Exploring Compositionality in Vision Transformers using Wavelet Representations

Anonymous Author(s)
Affiliation
Address
email

Abstract

While insights into the workings of the transformer model have largely emerged by analysing their behaviour on language tasks, this work investigates the representations learnt by the Vision Transformer (ViT) encoder through the lens of compositionality. We introduce a framework - analogous to the compositionality setting proposed for representation learning in Andreas (2019) - to test for compositionality in the ViT encoder. Crucial to drawing this analogy is the Discrete Wavelet Transform (DWT), which is a simple yet effective tool for getting input-dependent primitives in the vision setting. By examining the ability of composed representations to reproduce original image representations, we empirically test the extent to which compositionality is respected in the representation space. Our findings show that primitives from a one-level DWT decomposition produce encoder representations that approximately compose in latent space, offering a new perspective on how ViTs structure information.

1 Introduction

Vision Transformers (ViTs), in their supervised (Dosovitskiy et al., 2021), self-supervised (Caron et al., 2021), and unsupervised (He et al., 2022) variants, have delivered state-of-the-art performance across various computer vision applications. Image classification (Dosovitskiy et al., 2021), object detection (Li et al., 2022), semantic segmentation (Strudel et al., 2021), and image captioning and generation (Radford et al., 2021) are a few examples. ViTs leverage the transformer architecture - originally popularized in natural language processing (NLP) tasks - to process images using self-attention. This utilisation of transformer architecture in computer vision has opened new avenues for understanding and processing visual data.

It is natural to wonder why ViTs deliver such performance despite their origins in language models. Given their prevalence as backbones for generating image embeddings for various downstream tasks, we focus our investigation on the representations themselves. Several works have investigated the inner workings of the ViT. (Raghu et al., 2021) show that the representations of ViT encoder layers are much more uniform than the CNN-based architectures. (Park & Kim, 2022) sheds light on the Multi-head Self Attention block and its optimization. (Bhojanapalli et al., 2021) test the ViT's robustness to input and model perturbations. Their correlation analysis led to interesting findings about ViT models organizing themselves into correlated groups. Our motivation is along the lines of such studies attempting to understand the representations learned by ViTs and make them more explainable. The main contributions of this paper are summarized as follows.

1. A framework for testing compositionality in ViT encoder representations, analogous to the framework proposed by Andreas (2019) for representation learning.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

2. The use of the Discrete Wavelet Transform (DWT) to generate basis sets (input-specific primitives) for images. To the best of our knowledge, previous works have not used this approach to analyse ViTs.

3. Promising empirical results that demonstrate compositionality in the encoder representations of the ViT. Our analysis reveals that ViT patch representations at the final encoder layer are compositional for the DWT primitives obtained by a one-level decomposition.

2 Background

2.1 Vision Transformers

Following the success of transformers Vaswani et al. (2017) in NLP, (Dosovitskiy et al., 2021) adapted them to vision tasks. ViTs divide an input image into patches, each of which is tokenized. Positional embeddings are added to each token embedding to preserve its spatial location. A special CLS token is appended to the input embeddings, and is used for the final classification. The dimension of all the patch representations remains constant throughout the encoder layers, which gives the ViT model flexibility.

2.2 Compositionality in Representation Learning

Representational compositionality has been a field of study since the days of the connectionist approach (Fodor & Pylyshyn, 1988; Chalmers, 1990). Its linguistic origins still make themselves known in current research, with most investigations focusing on representations in NLP tasks and models (Chen et al., 2023; Li et al., 2023; Dziri et al., 2023). (Janssen, 2001) defines the principle of compositionality as “the meaning of a compound expression is a function of the meaning of its parts and of the syntactic rule by which they are combined". The notions of *meaning* and *syntactic rules* in language model representations naturally lend themselves to the study of compositionality.

Formally, a compositional representation function learns a homomorphism between the input space and the its representation space (Andreas, 2019). A homomorphism $\phi : H \to G$ is a map between two groups $(H, \cdot)$ and $(G, \oplus)$, such that if $\phi(h_1) = g_1$ and $\phi(h_2) = g_2$ for $h_1, h_2 \in H$ and $g_1, g_2 \in G$, then $\phi(h_1 \cdot h_2) = g_1 \oplus g_2$.

The study of compositional nature of pretrained models is motivated, in part, by interpretability. A model that can break its input into meaningful pieces and reconstruct it in a human-understandable manner is more interpretable than a model that does not do so. With interpretability in mind, we pursue our investigations into the representations learned by ViT.

