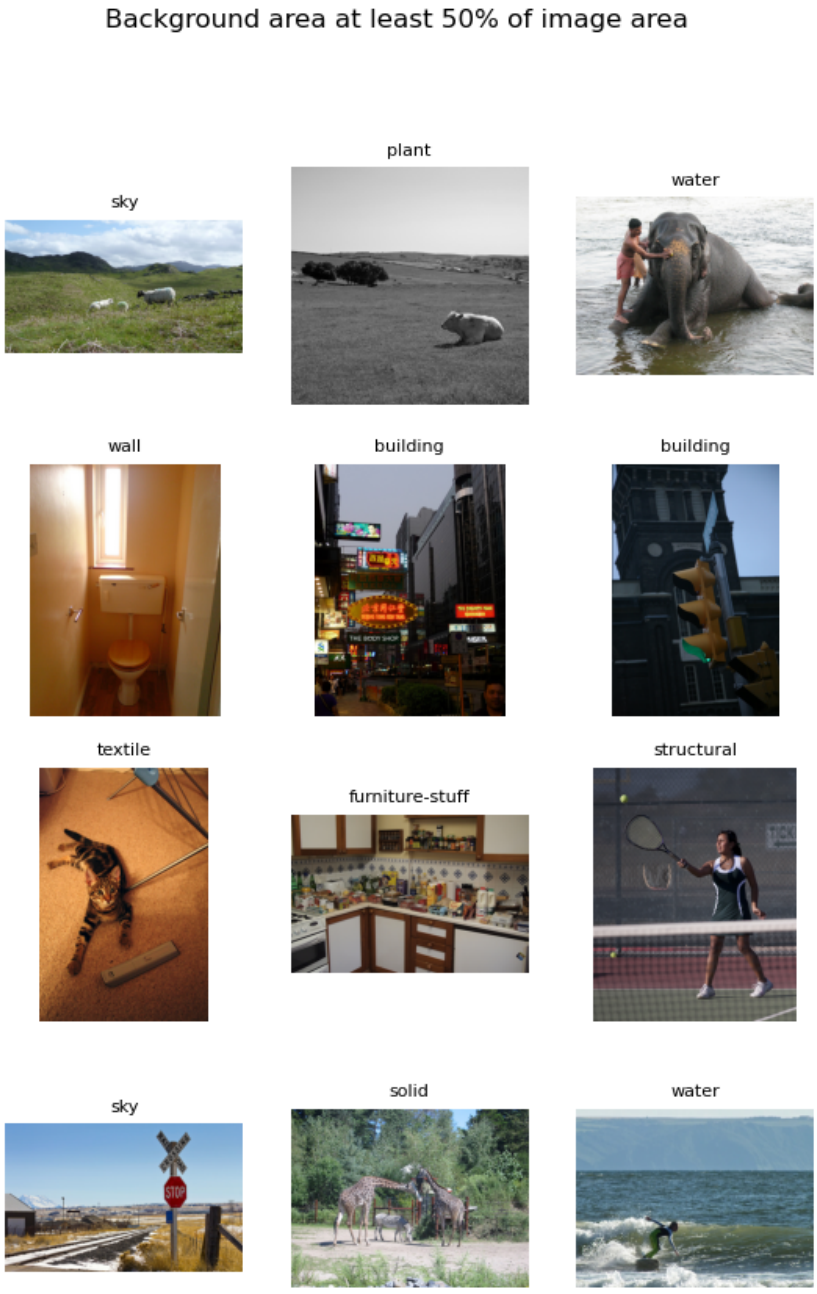

Figure 4.2. Samples from the COCO Background dataset training set with the threshold set to 50%. The image title indicates the corresponding label.

Figure 4.3. Samples from the COCO Background dataset training set with the threshold set to 33%. The image title indicates the corresponding label.

4.1.2 Food 101

To assess the validity of the analysis we test the complementary nature of the previous approach with the Food 101 dataset [2].

The Food 101 dataset is selected based on its subjects exhibiting structured patterns, aligning with our overarching goal of understanding scale-specific information in image embeddings. This dataset encompasses a diverse range of food items, each characterized by distinct visual textures and arrangements. It consists of 101 food categories, with 101'000 images. For each class, 250 manually reviewed test images are provided as well as 750 training images.

A critical consideration in dataset selection is the level of background variability. To ensure the model's classification isn't influenced by background cues, this dataset is selected because it has minimal background variability. This minimizes the reliance on background information, forcing the model to focus on the intrinsic patterns of the food items for accurate classification.

The aim is to uncover how disentangling scale-specific information, particularly focusing on high-frequency components, influences the model's performance in classifying food items with distinct visual structures.

For image examples please refer to Section 4.1.2.