In the NLP domain, such investigations usually decompose the input space into a dictionary of words, which acts as the fundamental set used to represent all sentences. However, it is difficult to construct a dictionary of *visually meaningful* images in the image domain, since the image space is continuous. In other words, we cannot construct a dictionary with infinite cardinality. This difficulty is additionally compounded by the uninterpretable nature of the canonical basis in the image space - the set of $H \times W \times C$ matrices with every element being 0 except for a single 1 at some position. Thus, we propose a different approach to decompose an image into its visually meaningful primitives, turning to analytical tools from signal processing.

2.3 Discrete Wavelet Transform (DWT)

While the Fourier series and Fourier transform are excellent tools to analyze the frequency spectrum of images, they do not provide localization in the pixel domain. Simply put, the Fourier spectrum of an image is not visually meaningful. The DWT Daubechies (1992) stands out among time-frequency analysis tools due to its unique ability of time-frequency localization. Specifically, applying the DWT to an image decomposes it into sub-bands, which form an ideal input-dependent primitive. The invertibility of the sub-band decomposition enables lossless reconstruction, making the DWT our tool of choice for compositionality analysis. After the introduction of ViTs, the DWT has been used for lossless downsampling to address their efficiency-vs-accuracy tradeoff (Yao et al., 2022). Zhang et al. (2024) also employs the DWT to improve the quality of the input in a transformer-based network. However, to our knowledge, the DWT has not been used to explore compositionality in ViTs.

2.4 Compositionality of Image Representations

When inputs belong to the pixel space, and a neural network learns input representations, the groups across which compositionality is studied are vector spaces. These spaces need to be equipped with a binary operation that satisfies the group axioms, the natural choice being vector addition.

A homomorphism between two vector spaces V and W reduces to a linear map $T : V \to W$. This map is fully defined by how it trasforms the basis set $\{v_1, v_2, ...\}$, which we refer to as primitives. An ideal composition learner would preserve how parts combine - like the way vector addition in pixel space, to representation space. However, such behaviour is typically not observed in real models, primarily because of the deep nesting of non-linearities. Thus, we now focus on *learning* how to combine parts in the representation space, rather than assuming it's just addition.

To quantify this, we study how the wavelet sub-band representations evolve through the network. In the latent space, we recompose the primitives' representations (like we do in pixel space) and compare the result with the original image. This helps reveal how compositional the learned representations are. To make this analysis manageable, we focus on compositionality in the last encoder layer.

3 Compositionality Analysis

3.1 Drawing Parallels from Existing Works

The inspiration for a framework to study compositionality in ViTs stems from the work by Andreas (2019). The paper offers a framework to measure compositionality in deep learning models, particularly neural networks. In the context of this paper, compositionality refers to the ability of a system to represent complex ideas using simpler concepts. It introduces a metric to measure how well an explicitly compositional model $\hat{f}_\eta$ can approximate a complex model f. To draw parallels, we summarize our understanding of their framework, with corresponding analogies to our approach:

1) Representations: They define a model $f : \mathcal{X} \to \Theta$, where $\mathcal{X}$ is the input space (e.g., images), and Θ is the representation space. The output representations $\theta \in \Theta$ produced by f are analysed for compositional behaviour.

Analogy: In our work, f refers to the ViT model, $\mathcal{X}$ is the set of input images, and Θ is the space of encoder representations. Each θ represents the ViT's internal encoding of an image.

2) Derivations: Derivations $\mathcal{D}$ are recursively constructed from a finite set of primitives $\mathcal{D}_0$ using a binary bracketing operation $\langle \cdot, \cdot \rangle$. If $d_i, d_j \in \mathcal{D}$, then $\langle d_i, d_j \rangle \in \mathcal{D}$. A derivation oracle $D : \mathcal{X} \to \mathcal{D}$ maps each input to its derivation tree.

Analogy: In our framework, derivations correspond to wavelet decompositions. The DWT acts as the oracle D, constructing a tree (Figure 1) from wavelet sub-bands. Although the set of sub-bands is infinite (2.2), combining valid sub-bands (primitives) still yields a valid derivation.

3) Compositionality: The model f is compositional if it preserves the structure of composition from the input space to the representation space, i.e. it is a homomorphism from input space to representation space. A composition operation on representations $*$: $\theta_a * \theta_b \mapsto \theta$ is defined such that for any input x with derivation $D(x) = \langle D(x_a), D(x_b) \rangle$, we have:

$$f(x) = f(x_a) * f(x_b)$$

While exactly compositional primitives may not exist, can a set of candidate primitives approximate the model's internal representation? If f can be approximated by a learnable compositional model $\hat{f}_\eta$ with parameters η, then the approximation itself could serve as a measure of compositionality for f.

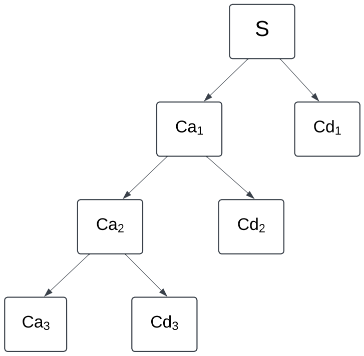
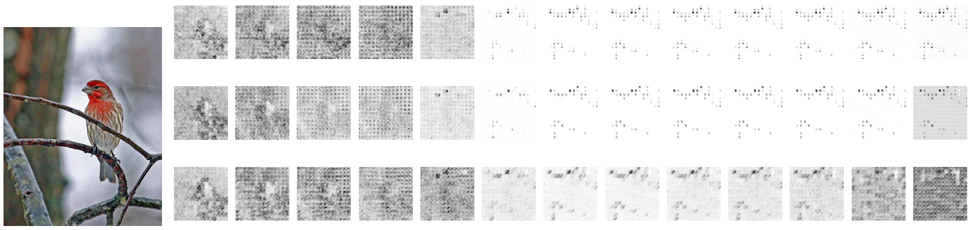


Figure 1: Tree structure of Discrete Wavelet Transform. S represents the input signal. Ca_i, Cd_i represent the approximate and detail coefficients of i^{th} level.

Analogy: In our case, the model $f : \mathcal{X} \longrightarrow \Theta$ is the ViT model. We consider an intermediate encoder layer l such that $f_l : \mathcal{X} \longrightarrow \Theta$ operates from the input space to the encoding layer l. Then, the compositional model



(a) Image

(b) SSIM maps

Figure 2: SSIM maps for each channel (R,G,B). For each encoder layer output, the original image's representation is compared with the composed image representation. The SSIM maps shown here are **after** comparison. There is no immediate notion of compositionality present visually.

$\hat{f}_\eta(d) : \mathcal{D} \longrightarrow \Theta$ can be viewed as an approximation of the encoder layer representations f_l. With this perspective, we can study the compositionality of any encoder layer of the ViT architecture.

3.2 Capturing Compositionality

The wavelet reconstruction in the image space gives back the original image without any loss of information. Thus, a natural approach to check the model's compositionality would be to examine how the wavelet reconstruction behaves in the representation space of the encoder layers. We analyse if such composition of the reconstructed encoder layer representations approximates the encoder layer representations of the original image. We identify two metrics, Structural Similarity Index (SSIM) (Wang et al., 2004) and Centered Kernel Alignment(CKA) (Kornblith et al., 2019). These metrics compare the image's encoder layer representation of the original image with that of the composed reconstruction. SSIM is a perceptual metric and takes into account local patterns of pixel intensities, their correlation, and spatial arrangements. CKA is used to compare the similarity between two sets of high-dimensional feature vectors (often from neural network layers). Using these metrics, we conduct the below analysis:

1. We use the SSIM map (Wang et al., 2004) to visualize structural similarities between the original and the composed representations. To do this, we reshape the encoder layer representation from $E_L(I)^{N-1 \times D}$ to $E_L(I)^{W \times H \times C}$ where N is the number of tokens ($N - 1$ to exclude the `CLS` token), and D is the encoder layer's hidden dimension. We measure the SSIM across the channels.

2. We plot the CKA (Kornblith et al., 2019) scores between the image representation and composed representation across all encoder layers. For this analysis, we sample 10k images(10 images per class) from the imagenet-1k dataset and average the CKA scores over all encoder layers.

Figure 2 shows the SSIM maps computed for a sample image, and Figure 3 presents the CKA scores averaged over 10K images from the ILSVRC validation set using ViT-Base representations. Neither the maps nor the scores provide any evidence of compositionality or structural similarity between the representations. It is unlikely that simply adding the individual wavelet representations would exactly give the original image's representation. This invites the possibility that reconstruction of these primitives in the representation space differs from reconstruction in the image space. Motivated by this, we investigate whether such a composition function can be learned to better approximate the true representations by relaxing the constraint that each wavelet representation has to be equally weighted.