Food 101 samples

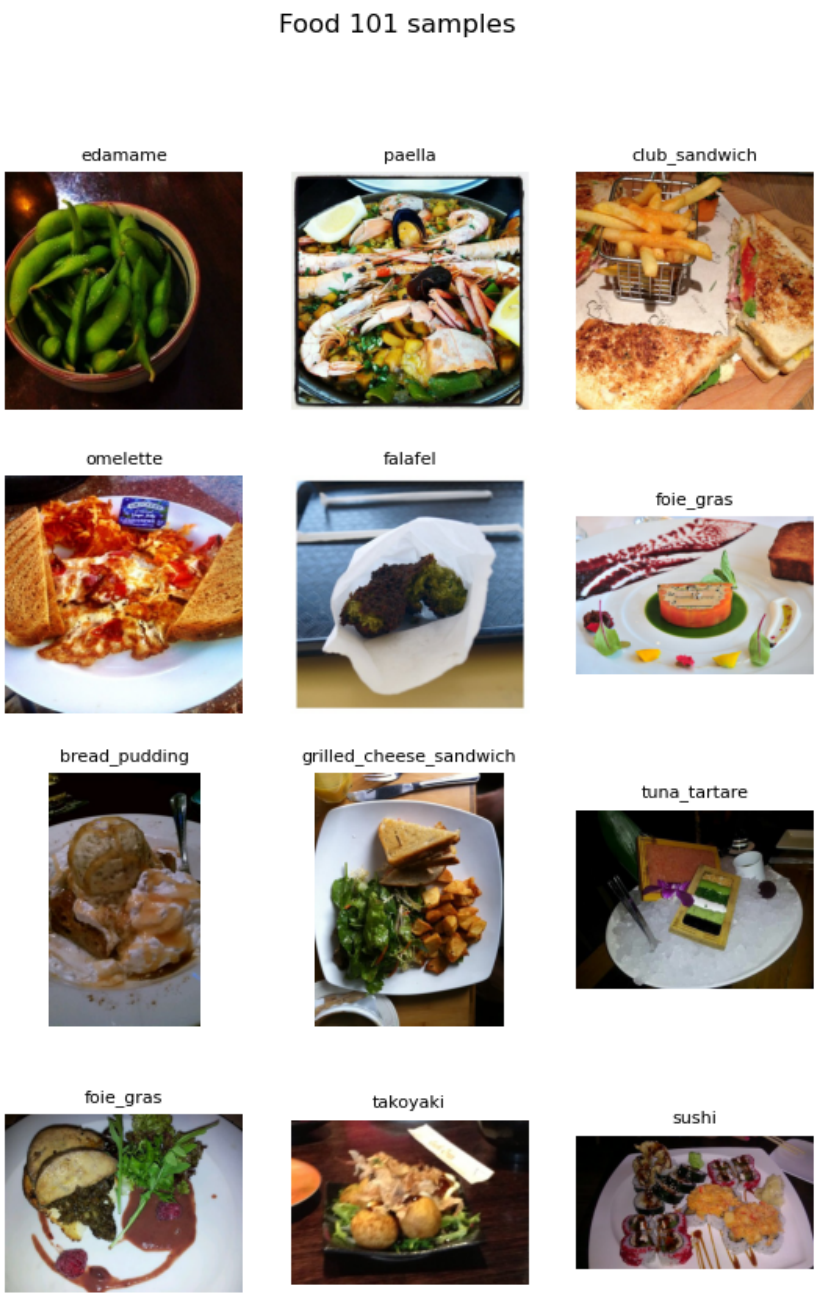

Figure 4.4. Samples from the Food 101 dataset training set. The image title indicates the corresponding label.

4.2 Setup

4.2.1 Model size

To explore the impact of model size on the spectral analysis of Visual Transformers, two distinct model variants were employed: ViT-MAE Base and ViT-MAE Huge. The ViT-MAE Base variant, with 86 million parameters and an embedding size of 196, represents the baseline configuration. The ViT-MAE Huge variant has a substantial increase in both parameters and embedding size, with 632 million parameters and an embedding size of 256. This larger-scale model enables a more discretized alaysis of the frequency spectrum due to the expanded embedding size. The rationale behind incorporating ViT-MAE Huge lies in the potential for capturing finer-grained details and variations in the visual input, given the increased capacity of the model.

The use of two model variants aims to verify the coherence of experimental findings under a two model sizes.

4.2.2 Filter configuration

For all three filter configurations – Filter Bypass, Auto-Filter, and Band-Pass – experiments were conducted.

The Filter Bypass mode serves as a baseline for the experiment, representing unfiltered linear probing. In this mode, no spectral filtering is applied to the activations of individual neurons.

The auto-filter mode introduces a dynamic approach to spectral filtering, providing tailored sigmoid-scaled weighting of each frequency for the image classification task. This mode aims to adaptively adjust filter weights, allowing the model to focus on frequencies that contribute the most to the classification task.

The band-pass mode enables the use of only a subset of frequencies for image classification. This configuration verifies whether a specific subset of frequencies is both suitable and sufficient for the image classification task. By restricting the model to operate within a defined frequency band, this mode explores the potential efficiency of utilizing a focused range of information.

4.2.3 Hyper parameter choice

We identified the optimal hyperparameters, as shown in Table 4.1, by aligning with the choices outlined in [10] and [6]. With slight adjustments, we found that adopting He's hyperparameter choices yielded more favorable results. To maintain a fair comparison across experiments, the hyperparameter configuration remained constant.

Hyperparameter	Value
Optimizer	SGD
Learning Rate	0.1
Momentum	0.9
Weight Decay	0
Train Batch Size per Device	16
Eval Batch Size per Device	16
Number of Training Epochs	30
LR Scheduler Type	Cosine
Warmup Ratio	0.1

Table 4.1. Adopted hyperparameter values

4.3 Results

4.3.1 Probing accuracy analysis

To evaluate the effectiveness of the approach, we delve into the probing accuracy for each model size, considering various filter configurations. This analysis aims to discern the most impactful frequency-related features for image classification.

Some highlighted experiments results are showcased below. For the complete and interactive version please refer to the thesis project repository.

The auto-filter mode, while not leading to significant improvements, exhibits a slight performance boost. This marginal enhancement is attributed to the increased capacity introduced by this filtering mode. The only instances where this doesn't hold true is depicted in Figure 4.10 AND 4.8. This models are the VIT-MAE Huge trained on the COCO Background dataset with threshold set to 50% and 75%. However, this particular cases are not deemed significant enough to lead to different conclusions.

Conversely, using four equally allocated bands results in diminished performance. This scenario mirrors the linear probing approach, as the network's capacity remains the same. Again, there is only one instance where this doesn't hold true, it is for the VIT-MAE Base model trained on the COCO Background dataset with threshold set to 75%. This case is depicted in Figure 4.9. As before, this particular case isn't deemed significant enough to lead to different conclusions.