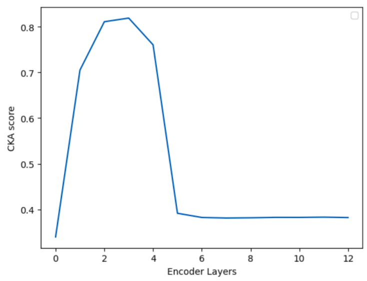


Figure 3: CKA scores of original vs composed representations at various encoder layers of ViT-B averaged over 10K images.

3.3 Compositionality Framework for ViTs

To generate a set of primitives for the pixel space, we turn to the 2D Discrete Wavelet Transform (DWT). The DWT has long been employed as a tool for space-frequency analysis due to its invertibility and for its unique ability to capture spatial resolution. Its exact reconstruction property makes it a great tool for our study of compositionality. Given any $W \times H$ image I, it can be represented in terms of its wavelet coefficients as,

$$I_{W\times H} = \sum_{i=0}^{W-1}\sum_{j=0}^{H-1} A_{M,i,j}\phi_{M,i,j} + \sum_{m=1}^{M}\sum_{i=0,j=0}^{W-1,H-1}\sum_{k=1}^{3} D^k_{m,i,j}\psi^k_{m,i,j} \tag{1}$$

where $A_{M,i,j} = \langle I_{W\times H}, \phi_{M,i,j}\rangle$ and $D^k_{m,i,j} = \langle I_{W\times H}, \psi^k_{m,i,j}\rangle$ are the approximation and detail coefficients respectively. k is the sub-band index, and the functions ϕ and ψ are the scaling (approximation) and wavelet (detail) wavelet bases. An orthogonal decomposition is assumed in this work. The first term corresponds to the approximation of the image at level M, while the second term represents all of the detail coefficients from level 1 to M. Together, they enable perfect reconstruction of the original image.

Let $E_l : \mathbb{R}^{W\times H\times C} \longrightarrow \mathbb{R}^{N\times D}$ be a function that takes an input image of dimension $W \times H \times C$ and outputs a set of N (number of patches + 1) token vectors of dimension D, produced by the l^{th} layer of a vision transformer with L encoder layers (i.e., $1 \le l \le L$).

We investigate the following question to assess whether a ViT model exhibits compositionality:

$$\sum_{l=1}^{L}\left\| E_l(I) - \left(E_l(I_{LL}) + \sum_{m=1}^{M}\sum_{k=1}^{3}(E_l I^{(m,k)}_{\text{detail}}) \right)\right\|_2 = 0? \tag{2}$$

Here, $I_{LL} = \sum_{i,j} A_{M,i,j}\phi_{M,i,j}$ denotes the low-frequency (approximation) image at level M, and $I^{(m,k)}_{\text{detail}} = \sum_{i,j} D^k_{m,i,j}\psi^k_{m,i,j}$ denotes the k^{th} directional high-frequency detail (horizontal, vertical, diagonal) at level m. . $E_l(\cdot)$ is the ViT encoder output at layer l. That is, we check if the representation of the original image at encoder layer l can be reconstructed by summing up the representations of its wavelet components. The preliminary analysis presented in Figs. 2 and 3 shows that this equality does not hold, suggesting a lack of compositionality under simple addition.

Given the highly non-linear nature of the ViT model and the high dimensionality of its representations, we relax the strict equality and ask whether a learnable function can approximate the original image's encoder layer representation:

$$E_l(I) \approx g_\eta\left(E_l(I_{LL}),\ \left\{ E_l\left(I^{(m,k)}_{\text{detail}}\right)\right\}_{\substack{1\le m\le M\\ 1\le k\le 3}}\right) \tag{3}$$

where $I_{LL} = \sum_{i,j} A_{M,i,j}\phi_{M,i,j}$ and $I^{(m,k)}_{\text{details}} = \sum_{i,j} D^k_{m,i,j}\psi^k_{m,i,j}$. That is, can we approximate original representations at layer l of the encoder by applying a learnable composition function $g_\eta(.)$ (with parameters η) on the primitive representations of the image? To emphasize, $g_\eta(.)$ attempts to find the best possible *linear combination* of the primitive representations.