This observations hold true across various hyperparameter configurations and model sizes, underscoring the robustness of the findings.

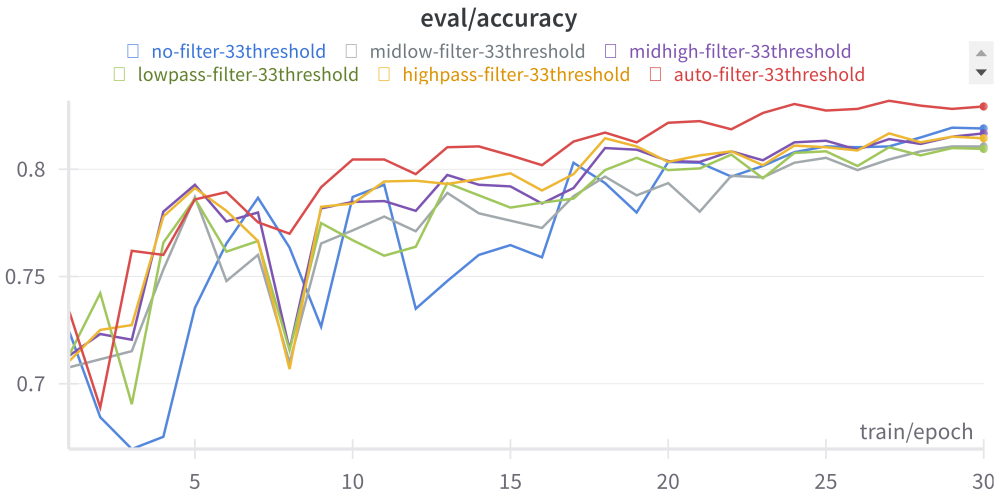

Figure 4.5. Evaluation accuracy of the VIT-MAE Base model trained on the COCO Background dataset for all filter configurations. COCO Background threshold set to 33%.

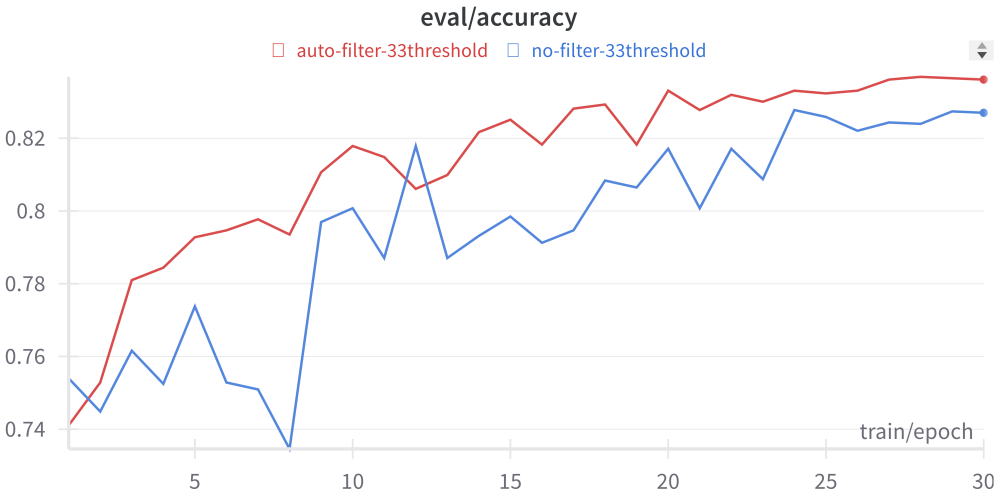

Figure 4.6. Evaluation accuracy of the VIT-MAE Huge model trained on the COCO Background dataset for all filter configurations. COCO Background threshold set to 33%.

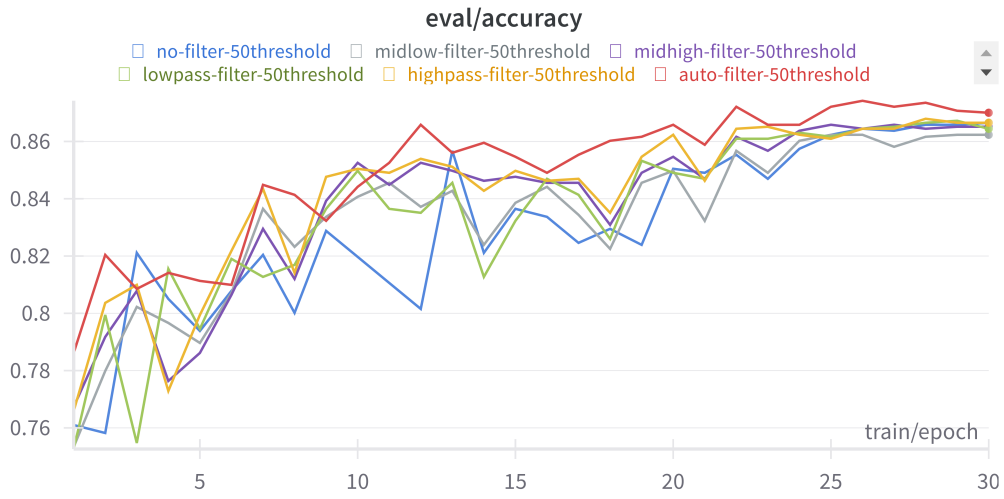

Figure 4.7. Evaluation accuracy of the VIT-MAE Base model trained on the COCO Background dataset for all filter configurations. COCO Background threshold set to 50%.

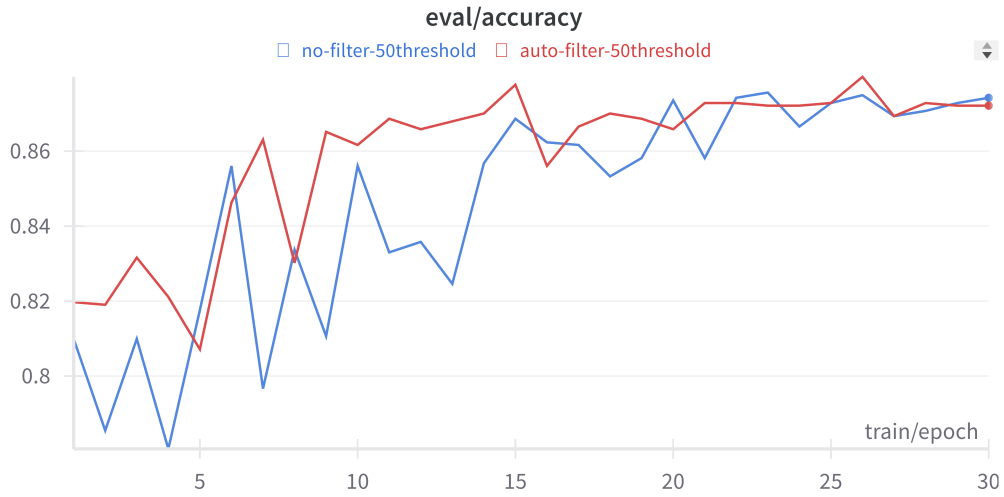

Figure 4.8. Evaluation accuracy of the VIT-MAE Huge model trained on the COCO Background dataset for all filter configurations. COCO Background threshold set to 50%.

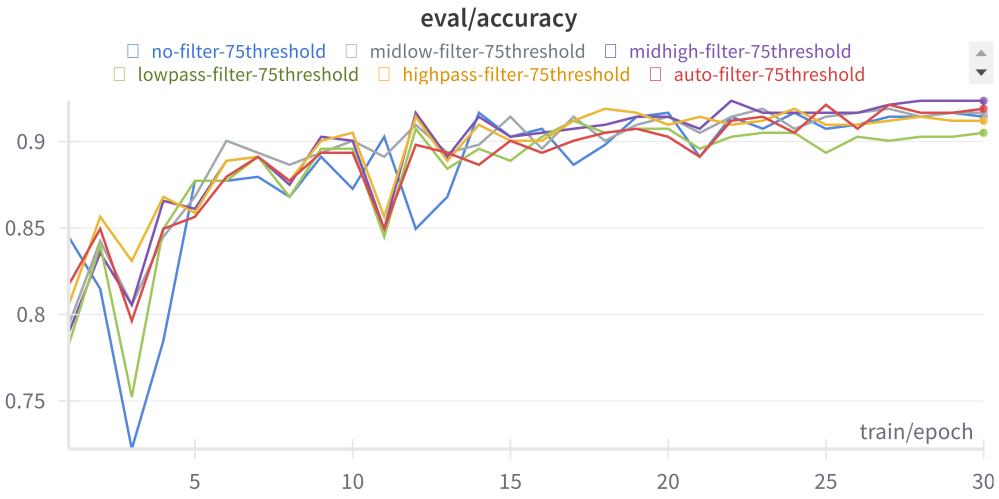

Figure 4.9. Evaluation accuracy of the VIT-MAE Base model trained on the COCO Background dataset for all filter configurations. COCO Background threshold set to 75%.

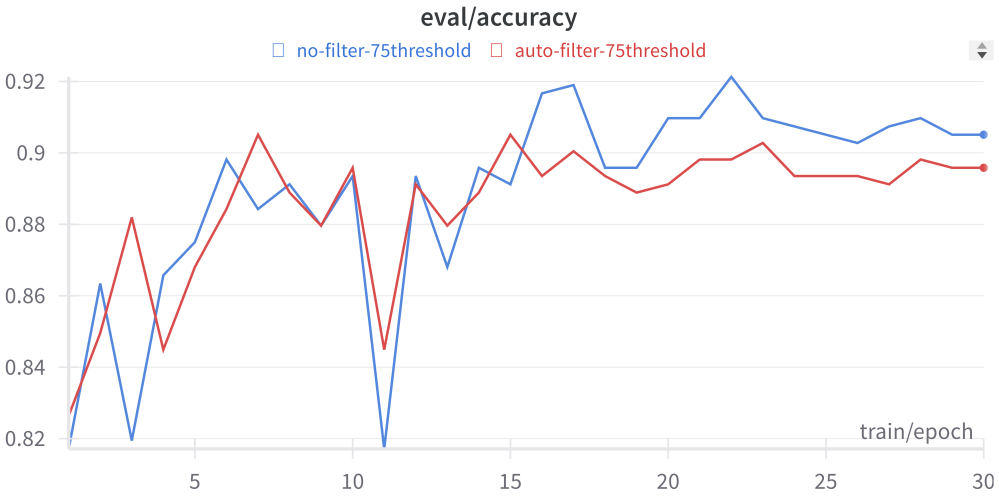

Figure 4.10. Evaluation accuracy of the VIT-MAE Huge model trained on the COCO Background dataset for all filter configurations. COCO Background threshold set to 75%.