We argue that popular distance metrics between these two high-dimensional representations might not be a reliable way of measuring similarity due to the curse of dimensionality. Instead, we aim to minimize the loss between the final layer classifier output logits of the original image's final cls token and the approximate *linearly combined* final layer cls token. Since we are not modifying any of the ViT model's parameters while training this composition function, we affirm that all our analyses are post-hoc and still viable probes for understanding the pretrained ViT model. Hence, our reformulated question to evaluate whether compositionally holds becomes:

$$\eta^* = \arg\min_\eta \mathcal{L}\left(E_c\left(E_l(I)_{[\text{CLS}]}\right),\ E_c\left(g_\eta\left(\left\{ E_l(\tilde{I}_p)_{[\text{CLS}]}\right\}_{p=1}^{n}\right)\right)\right) \tag{4}$$

where: Here, $\mathcal{L}$ is the loss, $E_l(I)_{[\text{CLS}]} \in \mathbb{R}^D$ is the CLS token from the original image at encoder layer l, $\tilde{I}_p$ denotes the p^{th} primitive (either $\sum A_{M,i,j}\phi_{M,i,j}$ or $D^k_{m,i,j}\psi^k_{m,i,j}$), $E_l(\tilde{I}_p)_{[\text{CLS}]}$ is its corresponding CLS token, and $g_\eta : \mathbb{R}^{n\times D} \rightarrow \mathbb{R}^D$ is a learnable linear function mapping n primitives to a single vector.

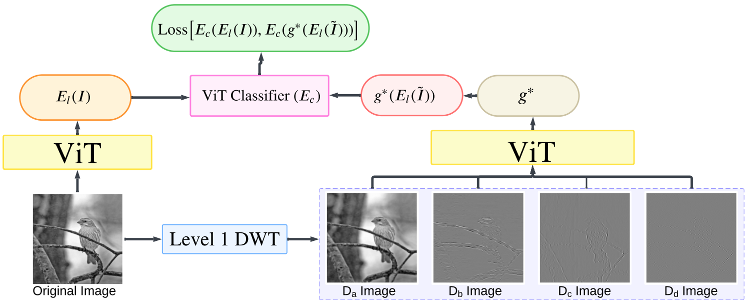


Figure 4: Overview of the proposed compositionality framework for ViTs. The figure presents learning the composition function for Level 1 DWT decomposition. D_a, D_b, D_c, D_d are the coefficients of the wavelet decomposition discussed in 1.

3.4 Applications of this framework

Our proposed framework demonstrates that the representations at the final encoder layer exhibit compositional behaviour using DWT primitives obtained by a one-level decomposition. Now, we validate this framework's utility by evaluating it under commonly encountered real-world distortions. Sepcifically, we assess compositinality in the presence of additive noise and compression. Our experiments show that our compositional framework is robust, and holds even in the presence of such distortions.

4 Experimental Setup and Results

We conduct our experiments on the ImageNet-1k dataset (Deng et al., 2009), sampling 50 images per class (50,000 total) and splitting them into 60:20:20 train/val/test sets. To train the composition function g^*, we first generate wavelet primitives via the DWT and pass them through the ViT to obtain their encoder layer outputs $E_l(\tilde{I})$. The CLS tokens from these outputs serve as inputs to g^*, which outputs a composed CLS token. This is fed to the ViT classifier, and the target is the original image's classification output—not the image label—as out aim is to measure how well the composed token reproduces the original representation. Only g^* is trained (using cross-entropy loss), while ViT weights remain frozen. Models are trained for 100 epochs using SGD with a learning rate of 0.001.

We restrict our analysis to two levels of DWT decomposition using two wavelet bases: Haar and db4. To assess generalizability, we evaluate two ViT variants—ViT-B and ViT-L—both pretrained on ImageNet-21k (14M images, 21k classes). We relax the equal-weight constraint from Section 3.2 and explore three adaptive weighting strategies for the composition function g^*:

1. **Convex:** weights satisfy $\sum_i \eta_i = 1$, $\eta_i \geq 0$;
2. **Conic:** non-negative weights $\eta_i \geq 0$;
3. **Unconstrained:** no restrictions on η.

We used the same subset of images from the ImageNet-1k to learn the composition function g^* for these three variations. While the framework can be used to study any encoder layer in the model, we restrict our analysis to the last layer, whose outputs are often inputs to downstream tasks.

4.1 Composition Approximation: Accuracy of Learned Model on ground truth

Our initial analysis brings us back to our first question (eq. 2) posed in section 3.3: whether wavelet based primitive representations, satisfy compositionality under simple summation. We compare the classification accuracy of the representations composed following simple summation (eq. 2) and that of the learned composition model. Table 1 compares the classification accuracies for three cases (i) original ViT's output, (ii) Output from summing the individual wavelet decomposition representations, and (iii) Output from the proposed learned composition model. Please note, these accuracies are cclaulcated on the ground truth. These results clearly demonstrate how the learned representations perform significantly better than just the summed representations. Notably, the performance for level 1 decomposition is almost on par with the original ViT model's accuracy. These findings confirm that the learned composition function g^* is effective in capturing compositionality of level 1 wavelet primitives.

			Learned		
Model	Original	Summed	Unconstrained	Conic	Convex
ViT-B (Haar-level 1)	0.792	0.13	0.775	0.775	0.771
ViT-B (db4-level 1)	0.792	0.13	0.777	0.775	0.772
ViT-L (Haar-level 1)	0.809	0.18	0.797	0.795	0.795
ViT-B (Haar-level 2)	0.83	0.005	0.51	0.5	0.48
ViT-B (db4-level 2)	0.83	0.005	0.51	0.51	0.48
ViT-L (Haar-level 2)	0.82	0.003	0.63	0.62	0.59

Table 1: Accuracies of original representations vs. summed representations vs. learned compositions. Note that the learned representations perform significantly better than just the summed representations.

Model	Unconstrained	Conic	Convex
ViT-B (haar-level 1)	0.87	0.87	0.86
ViT-B (db4-level 1)	0.9	0.9	0.89
ViT-L (haar-level 1)	0.92	0.91	0.91
ViT-B (haar-level 2)	0.53	0.51	0.49
ViT-B (db4-level 2)	0.69	0.68	0.61
ViT-L (haar-level 2)	0.65	0.64	0.61

Table 2: Relative accuracy of the learned composition models. Note that the target for the composed representation is the output predicted by the original image classifier (not the ground truth label).

4.2 Composition Approximation: Understanding the Learned Model Weights

Model	Unconstrained	Conic	Convex
ViT-B (haar)	[2.02, -0.18, 0.43, 0.18]	[1.67, 0.34, 0.57, 0.02]	[0.66, 0.11, 0.10, 0.12]
ViT-B (db4)	[2.02, 0.1, -0.15, -0.16]	[1.65, 0.12, 0.63, 0.03]	[0.62, 0.09, 0.25, 0.03]
ViT-L (haar)	[1.93, 0.16, -0.02, 0.25]	[1.81, 0.28, 0.13, 0.44]	[0.68, 0.1, 0.05, 0.16]

Table 3: Weights learned by the proposed composition model (g^*) for level 1 wavelet decomposition.

Model	Unconstrained	Conic	Convex
ViT-B (haar)	[1.32, 0.35, -0.07, -0.14, 0.65, -0.20, 0.21]	[1.88, 0.61, 0.35, 0.17, 0.10, 0.10, 0.44]	[0.42, 0.13, 0.05, 0.13, 0.07, 0.10, 0.06]
ViT-B (db4)	[1.52, -0.18, 0.06, 0.30, 0.35, 0.16, -0.21]	[1.64, 0.40, 0.12, 0.02, 0, 0.03, 0]	[0.43, 0.11, 0.08, 0.07, 0.14, 0.06, 0.08]
ViT-L (haar)	[1.52, -0.01, -0.21, 0.29, 0.06, -0.01, 0.34]	[1.82, 0.29, 0.32, 0.17, 0, 0.33, 0.23]	[0.40, 0.11, 0.10, 0.08, 0.09, 0.06, 0.13]

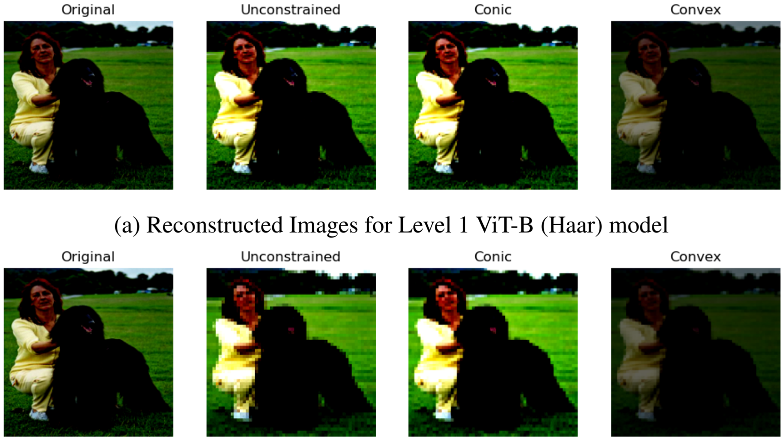
Table 4: Weights learned by the proposed composition model (g^*) for level 2 wavelet decomposition.

To evaluate how accurately our learned composition function g^* approximates the original image's representation, we compute the relative accuracy (by considering the **original model's (ViT)** output logits as the ground truth, or reference target). Table 2 presents those results. Interestingly, the relative accuracies are similar across different constraints (convex, conic, and unconstrained variations of g^*). To investigate this further, we look at the learned model weights in Table 3 and Table 4, which indicate the relative importance weights assigned to different sub-bands. Across all settings, the learned model g^* weighs the approximation (Low-pass filtered image) coefficient i.e the first value significantly more than the other coefficients in the representation space.

Notably, there is no discernible pattern among the learned weights under different constraints. There is considerable variation among the weights assigned for different g^*'s, but their performance is quite similar. There could be multiple such compositions for an encoder layer, which leads to further questions about the representation space.

4.3 Composition Approximation: Learned Reconstructed Image Analysis

In this subsection, we investigate how the weights learned by the proposed composition function (g^*) influence the reconstruction of the original images, when these weights are applied to the primitives (wavelet sub-bands) in the image space. Simply put, we transfer the weights learned from the ViT encoder embeddings space to their corresponding wavelet sub-bands in the image space, thereby reconstructing a weighted image in the image space. We consider a subset of 200 images to conduct this analysis. Table 5 presents the classification accuracy of the ViT model on the reconstructed images. Note that although there is a significant drop in level 2 accuracies, the learned weights translate back well for level 1 decomposition. Figure 5 visualizes the reconstructed images using Level 1 ViT-B (haar) model and Level 2 ViT-B (haar) model. Interestingly, although the convex combination of the sub-bands in the image space significantly affect the pixel intensities, their performance is at par with other learned models.



(a) Reconstructed Images for Level 1 ViT-B (Haar) model

(b) Reconstructed Images for Level 2 ViT-B (Haar) model

Figure 5: Reconstructed images obtained by applying the learned composition model g^*'s weights to the corresponding sub-bands of the original image in the input space.

Model	Original	Unconstrained	Conic	Convex
ViT-B (haar-level 1)	0.79	0.72	0.72	0.76
ViT-B (db4-level 1)	0.84	0.81	0.81	0.82
ViT-B (haar-level 2)	0.84	0.58	0.48	0.64
ViT-B (db4-level 2)	0.84	0.49	0.51	0.65
ViT-L (haar-level 1)	0.83	0.82	0.82	0.80
ViT-L (haar-level 2)	0.83	0.63	0.68	0.71

Table 5: Classification accuracy of ViT-B on reconstructed images generated by applying the learned weights of the proposed composition model (g^*) to the corresponding sub-bands in image space.

4.4 Composition Approximation: Error Analysis

While the classification performance of the proposed compositional model presented thus far provides a broad picture of compositionality, a natural question about error in composition arises. Here, we compare the compositional model's predictions - particularly its misclassification - with that of the original model. Since our downstream task is image classification, we analyze the composition error via prediction discrepancies. It would be interesting to explore other ways to study the error in composition in the future. In this preliminary experiment, we sample 1000 images from the Imagenet-1K dataset, and the performance of both the original and compositional model (level 1 DWT decomposition) is evaluated as follows:

- Percentage of images where the original model is accurate and the learned model is inaccurate ($\text{Err}_{\text{Learned}\neg\text{Org}}$).

Model	$\text{Err}_{\text{Learned}}$	Err_{Org}	$\text{Err}_{\text{Learned}\neg\text{Org}}$	$\text{Err}_{\text{Org}\neg\text{Learned}}$	Err_{both}
ViT-$\text{B}_{\text{Unconstrained}}$(Level-1 haar)	19.7%	17.1%	3.8%	1.2%	15.9%
ViT-B_{Conic}(Level-1 haar)	19.7%	17.1%	3.8%	1.2%	15.9%
ViT-B_{Convex}(Level-1 haar)	20.4%	17.1%	4.3%	1%	16.1%

Table 6: We report the Errors on the test set using the learned composition model. The reported percentages are calculated on the 1000-image subset. Interestingly, there is a fraction of samples on which the learned composition performs better than the original model.

- Percentage of images where the learned model is accurate and the original model is inaccurate ($\text{Err}_{\text{Org}\neg\text{Learned}}$).
- Percentage of images where both models are inaccurate (Err_{both}).

The results in Table 6 provides a comparative analysis of prediction error between the learned model and the original ViT. As anticipated, the learned model commits relatively more errors than the original. However, it is noteworthy that the learned model also performs better on some images. This preliminary analysis provides sufficient motivation to further analyse the role of individual wavelet representations towards the model's prediction.

4.5 Composition Approximation: Effect of distortion on images

To further cement the practical utility of our proposed framework, we explore whether it persists when images are subjected to distortions. We measure how JPEG compression and additive Gaussian noise affect classification accuracy of our learned model.

Image Type	Image Accuracy	Learned Accuracy		
		Unconstrained	Conic	Convex
Original Images	0.792	0.775	0.775	0.771
Compressed Images	0.628	0.603	0.603	0.599
Noisy Images	0.593	0.565	0.565	0.563

Table 7: Comparison of classification accuracies for original, compressed, and noisy images. Learned accuracies are obtained from the output of the composition model g^*.

Table 7 clearly demonstrates that our framework is robust to distortions, and compositoniality holds even with compressed and noisy images.

5 Conclusion and Future Work

Our work explores notions of compositionality present in the ViT encoder layer representations. We present a general framework to measure compositional behaviour in encoder layers of ViT-based architectures. Fundamental to this framework is the use of the DWT representation as an input-dependent primitive. Our findings indicate the possibility of compositional behaviour in the ViT model. Specifically, we provide evidence for compositionality in the last encoder layer when primitives induced by a one-level DWT decomposition are applied. While our present analysis is restricted to the final encoder layer, we aim to explore all the encoder layers for potential compositionality. We hope this work leads to further analysis for explainability in ViT's.

6 Reproducibility

The code for implementing the proposed compositionality framework is provided at Compositionality-in-ViTs

References

Jacob Andreas. Measuring compositionality in representation learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL `https://openreview.net/forum?id=HJz05o0qK7`.

Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10231–10241, 2021.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

D. Chalmers. Why fodor and pylyshyn were wrong : the simplest refutation. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society, Cambridge*, pp. 340–347, 1990. URL https://cir.nii.ac.jp/crid/1570854174742444672.

Jiaao Chen, Xiaoman Pan, Dian Yu, Kaiqiang Song, Xiaoyang Wang, Dong Yu, and Jianshu Chen. Skills-in-context prompting: Unlocking compositionality in large language models. *ArXiv*, abs/2308.00304, 2023. URL https://api.semanticscholar.org/CorpusID:260351132.

Ingrid Daubechies. Ten lectures on wavelets. *Society for industrial and applied mathematics*, 1992.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Fkckkr3ya8.

Jerry A. Fodor and Zenon W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1):3–71, 1988. ISSN 0010-0277. doi: https://doi.org/10.1016/0010-0277(88)90031-5. URL https://www.sciencedirect.com/science/article/pii/0010027788900315.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Theo M. V. Janssen. Frege, contextuality and compositionality. *Journal of Logic, Language, and Information*, 10(1):115–136, 2001. ISSN 09258531, 15729583. URL http://www.jstor.org/stable/40180264.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited, 2019. URL https://arxiv.org/abs/1905.00414.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pp. 280–296. Springer, 2022.

Yingcong Li, Kartik K. Sreenivasan, Angeliki Giannou, Dimitris Papailiopoulos, and Samet Oymak. Dissecting chain-of-thought: Compositionality through in-context filtering and learning. In *Neural Information Processing Systems*, 2023. URL https://api.semanticscholar.org/CorpusID:265051253.

Namuk Park and Songkuk Kim. How do vision transformers work? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=D78Go4hVcxO.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=R-616EWWKF5.

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7262–7272, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Ting Yao, Yingwei Pan, Yehao Li, Chong-Wah Ngo, and Tao Mei. Wave-vit: Unifying wavelet and transformers for visual representation learning, 2022. URL https://arxiv.org/abs/2207.04978.

Shengli Zhang, Zhiyong Tao, and Sen Lin. Waveletformernet: A transformer-based wavelet network for real-world non-homogeneous and dense fog removal, 2024. URL https://arxiv.org/abs/2401.04550.