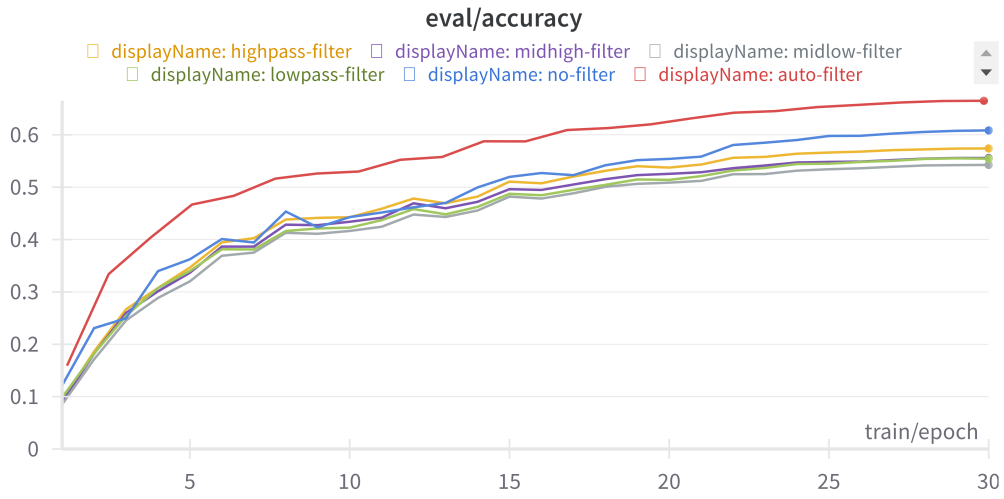

Figure 4.11. Evaluation accuracy of the VIT-MAE Base model trained on the Food 101 dataset for all the tested filter configurations.

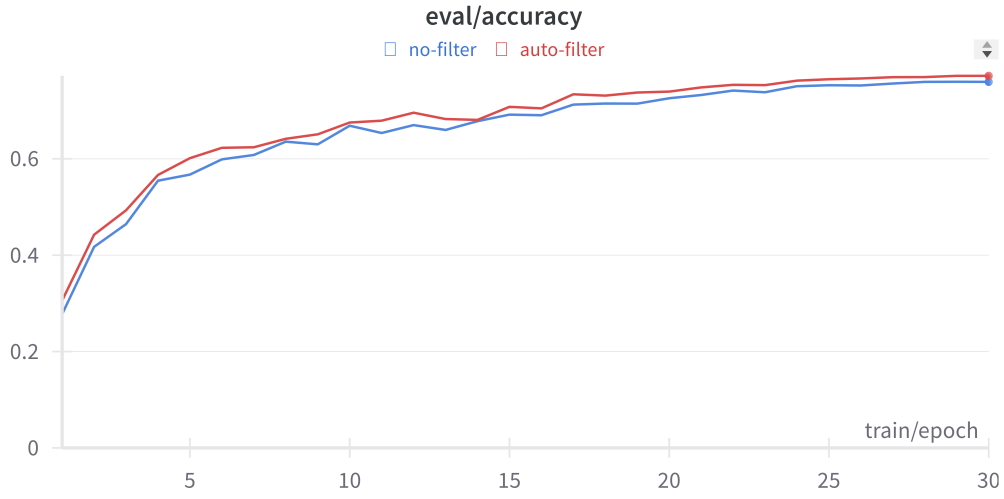

Figure 4.12. Evaluation accuracy of the VIT-MAE Huge model trained on the Food 101 dataset for all the tested filter configurations.

4.3.2 Inspecting Filter Weights

An examination of filter weights did not reveal a clear contribution by distinct frequency bands. Despite variations in thresholds and model sizes, the evidence from spectral profiling remains inconclusive regarding the explicit influence of individual frequency bands on the model's performance.

When the COCO Background threshold is set to 33%, the filter weights show a greater magnitude for high frequencies. However this result does not maintain consistency across different thresholds and model sizes. More importantly there is no evidence of better performance from the respective measurments represented in Figures 4.5 and 4.6.

The lack of a discernible pattern in filter weights persists across different thresholds and model sizes, indicating that the spectral profiling outcomes are coherent and not heavily dependent on specific experimental configurations.

All the filter weight histograms produced in the experiments are showcased below. For the interactive version please refer to the thesis project repository.

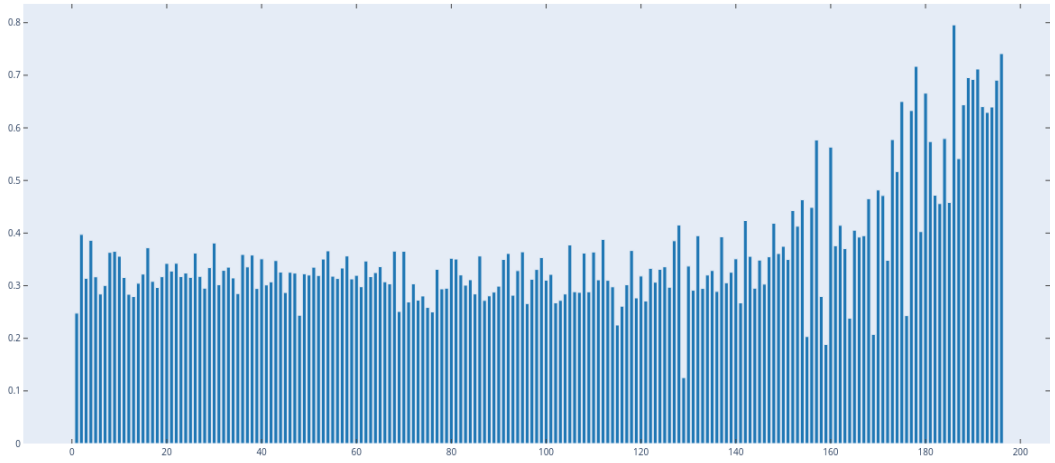

Figure 4.13. Filter weights of the VIT-MAE Base model trained on the COCO Background dataset for the auto-filter configuration. COCO Background threshold set to 33%. The x-axis refers to the frequency index.

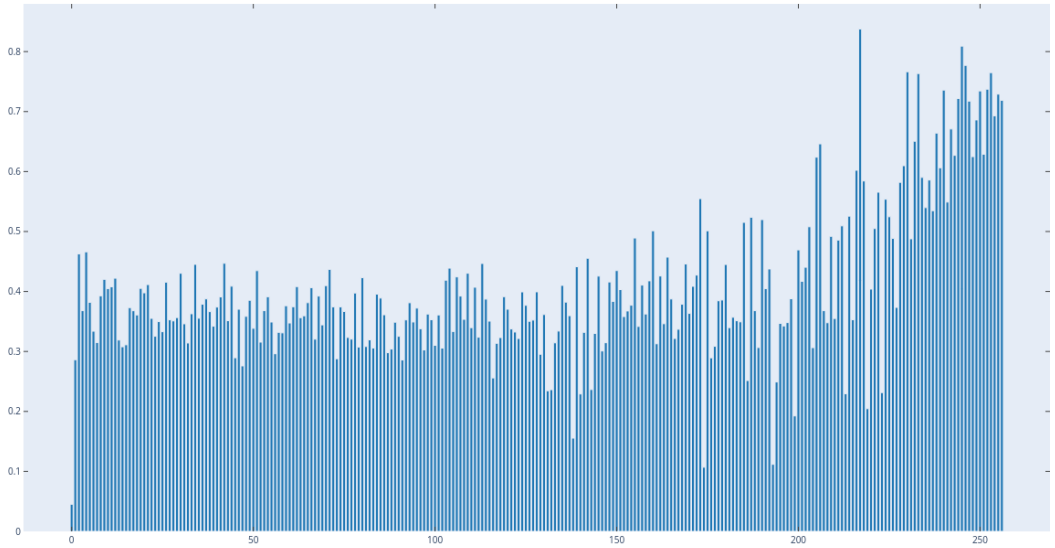

Figure 4.14. Filter weights of the VIT-MAE Huge model trained on the COCO Background dataset for the auto-filter configuration. COCO Background threshold set to 33%. The x-axis refers to the frequency index.

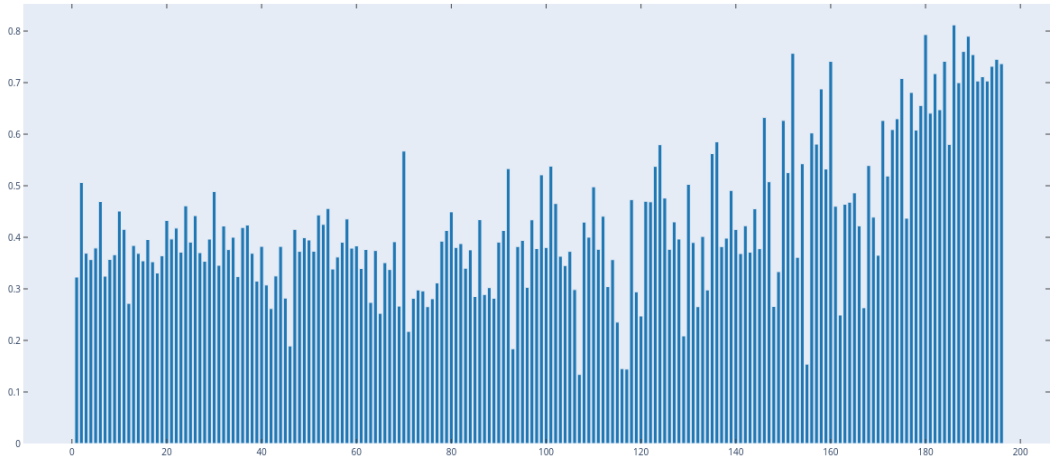

Figure 4.15. Filter weights of the VIT-MAE Base model trained on the COCO Background dataset for the auto-filter configuration. COCO Background threshold set to 50%. The x-axis refers to the frequency index.

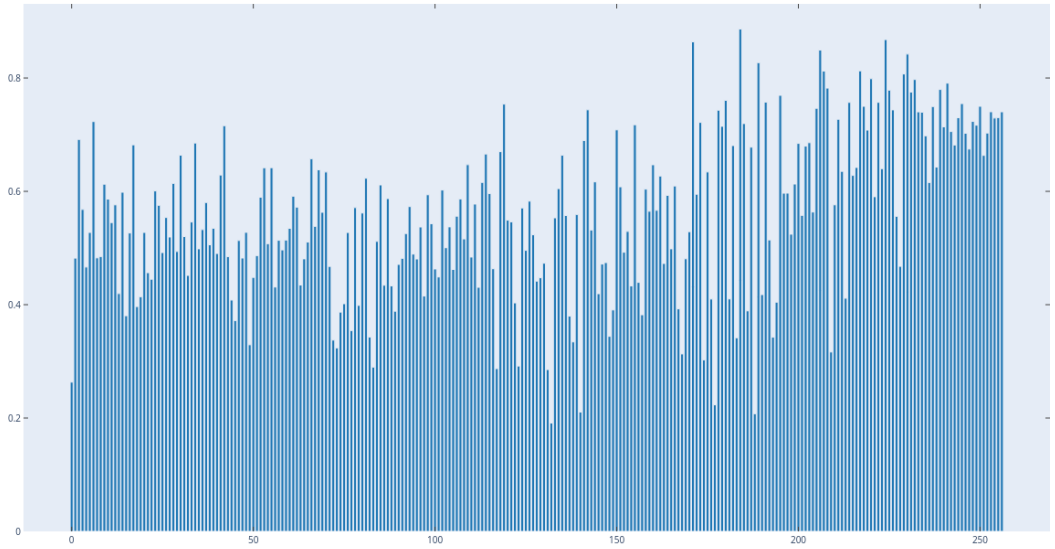

Figure 4.16. Filter weights of the VIT-MAE Huge model trained on the COCO Background dataset for the auto-filter configuration. COCO Background threshold set to 50%. The x-axis refers to the frequency index.

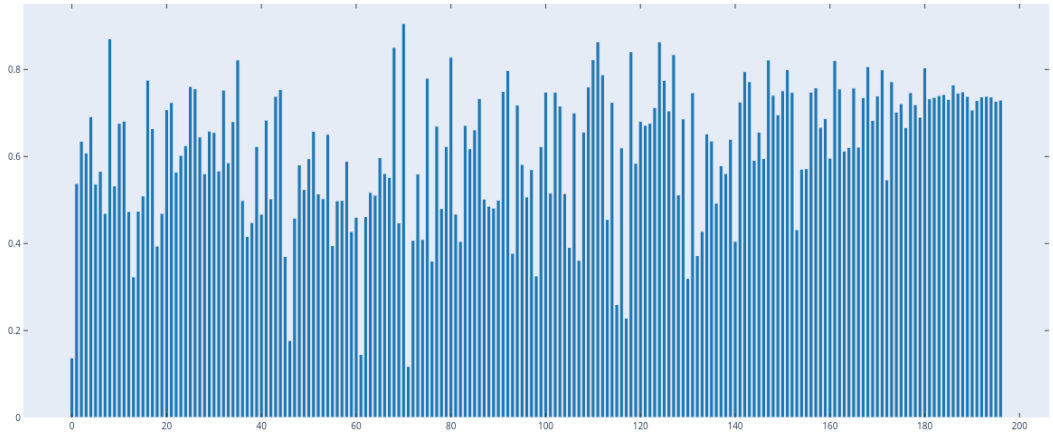

Figure 4.17. Filter weights of the VIT-MAE Base model trained on the COCO Background dataset for the auto-filter configuration. COCO Background threshold set to 75%. The x-axis refers to the frequency index.

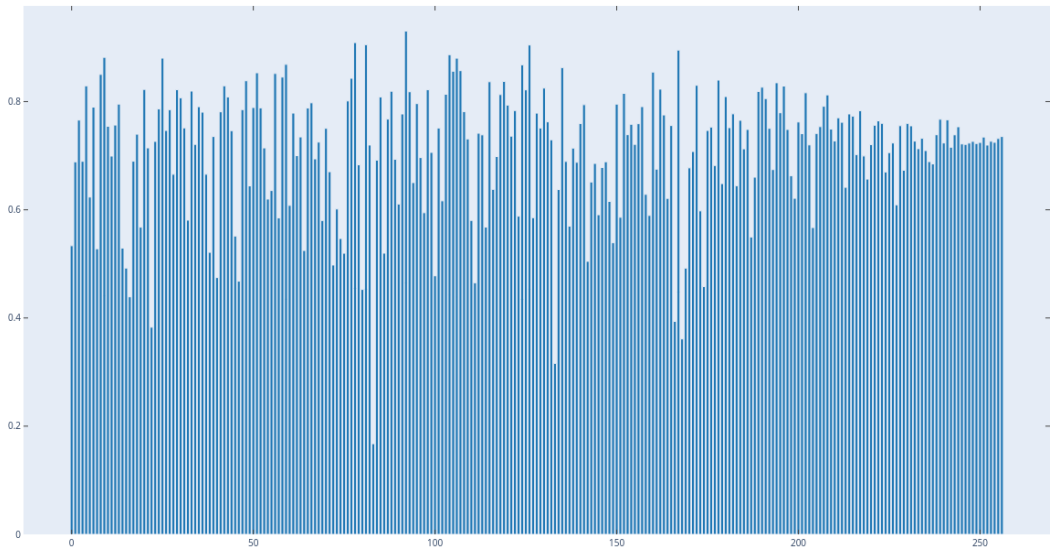

Figure 4.18. Filter weights of the VIT-MAE Huge model trained on the COCO Background dataset for the auto-filter configuration. COCO Background threshold set to 75%. The x-axis refers to the frequency index.

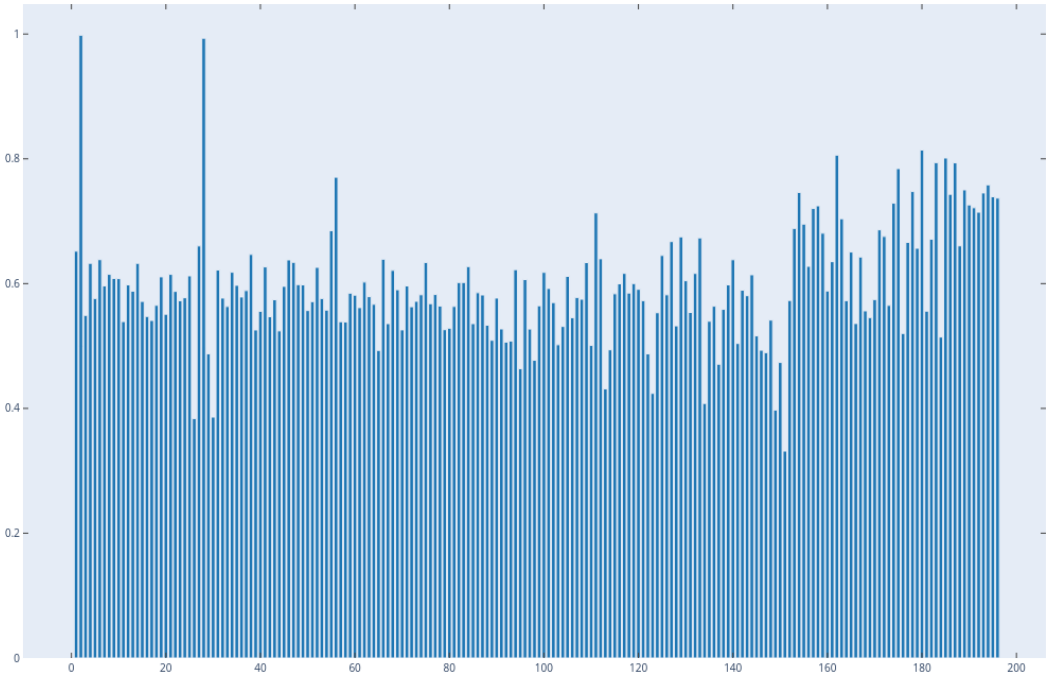

Figure 4.19. Filter weights of the VIT-MAE Base model trained on the Food 101 dataset for the auto-filter configuration. The x-axis refers to the frequency index.

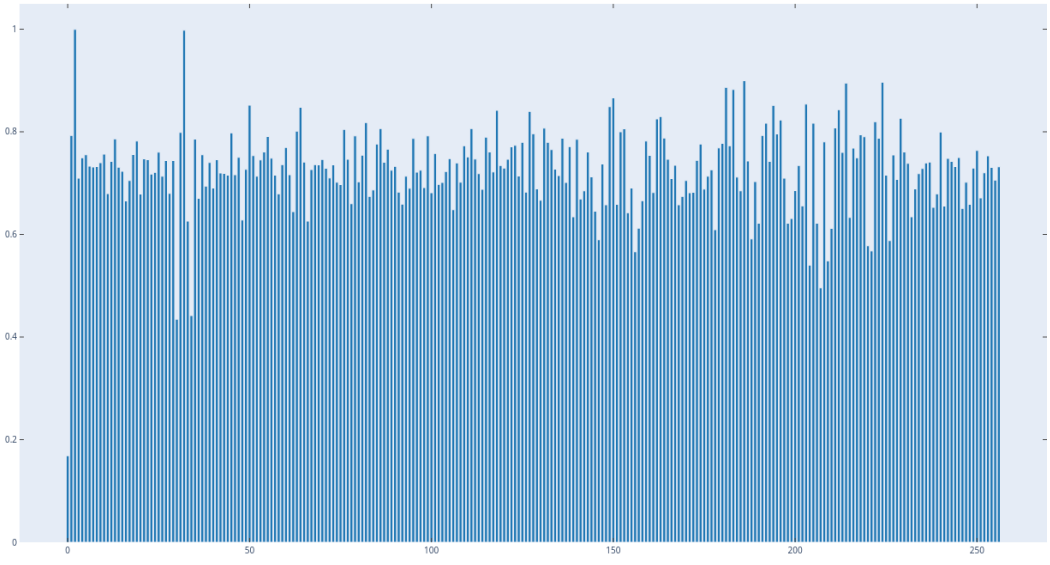

Figure 4.20. Filter weights of the VIT-MAE Huge model trained on the Food 101 dataset for the auto-filter configuration. The x-axis refers to the frequency index.

Chapter 5

Conclusions

One potential explanation for the observed phenomena in the COCO Backgrounds dataset lies in the structure variability of the image subjects. A portion of the dataset contains subject with noticable structure variability. We did not find however a better solution in choosing the image subjects for sucha a classification task.

Another potential explanation for the observed phenomena lies in the "quality" of contextualized embeddings as discussed in Section 3.3.1. This aspect prompts us to delve deeper into the interplay between contextualized embeddings and the successful disentanglement of scale-specific information.

Analogous to NLP, where phrases consist of semantically distinct words, we suggest viewing images as compositions of semantically distinct subjects. Departing from arbitrary square patches, we propose representing images through sets of bounding boxes identified in image annotations. This approach allows the embedding of semantically distinct elements separately, mirroring the structure found in natural language.

However, this innovative training approach introduces challenges, particularly in handling varying patch sizes. In addition, similar to BERT's strategy for handling different sentence lengths with padding, a BERT-like approach is needed to address varying patch numbers.

Future research could explore this proposed training approach.

Bibliography

[1] Guillaume Alain and Yoshua Bengio. *Understanding intermediate layers using linear classifier probes*. 2018. arXiv: 1610.01644 [stat.ML].

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. "Food-101 – Mining Discriminative Components with Random Forests". In: *European Conference on Computer Vision*. 2014.

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. *COCO-Stuff: Thing and Stuff Classes in Context*. 2018. arXiv: 1612.03716 [cs.CV].

[4] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[5] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].

[6] Kaiming He et al. *Masked Autoencoders Are Scalable Vision Learners*. 2021. arXiv: 2111.06377 [cs.CV].

[7] John Hewitt and Percy Liang. "Designing and Interpreting Probes with Control Tasks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2733–2743. DOI: 10.18653/v1/D19-1275. URL: https://aclanthology.org/D19-1275.

[8] Geoffrey E. Hinton et al. *Improving neural networks by preventing co-adaptation of feature detectors*. 2012. arXiv: 1207.0580 [cs.NE].

[9] Tsung-Yi Lin et al. *Microsoft COCO: Common Objects in Context*. 2015. arXiv: 1405.0312 [cs.CV].

[10] Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. *Spectral Probing*. 2022. arXiv: 2210.11860 [cs.CL].

[11] Alec Radford et al. "Improving language understanding with unsupervised learning". In: (2018).

[12] K.R. Rao and P. Yip. *Discrete Cosine Transform: Algorithms, Advantages, Applications*. Elsevier Science, 2014. ISBN: 9780080925349. URL: https://books.google.it/books?id=fWviBQAAQBAJ.

[13] Alex Tamkin, Dan Jurafsky, and Noah Goodman. *Language Through a Prism: A Spectral Approach for Multiscale Language Representations*. 2020. arXiv: 2011.04823 [cs.CL].

[14] Ian Tenney, Dipanjan Das, and Ellie Pavlick. "BERT Rediscovers the Classical NLP Pipeline". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4593–4601. DOI: 10.18653/v1/P19-1452. URL: https://aclanthology.org/P19-1452.

[15] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL].