UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Computer Science Curriculum

Braian Olmiro Dias

Content based analysis of compositionality in Vision Transformers

Master's Thesis (30 ECTS)

Supervisor(s): Tarun Khajuria, MSc

Tartu 2023

Content based analysis of compositionality in Vision Transformers

Abstract:
Neural Network models have achieved state of the art results in various tasks related to vision and language, there are still questions regarding their logical reasoning capabilities. In particular, its not clear whether these models can reason beyond using analogy. For example, in an image captioning model, the model can either learn to correlate a scene representation to a caption i.e. text space, or the model could learn to bind objects explicitly and the utilise the explicit composition of individual representations. The inability of models to perform the later has been related to their failures to generalise on wider scenarios in various tasks. Transformer based models have achieved high performance in various language and vision tasks. Their success has been accredited to their ability to model long range relations between sequences. But in vision transformers there has been a discussion that the use of patches as tokens and the interaction between them, gives them an ability to flexibly bind and model compositional relations between various objects at different distances. Hence, showing aspects on explicit compositional abilities. In this thesis, we perform experiments on the Transformer (VIT) based vision encoder of an image captioning model. In particular we probe the internal representation of the encoder at various layers to examine if a single token captures the representation of 1) an object 2) related objects in scene 3) composition of two objects in the scene. In our results we find some evidence to hint binding of object properties into a single token as the image is processed by the transformer. Further, this work provides a list of methods to create and setup a dataset to study internal compositionality in Vision Transformers models and suggests future lines of study to expand this analysis.

Keywords:
Transformers, Vision transformer, compositionality, computer vision, neural networks, image captioning models

CERCS: P176 - Artificial Intelligence

Sisupõhine analüüs Vision Transformerite kompositsioonilisusest

Lühikokkuvõte:

Närvivõrgu (Neural Network) mudelid on saavutanud tipptasemel tulemusi mitmesugustes ülesannetes seoses nägemise ja keelealase tööga, kuid nende loogilise mõtlemise võime kohta on endiselt rida küsimusi. Pole selge, kas need mudelid suudavad analoogia kasutamisest kaugemale mõelda. Näiteks, kas pildi pealkirjastamise mudelis suudab mudel õppida korreleerima stseeni esitust pealkirjaga, st tekstiruumiga, või suudab mudel õppida objekte siduma selgesõnaliselt ja kasutama individuaalsete esituste selget kompositsiooni. Mudelite suutmatus viimast teostada on seostatud nende ülesannete laiemate stsenaariumide generaliseerumise ebaõnnestumisega. Transformer-põhised mudelid on saavutanud kõrge jõudluse erinevates keele- ja nägemisega seotud ülesannetes. Nende edu on omistatud nende võimele modelleerida pikamaa suhteid jadade vahel. Kuid Vision Transformer käsitlevas arutelus on öeldud, et paikade (patches) kasutamine märkidena (token) ja nende vaheline koostoime annab neile võime paindlikult siduda ja modelleerida kompositsioonilisi suhteid erinevate objekide vahel erinevatel kaugustel. Seetõttu näitavad need mudelid selget kompositisoonivõimet. Käesolevas lõputöös viime läbi katseid pildi pealkirjastamise mudeli Transformer (VIT) põhise nägemise kodeerija peal. Eelkõige uurime sisemist esitust kodeerija eri kihtidel, et uurida, kas üks märk kajastab esituse 1) objekti 2) stseenis seotud objekte 3) kahte stseenis oleva objekti kompositsiooni. Meie tulemuste põhjal leiame mõningaid tõendeid, mis viitavad, et transformaatorit töödeldes seotakse objekti omadused ühe märgiga (token). Lisaks pakub see töö meetodite loetelu andmestiku loomiseks ja seadistamiseks, et uurida Vision Transformer mudelite sisemist kompositsioonilisust ning soovitab selle analüüsi laiendamiseks tulevasi uuringusuundi.

Võtmesõnad:
Transformerid, Vision transformer, kompositsioonilisus, arvutinägemine, närvivõrgud, pildi pealkirjade mudelid

CERCS: P176 - Tehisintellekt

Contents

1 Introduction

Compositionality is a fundamental aspect in human intelligence that allows us to understand sentences from a set of words and the material world from a set of objects. Humans are capable of learning complex concepts by combining known basic ones both in the language and vision domain. The sentence "Two boys in the rain with the youngest holding a large umbrella" contains not only a description of objects portrayed in the scene(nouns) but the attributes which describe them (adjectives) and the relation between the parts (a verb in this case - boys holding a large umbrella). The composition of these parts represent the whole scene and can be easily interpreted by humans both through reading the description or by looking at the image. Computers, however, may or may not perceive this scene the same way.

Recent developments in Computer Vision (CV) and Natural Language Processing (NLP) are based on neural networks with less inductive bias along with self-supervised learning and massive datasets[SVB+21, CSDS21, TSF+16]. Transformers[VSP+17] are capable of learning relationships between data to perform several tasks such as image classification and machine translation at high level and quickly became the state-of-the-art architecture for both CV and NLP.

Despite its success, little is known about what such models learn and what type of basic knowledge is built internally in order to execute a complex task such as image captioning. While popular benchmarks for computer vision tasks[RDS+15] are the standard method for testing the algorithm's performance, they do not encompass compositional understanding. For example, is a captioning model capable of distinguishing the sentences "the horse is eating the grass" and "the grass is eating the horse"? By doing that, does it encode compositional representation of the concepts internally?

Based on the literature about compositionality in neural networks, we explore the internal representation of concepts in a multi-modal network that combines visual and text information to perform image captioning. Specifically, we test whether tokens in the Vision transformer[DBK+20] encoder acquire representations combining the information from two or more objects. Next, we devise a few experiments containing the following relationships:

- **Person** *has/used* **Accessory**
- **Kitchenware** *on* **Dining table**
- **Person** *eats/holds* **Food table** and **Food** *on* **Dining table**

Lastly, we probe image captioning models based on transformers to understand if they are able to encode two concepts internally in order to derive complex answers.

Our analysis may provide clues on what vision models learn about visual perception helping us understand why models succeed or fail.

Our main contributions are:

- Pixel-wise (or token-wise) compositionality analysis of Transformers-based caption models;

- Evaluation of different methods to extract COCO[LMB+14] subsets for analysing compositionality, in particular, interaction between the objects in an image.

In section 2 we explore the literature regarding Transformers, image captioning and decoding. In section 3 we briefly describe other related works on compositionality in transformers. Section 4 shows the setup of our experiments in regards to the network used, dataset, internal representation and probing. Then on section 5 we share our results and in section 6 we share our main ideas about this work and how we can improve it.

2 Background

In this section we will cover the core concepts used in our work, how they connect to each other in the image captioning task, and how we plan to test compositionality in the Vision Transformer-based captioning model.

2.1 Neural networks

Neural networks are a type of machine learning algorithm loosely based on the structure and function of the human brain. It consists of interconnected nodes (neurons) organized in layers that process information and can learn how to represent complex functions. A feed forward neural network (FFNN) processes information in only one direction, from input layer through one or several hidden layers to the output layer, without feedback loops. In supervised learning we use the backpropagation algorithm to calculate the gradient of the error with respect to each node, and then use an optimization algorithm, such as gradient descent, to update the original weights of each node to minimize the error. This process of learning how to produce better outputs is referred to as the training process of a neural network. Different neural network architectures are used today in a variety of tasks, such as image classification and natural language processing (NLP).

2.2 Transformers neural networks

Transformers[VSP+17] were a major breakthrough in NLP tasks such as language translation. It employs self-attention to capture long range dependencies between words, calculating the importance of the elements in the sentence with respect to each other. In language translation tasks, the embedding of each word (token) has a score calculated through query (q), key (k) and value (v) vectors, where the result is the product of the query vector (a token in a certain position) with the key vector of the word we are scoring. The score allows the network to understand how much focus other parts of the input sequence need in reference to a word in a certain position. The scores are passed through a softmax function to obtain weights on the values. This operation can be done by packing the queries into a matrix Q, and key and values into matrices K and V, respectively. The attention is calculated as in equation 1:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

Where $\sqrt{d_k}$ is the dimension of the key vector k and query vector q.

The result of the attention function is multiplied with the value vector, effectively removing focus of irrelevant words based on self-attention. Lastly they are summed up to produce the output of the self-attention layer. In a vanilla Transformers Encoder (Figure

1) block there is a self-attention layer followed by a Feed Forward Neural Network, with the addition of skip connection and normalization to improve performance. In the original implementation of Transformers used for machine translation, there is a decoder which consists of a stack of layers performing self-attention with an additional block that performs attention calculation over the output of the encoder block.

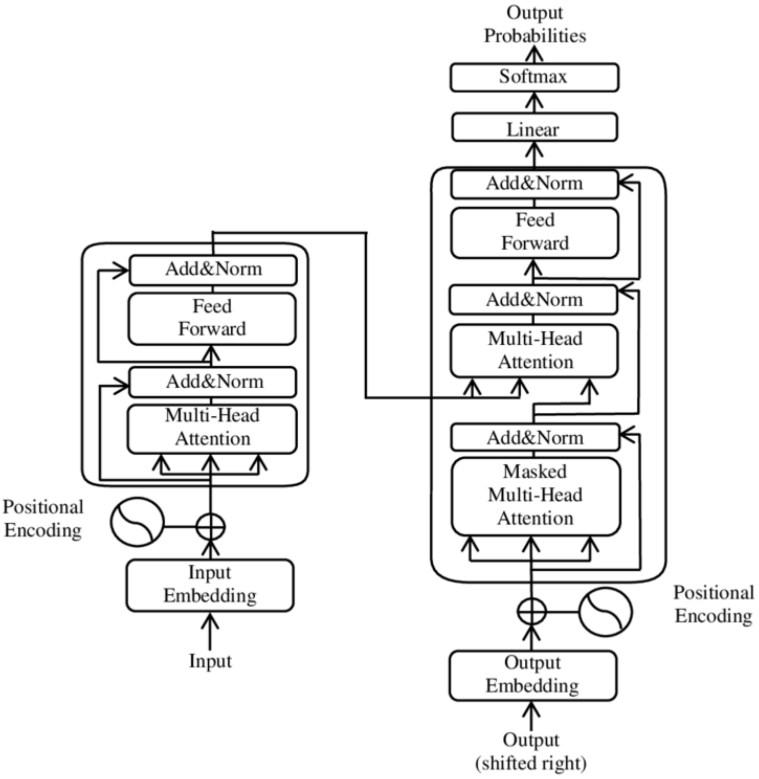


Figure 1. The transformer model architecture originally proposed for machine translation.

2.3 Vision Transformers

Vision Transformers[DBK$^+$20], ViT, apply the same core concepts of self-attention in the image domain, by splitting the image into several fixed-size tokens and computing the importance of them with respect to each other. These tokens are flattened, becoming a series of 1D vectors, and then a special token called *CLS* is inserted in the beginning of the sequence of patches. This sequence of vectors is linearly projected and enriched with positional embeddings, and then fed into a series of stacked Transformer Encoders.

The vanilla Transformer Encoder employs several self-attention layers in parallel, called multi-head attention, that allows the model to attend to information from different representation sub-spaces at different positions. The output of each block is called a hidden state and it has the same dimensionality as the input sequence. The input is processed by each Transformer block, sequentially, until it reaches the last one. In the original ViT used for image classification (Figure 2), an extra FFNN is added to process the *CLS* token and output class scores. This token can be seen as a learnable representation of the entire image, and its position at the beginning of the sequence ensures it receives special attention from the transformer during processing. The original ViT has several configurations with 12, 24 and 32 stacked Transformers blocks, called ViT-Base, ViT-Large and ViT-Huge, respectively.

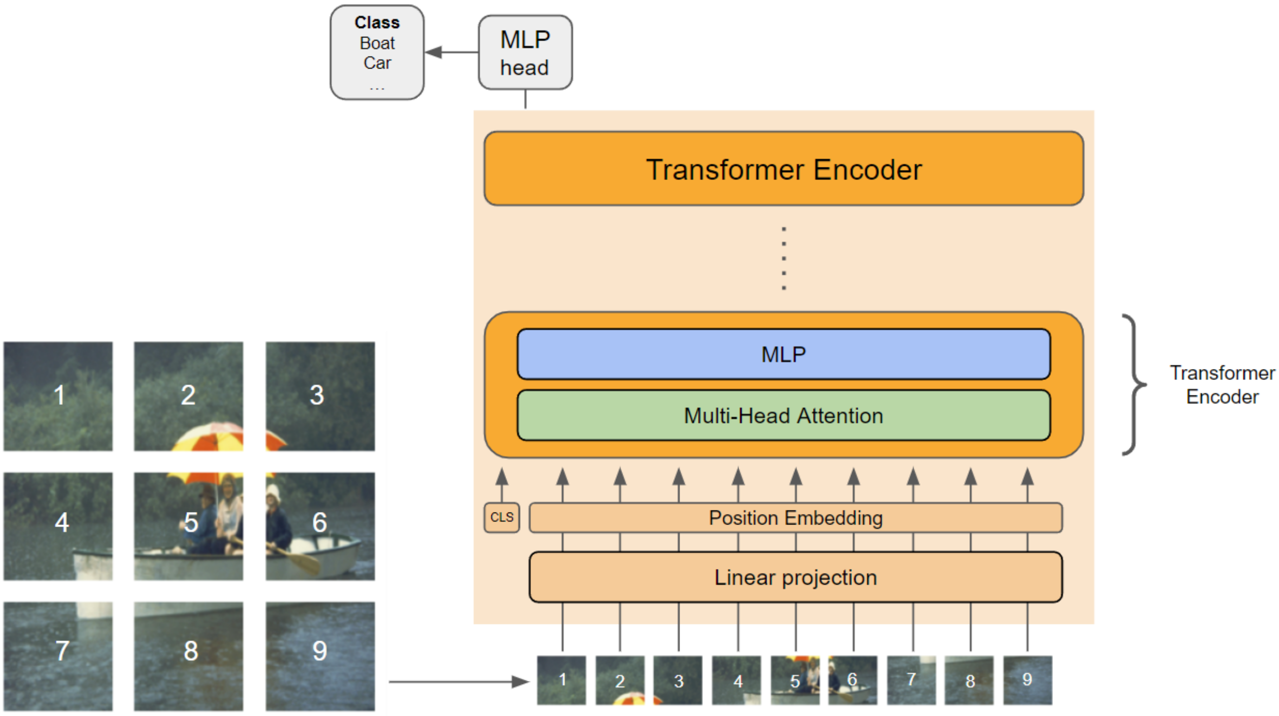


Figure 2. Vision Transformer (ViT) proposed for image classification.

2.4 Image captioning

Image captioning task is an intersection of NLP and computer vision, where the model has to describe the input image in natural language. In computer vision, CNNs[KSH17, SZ14] are a type of neural network that uses specialized layers that perform convolutions. They became popular and fomented the development of vision based models to human level in many tasks. CNNs are able to identify features and patterns in images and are used to output a fixed length vector that can be used for other downstream tasks such as object classification and detection. Some works[VTBE15] took advantage of CNNs to embed images into a feature vector, and then use a language model on top of that to predict captions. In [VTBE15], an encoder decoder approach is used with CNN as encoder followed by a recurrent neural network (RNN) decoder that generates sentences. Later, [XBK$^+$15] added an attention mechanism to the RNN decoder and improved the original work.

Current state of the art image captioning models, such as [LXT$^+$22] and OFA[WYM$^+$22], use an encoder-decoder architecture with different implementations of Transformers. In a simplified architecture, the encoder model encodes the image into a sequence of embeddings, while the decoder generates the caption. Cross-attention is frequently used to combine the cross-modal (image, text) embeddings and its calculations are the same as self-attention, but with two inputs.

In our work we use an encoder-decoder transformers-based captioning model[Kum22] with a pre-trained ViT-Base as the Encoder and a pre-trained GPT-2[RWC$^+$19] decoder. A cross-attention module is added to calculate the importance of the output of the Transformer encoder (ViT) in regards to the text processed by the decoder. Figure 3 shows the architecture of the model and in section 4.1 we explore how we chose this network. The whole model is then fine tuned using the COCO dataset[LMB$^+$14] including images and captions. This architecture is simple but robust, since it employs two components which achieve near state-of-the-art results in their vision and language tasks separately. Other vision and language transformers that can be used for image captioning, such as VisualBERT[LYY$^+$19], ViLBERT[LBPL19], use region features extracted from intermediate layers of a Faster R-CNN[RHGS15] to select image regions and feed into BERT[DCLT18].

2.5 Decoding and probing methods

Decoding, in general terms, is the process of translating information from one format into another that is more useful or readable. In our daily lives it happens when a compressed video plays on our computer or a compressed audio file is uncompressed to be audible. Brain neural activity decoding uses machine learning to recognize brain activities via Electroencephalogram (EEG) or functional magnetic resonance imaging (fMRI)[DCDN22], for example. A variety of deep neural networks have been used to decode brain activity, such as decoding natural images from fMRI data. Along the same lines, our work tries to decode visual compositionality from internal representations of Transformers.

In neural networks, probing has been used to better understand the internal dynamics of different architectures [AB16, HL19] and can be a powerful tool to infer whether a certain property was encoded in the representation. For that, a separate supervised model is trained to predict, or decode, properties from internal representations. it is assumed a high evaluation accuracy will imply the probe has evidence the property was indeed encoded internally. Language models have been explored[BDD+17] to assess the quality of its internal representations for decoding part-of-speech, semantic and morphological tagging tasks. In these tasks the aim is to understand whether the neural network, often seen as a black box, is encoding concepts from language that humans know of and are useful for language understanding. Insights obtained from probing may help us understand the reason models succeed or fail and what they learned.

Based on the concepts presented above, our work uses probes to test whether a vision transformer-based image captioning model is capable of representing internally compositionality, that is the meaning of the whole is a function of the meanings of its parts. We devise three tasks and analyze the evaluation accuracy of our probes under different settings to validate our hypothesis. In the next section we explore the literature related to compositionality in vision and language models.

3 Related work

Recently the interest in understanding general properties of vision transformers has grown and several works explore whether these models are capable of learning complex relations from data. Humans are able to interpret a natural scene as a function of its parts: we parse an image to understand the single objects present in the scene, what attributes they have, how they interact with each other and in which order they interact. In language, we are also capable of interpreting individual parts of text and composing complex representations from it. As an example, we naturally understand both from vision and text that the sentence “there is a book on the table” contains an object (book) on top (relation) of another object (table). This simple example can be challenging for computers to understand, despite the high accuracy deep learning models can achieve in tasks such as image classification, object detection and image captioning. We theorize models combining two streams of information, image and text, trained for tasks such as Visual Question Answering (VQA) and image captioning, should understand basic concepts and relationships in order to perform their tasks competently.

In this section we briefly explore a few popular datasets used in vision and language models, compositionality in neural networks and additional work related to reasoning.

3.1 Datasets for vision and language models

Vision and Language models (VLMs) have made incredible progress in recent years boosted by the usage of massive datasets used in self-supervised learning. These datasets all have millions of images sourced mainly from the web. **CC-12M**[CSDS21]: The Conceptual Captions 12M dataset is a large-scale database of image captions, consisting of over 12 million images and their captions. The images in the dataset are sourced from the web and cover a wide range of concepts and topics. Each annotated image provides a rich source of training data for image captioning and other vision and language tasks. **YFCC100M**[TSF$^+$16]: The Yahoo Flickr Creative Commons 100 Million dataset is a massive collection of images and videos, consisting of over 100 million items. The dataset was created by collecting publicly available content from Flickr and other sources. The dataset has been used for training and evaluating vision and language models. **Laion-400m**[SVB$^+$21]: A collection of images and corresponding text descriptions, consisting of over 400 million examples. The dataset was also created by mining images and text from the web and other sources, but then filtered using CLIP[RKH$^+$21].

3.2 Benchmarks for compositionality in neural networks

Along with the development of VLMs came the interest in understanding what these high-performance neural networks learn about vision. Deep learning is by no means easy to interpret, and such complex models accentuate the problem of understanding the decisions made by machine learning algorithms. Insights on this area may help us comprehend why models succeed or fail and what they learned. Currently there is no standard benchmark for testing compositionality in neural networks. However, recent works propose different datasets for testing several properties related to compositionality and reasoning in vision and language models.

3.2.1 Winoground

Winoground[TJB$^+$22] proposes a benchmark to evaluate visio-linguistic compositional reasoning, where two captions containing the same set of words describe two different scenes. It measures how well models can distinguish between the sentences such as, "It's a fire truck" and "it's a truck fire", when two images depicting the exact scene are presented. In order to have a good performance in the benchmark models must be able to acquire knowledge from text and images and encode compositional structure internally. Their work defines a combined metric considering:

- **Text score**: Measures whether a model can identify the correct caption;
- **Image score**: Measures whether the model can select the correct image given a caption.

They find that all vision and language models perform barely better than chance, with the highest scores being $\sim$50% and $\sim$70% worse than humans perform in text score and image score respectively. The hand-curated dataset proposed by their work might serve as a useful evaluation dataset for compositional reasoning, however it is limited in size with only 400 examples.

3.2.2 CREPE

CREPE[MHG$^+$22] (Compositional REPresentation Evaluation) introduces a compositionality evaluation benchmark based on two fundamental aspects of compositionality: systematicity and productivity. Systematicity evaluates how a model is able to recombine atoms (objects, relationships and attributes) and their composition (compounds). For example, a set of atoms in a scene can be palm, three, with, flowers and a compound derived from these atoms would be palm trees with flowers. Productivity measures how well a model can match a given image with the corresponding caption regardless of the complexity (number of atoms in the scene). In their work they use popular training sets CC-12M[CSDS21], YFCC-15M[TSF$^+$16], and LAION-400M[SVB$^+$21] to identify atoms and compounds as well as split image-text pairs in nine levels of complexity according to the number of atoms present in the caption. Tests with different models, such as FLAVA[SHG$^+$22] and CLIP[RKH$^+$21], demonstrate results in both aspects decrease along with input complexity.

3.2.3 Attribution, Relation and Order (ARO) benchmark

VLMs demonstrated high performance on important benchmarks however this does not indicate the model has understanding of compositional features in text or images. The work described in [YBK+22] is highly pertinent to the field of visual representation learning, their work has effectively identified failure modes in most vision-language models and suggested a reasonable benchmark for evaluating Attribution, Relation and Order (ARO). Humans naturally scan a scene and perceive relation between objects but it is not clear whether ML algorithms can understand at the same level. They test order understanding, such as "dog is behind the three" versus "three is behind the dog" and “the horse is eating the grass” versus “the grass is eating the horse”.

The same study also tests attribution of properties to objects, such as "the crouched cat and the open door" and "the open cat and the crouched door". Lastly, they test order sensitivity to evaluate whether the model chooses the correct ordering of words within a caption, as in “remarkable scene with a blue ball behind a blue scene” and “green ball with a remarkable chair behind a blue scene”. They conclude VLMs show poor compositional and order understanding despite the high accuracy in text-image retrieval evaluation on large scale datasets. One of their proposals to improve compositionality is to fine-tune VLMs with sampled strong alternative images and also targeted negative captions (swapping nouns or verbs, such as “The horse is eating the grass and the zebra is drinking the water” becomes “The zebra is eating the grass and the horse is drinking the water”). Their experiments with CLIP lead to better results on their proposed ARO benchmark without hurting the performance of the model.

3.2.4 Visual genome

Visual genome is an effort to combine COCO[LMB+14] and YFCC100M[TSF+16] datasets to create a dense set of descriptions of the scene and present it in a structured representation, denoted as a scene graph representation. The visual genome dataset is able to describe different regions of the image as subgraphs, connecting them to its corresponding region description (bounding boxes) and text description (captions). Their work provides structured data to understand interactions between objects in an image.

3.3 Additional work on reasoning and compositionality

Pre-trained vision and language models are regarded as capable of solving complex tasks, however [PGFC20] suggests that understanding their reasoning and grounding capabilities requires more target investigation on specific phenomena. In their work, they investigate whether VLMs can detect and categorize instances of objects by testing the capability of counting entities in an image in different pre-trained vision and language models. They report sub-optimal performance in all models investigated, even after fine tuning on counting data.

In [LYMP22] it is created a synthetic dataset containing simple objects, such as 3D geometric shapes with different colors, and investigates whether CLIP can bind adjectives to correct objects as opposed to representing the scene as a collection of concepts. They test it with a single composition ("red cube"), multiple ("red cube and blue sphere") composition, and relational composition ("cube behind sphere"). Results demonstrate CLIP performs well on single adjective-noun compositions but it does not achieve good results in the multiple adjective and relational composition tasks.

Our work takes inspiration from the literature reviewed on compositionality and investigates how a Vision Transformers encodes relationships between two objects in an image. Our hypothesis is that there are specific internal parts of the Transformer Encoder that are combining information of multiple objects to represent the composition. We try to use a linear probing method to find such representation where you can decode the composite class of two objects. In the next section we detail the experiments we devised to test our hypothesis and what results we expect from it.

4 Experiments

For the purpose of validating that Vision Transformers indeed encode compositional features we analyze ViT based models that perform different tasks related to images. We review a few neural network architectures applied to different computer vision tasks that require the network to learn the representation of objects. Next, we create a dataset containing images that can represent scenes, such as "**[noun]** *[verb]* **[noun]**" (e.g.: "a woman holding a yellow umbrella"), that can be used for analyzing the internal representation of the transformer encoder. Then we present different methods for selecting the token that might hold information about compositionality, and design a probe to test our hypothesis.

4.1 Network selection

We start by analyzing different architectures that employ Vision Transformer for image encoding that can be suitable for analysis. Works such as [NRK$^+$21] and [MBZ$^+$22] analyze the internal properties of vision transformers with respect to image classification task and shed some light on how patches interact with each other, and also shows how a few tokens are able to be more discriminative than others with regular self-attention[MBZ$^+$22]. [NRK$^+$21] shows ViT can encode information about shape for image classification. We understand a base ViT encoder trained for image classification might not have the capability of identifying multiple objects and, and consequently its relations, therefore we focus on Transformers-based segmentation models.

4.1.1 Instance and semantic segmentation models

Instance segmentation requires predicting pixel-wise class and instance information across an image, making a task where the network has to learn how to encode characteristics of objects with regards to their location, shape, color. Segmention Objects with TRansformers (SOTR)[GNQL21] is a transformer-based instance segmentation model that uses a Feature Pyramid Network (FPN) to extract low-level feature representation and Transformers for capturing long-range dependencies between objects. SEgmentation TRansformer (SETR)[ZLZ$^+$21] on the other hand, relies solely on a pure transformer for encoding the image in a sequence-to-sequence model, achieving state-of-the-art results in semantic segmentation. In general, segmentation models do not expand their capabilities of reasoning too far since the data used to train is composed of images with their locations, plus a limited number of classes in which the model will predict a score. Next, we proceed to multi-modal networks which combine information from two different modalities (e.g.: image-text) to perform tasks such as image captioning and visual question answering.

4.1.2 Image captioning models

Image captioning models, as described in the section 2, combine image and text to learn meaningful representations which can be used to generate a description (caption) that describes the visual content in an image. A ViT-based encoder splits the input image in patches of size 16x16, resulting in 196 tokens, and applies multi-headed attention, so it can learn global contextual information with respect to each patch of the image. This allows us to directly analyze which parts of the image were attended to in the transformer encoder. In some works the image transformer encoder takes as input visual representations of objects produced by a pre-trained object detection model, such as faster R-CNN[RHGS15], therefore the self-attention mechanism does not attend to all regions of the image directly.

That being said, we leverage the fact that a multi-modal network has to represent complex relationships learned from language, so it can output meaningful and readable sentences, with the addition of learning which regions of the image are connected to each part of the text. In our work we use an encoder-decoder transformers-based captioning model with a pre-trained ViT-Base as the Encoder and a pre-trained GPT2[RWC+19] decoder (Fig: 3). A cross-attention module is added to calculate the importance of the output of the Transformer encoder (ViT) in regards to the text processed by the decoder. This setup provides a clean testbed for experimentation and allows us to further investigate the internal representation of the Image encoder (ViT) and hypothesise whether it learns composition to produce its outputs. The model[Kum22] was fine tuned on COCO captions, and its performance scores will be shared in `https://github.com/bodias/vision-transformers-analysis`. Samples of images and predictions done by the model are shared in 17.

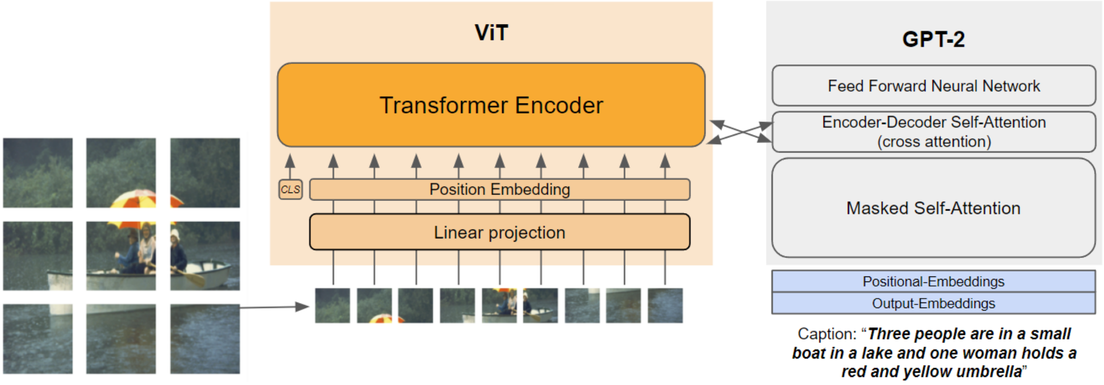

ViT
Transformer Encoder
CLS
Position Embedding
Linear projection
GPT-2
Feed Forward Neural Network
Encoder-Decoder Self-Attention (cross attention)
Masked Self-Attention
Positional-Embeddings
Output-Embeddings
Caption: "Three people are in a small boat in a lake and one woman holds a red and yellow umbrella"

Figure 3. Vision transformers-based image captioning model.

4.2 Dataset

Experimental evaluations are conducted on the COCO dataset[LMB+14] (Common Objects in Context), a large-scale image recognition, segmentation, and captioning dataset. It contains approximately 130,000 images, each annotated with a set of object categories, object segmentation, and captions. Images in the COCO dataset portraits a wide range of common objects and scenes in their natural contexts, and were collected from a variety of sources including Flickr and other public image collections. The dataset is then analyzed using image segmentation annotation with the main goal of understanding which object categories interact with each other (pairs) and how many images are available in the dataset for each pair. Initially, the only restriction applied to the data is a minimum presence of two distinct objects, discarding approximately 24 thousand images in this process. From this analysis of object pairs we come up with two different variations of datasets derived from COCO:

- **Single main and multiple secondary objects**, such as "Person and Accessories", portraying relations in the form of "Person [verb] object", "Person's [object]" or any other relation that involves a "person" with objects in the super category "accessory". We refer to this dataset as *Single object* dataset in our work;

- **Multiple main objects interacting with multiple secondary objects**, such as "Person and Food" and "Dining table and Food", where food is a set of more than one object. We refer to this dataset as *Multiple objects* dataset.

In the following session we describe in detail each dataset and how we plan to use it for testing compositionality in vision transformers.

4.2.1 Single object dataset

The first attempt was to capture the relationship of person and different objects commonly seen with people in images. The majority of the object pairs analyzed contain a person with another object, and several contain at least 1000 images per pair. In the COCO dataset there are supercategories and one of them, named accessory, contains only objects regularly associated with people: backpack, handbag, tie, umbrella. We then form 4 different compositions in the form of "person [interacts with] {backpack, handbag, tie, umbrella}. The main idea is to capture a person plus an object interacting with it. We refer to this subset of data as *Person-Accessory*.

Next, we locate another pair of objects in the main COCO dataset that contains at least 1000 images per pair, excluding images with "person" on it. We observe the object "Dining table" appears frequently with a group of "Kitchenware", such as {bottle, bowl, fork, knife}. We exclude images containing people in order to have a complete disjoint dataset to test and compare with the first one. We form 4 different compositions in the form of “dining table [interacts with] {bottle, bowl, fork, knife}". The main idea is to capture a dining table with an object interacting with it. We refer to this subset of data as *Dining table-Kitchenware*.

In both datasets, we are interested in finding a single token that represents the concepts mapped. Since we have two disjoint sets, Person-Accessory and Dining table-Kitchenware, we are probing them separately. Figure 4 shows examples of images and captions for both datasets generated with only one main object.

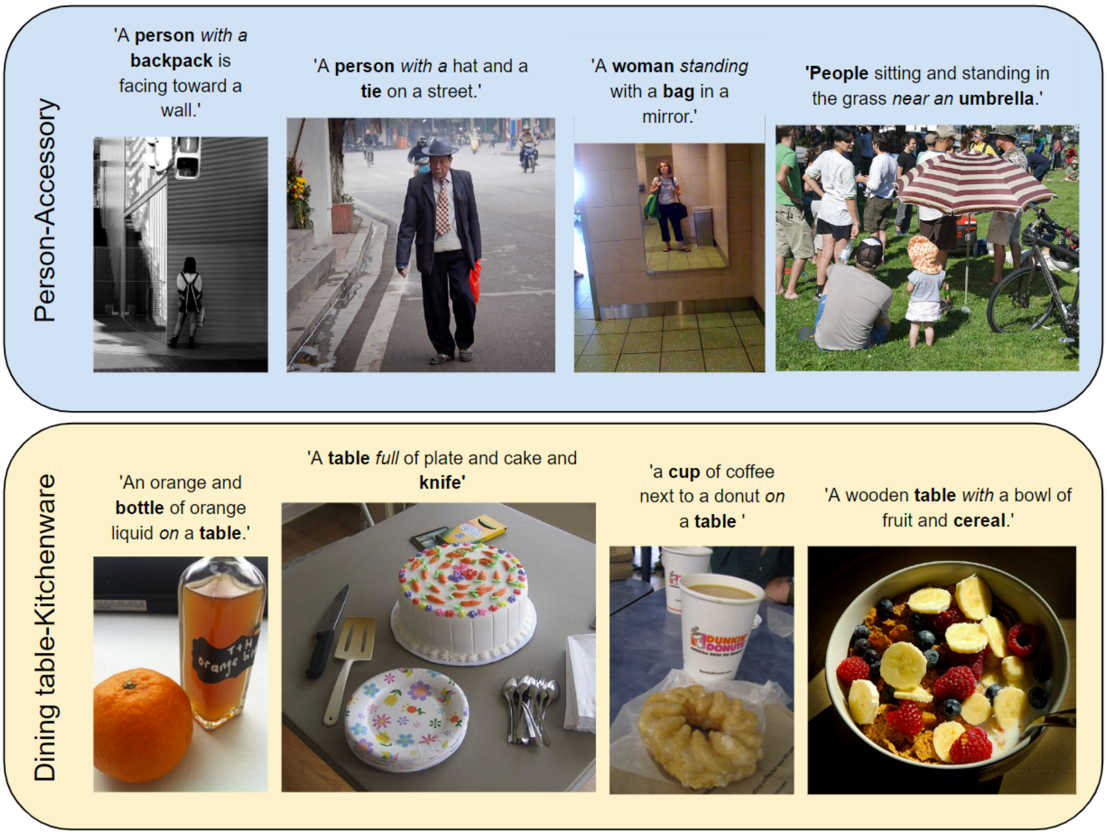

Person-Accessory
'A person with a backpack is facing toward a wall.'
'A person with a hat and a tie on a street.'
'A woman standing with a bag in a mirror.'
'People sitting and standing in the grass near an umbrella.'
Dining table-Kitchenware
'An orange and bottle of orange liquid on a table.'
'A table full of plate and cake and knife'
'a cup of coffee next to a donut on a table '
'A wooden table with a bowl of fruit and cereal.'

Figure 4. Samples of images filtered from COCO dataset to portray the relation *Person-Accessories* (upper) and *Dining table-Kitchenware* (bottom). Caption shown is one of the 5 captions available in the annotations. In bold, the objects we are interested in the image.

4.2.2 Multiple objects dataset

The second dataset should portrait a more complex scenario containing two main objects interacting with two secondary objects. From the pairs in the COCO dataset, a few could represent a meaningful relationship such as “knife and dining table” and “fork and dining table”, however, they are not frequent in the dataset. Another limitation was the fact images should be used only once to represent a relationship. In an image containing “person, backpack and umbrella”, the token selected to represent the pair “person and backpack” cannot be used to represent “person and umbrella”, otherwise the same input will have two different labels in the probe training data. This poses another limitation to the data available to use, leaving a few pairs to be used for testing the cross relation of objects.

After filtering the data to account the issues highlighted before, “person” and “dining table” object classes appear along with “cake” and “pizza” more than 1000 times, leading to 4 different compositions: “dining table and cake”, “dining table and pizza”, “person and pizza”, “person and cake”. We refer to this subset of the data as *Dining table/Person-Food* for simplicity. This setting allows us to choose different tokens from each object knowing there is more than one combination for it to occur. In other words, previous compositions may lead to the conclusion the token from the second object (accessory and kitchen appliances) is only encoding information from itself. We depict the relation of the objects under the settings above in Figure 5.

4.2.3 Caption based dataset filtering

Initially, we filtered images from COCO taking into consideration only the *category_id* from object segmentation annotations without using captions provided in the dataset. For example, we assign as Person-Umbrella images where both objects are shown in the image. Captions in COCO dataset are short texts that describe the image, and often would mention objects that are relevant to the scene portrayed in it. Since the model we are investigating was fine tuned on COCO data, we assume samples where the caption mentions a word that describes the object might lead to a better internal representation. The filter also helps remove images where objects are very small, partially occluded, or in the background.

With that in mind, a simple NLP pipeline is devised to extract nouns from captions and check which objects are present in the text. The pipeline consists of tokenization, followed by part-of-speech tagging and lemmatization, extracting only the tokens tagged as nouns. This results in a list of nouns per image, which we can use to detect different words that have the same meaning. For instance, person could be mentioned in the caption as [’man’, ’woman’, ’person’, ’girl’, ’boy’, ’couple’]. More sophisticated approaches could have been applied, such as similarity metric in an embedding space using pre-trained embeddings, instead of manually inspecting the occurrence of nouns per pair of objects.

We match the processed object names with the object category name from COCO to obtain a filtered version of our datasets. We later report in our results the effect of this filter in the decoding accuracy. The difference of both filtered and non-filtered datasets is

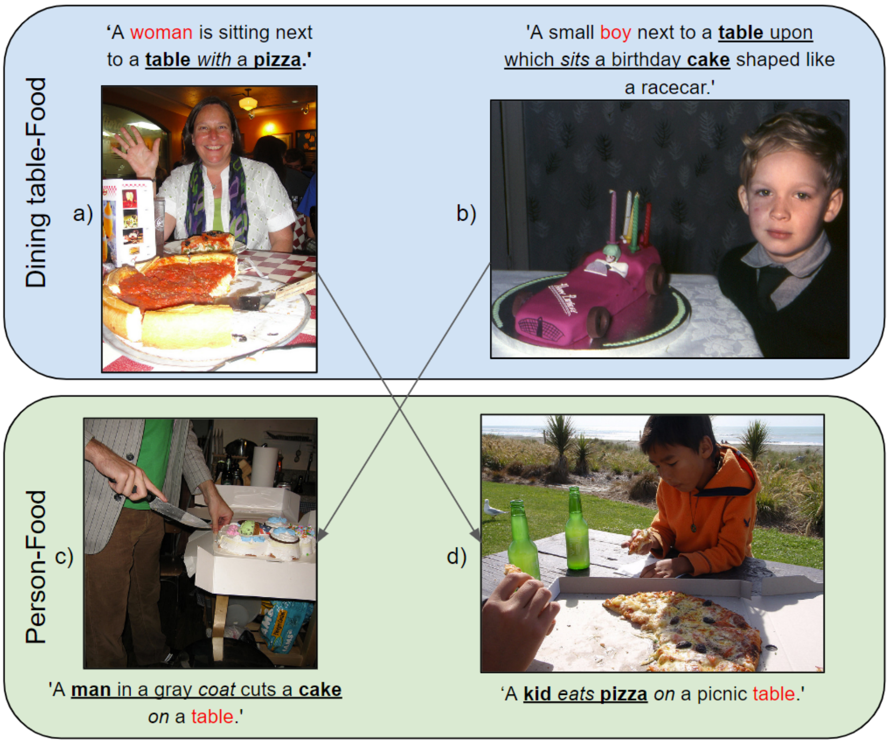

Dining table-Food
Person-Food
a)
b)
c)
d)
'A woman is sitting next to a table with a pizza.'
'A small boy next to a table upon which sits a birthday cake shaped like a racecar.'
'A man in a gray coat cuts a cake on a table.'
'A kid eats pizza on a picnic table.'

Figure 5. Multiple objects setting poses a more complex scenario where the same image might contain the encoded representation of multiple pairs of objects. For example, the image on the top-right (b) contains a person next to a cake, but also a dining table with a cake. In our results we analyze if one of the concepts is indeed encoded internally in the vision transformer. Underlined portions of the captions denote the relation the image is assigned to. We collect data from Person-Food and Dining-Table food in a manner the same image is not used in both sets at the same time.

exemplified in Figure 6.

4.3 Feature selection

Once datasets have been defined, we proceed selecting features which will represent the composition of two objects, and use them to train a simple neural network to act as a probe. From COCO annotations we have the ground truth segmentation which represents exactly where the object is located in the image. We use them to identify which image patches fed into the encoder overlaps with the location of the object being analyzed, and later use attention maps to define which tokens can be used to represent each object.

In Vision Transformers, attention maps refer to the weights applied to the different

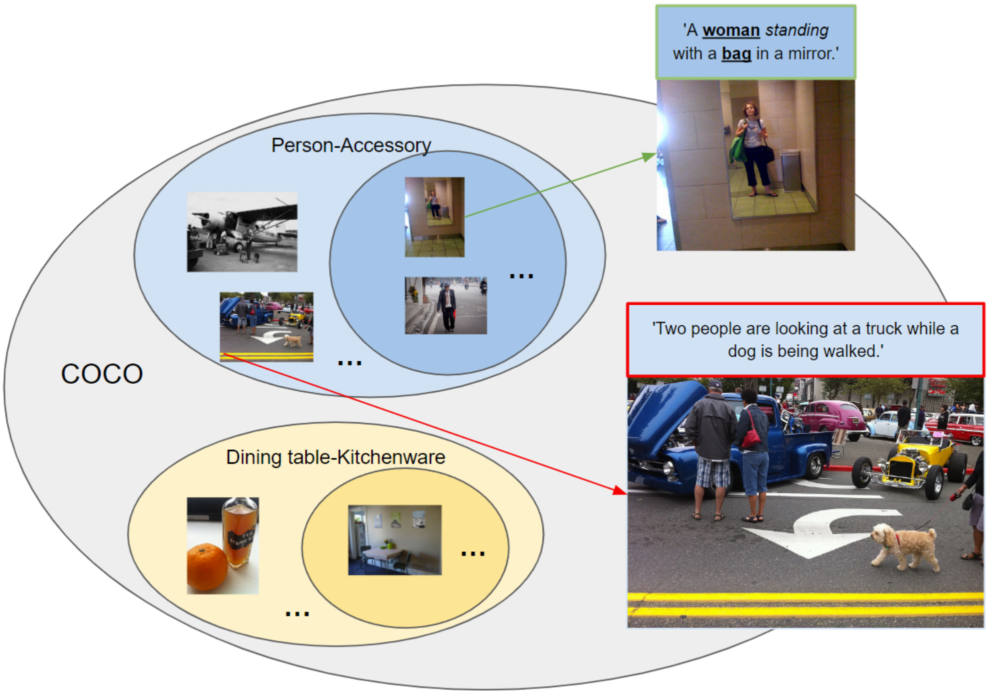

'A woman standing with a bag in a mirror.'
Person-Accessory
'Two people are looking at a truck while a dog is being walked.'
COCO
Dining table-Kitchenware
...

Figure 6. Different subsets of data. From the COCO dataset we subset images containing objects belonging to specific classes using *category_id* (i.e.: id of the object segmentation annotation). Then, we process the caption and filter images where the category name of objects match to the caption. Zoomed in on the right side are samples from the subset Person-Accessory, more precisely Person-Handbag. On the top (in green) is an example of an image filtered by caption. On the bottom an image where the caption does not mention handbag (or any variation, such as “bag”). We experiment with different datasets to understand their effect in our results.

patches of an input image (**token**) by the self-attention mechanism of the transformer, as described in section 2. During the self-attention process, the transformer calculates a set of attention weights for each patch in the input image, indicating how important each one is for generating the output features of the transformer. These attention weights can be visualized as a heatmap over the input image, highlighting the regions that are most relevant for the task the model was trained for. Attention maps are important in Vision Transformers because they enable us to interpret the internals of the transformer and provide insight into how it processes visual information.

Our work analyzes attention maps of the Vision transformed-based encoder used in the caption model to extract relevant information that later can be used for probing. We call feature extraction pipeline the sequence of the following steps:

Object instance selection: Ensure we are choosing a relevant instance of the object in the scene, in case there are many. Relevant token selection and extraction: In the subset of tokens belonging to the objects analyzed, choose a few of them to be used to represent the composition.

We used a computer with a 3.6GHz processor and 32GB of RAM, paired with a Nvidia RTX 3070 with 8GB of VRAM to run parts of the code on GPU, mainly the captioning model inference. Running all combinations takes approximately 4:45 hours with the aforementioned hardware, with one of the limiting factors being the size of the dataset has to fit into the 32GB of RAM available. Next we detail all the parts of the feature extraction pipeline.

4.3.1 Object instance selection

A common characteristic of the COCO dataset is that images might contain several instances from the same object (i.e.: several people in the image), leading to problems when selecting exactly which object should be used for feature extraction. Figure 7 exemplify this issue. Next we define three methods to select the correct object instance.

- **Largest object**: Considers the area of the object and chooses the largest one. Simple images with few objects tend to have the object focus of the scene as the largest. However, complex images with several objects might get confused between background and relevant objects. This approach takes advantage of how captions describe objects in the COCO dataset, where they tend to highlight a few salient objects and omit other parts. The impact of this fact in the analysis is explored in the discussion section;

- **Closest** : A variation of the previous one, choosing the second object from the composition (i.e.: umbrella, for the pair “person and umbrella”) based on euclidean distance from the centroid of the largest object. Considering the object pairs we chose, the second object should, in most of the cases, be closer to the main object to portrait the compositions we are looking for;

- **Attention based**: Selects the instance which received maximum attention in layer n, with $n = [1, ..12]$, among all the occurrences of the object. Transformer-based encoders contain several layers, where the input flows from layer 1 to layer 12 (ViT-B). Since we probe in each layer individually, the instance of the objects which gets maximum attention in that layer represents the most important based on self-attention.

In our experiments we combine method 1 and 2 and denote it non-attentive. We choose the largest main object (person and dining table in our datasets) and its closest second object. Figure 7 shows the main difference in these object selection approaches in complex images.

4.3.2 Relevant token selection and extraction

Given an input image, we are presented with several objects and we know their exact location from the segmentation annotation. In the previous step, we choose two objects portrayed in the image that represent a composition. Next, we feed the image into the captioning model to extract its attention matrices and hidden states per layer. The standard ViT-Base (ViT-B) encoder used contains 12 layers and 12 attention heads, hidden dimension size is 768. The attention matrices have size 197x197 (196 images tokens plus *CLS* token) containing the calculated attention for the given input, in each layer and each head, resulting in a tensor size (12, 12, 197, 197). We average the attention matrices across all heads[DBK$^+$20] since we are not interested in the particular behaviour of different heads. We analyze the attention the *CLS* token received with respect to all patches in the image. In the end we have attention maps which will highlight which tokens are receiving more attention.

We initiate the process of selecting the internal representation we are probing by using the location of the objects to produce a boolean mask that will highlight only the objects being analyzed. This mask is applied on top of the attention matrices to ensure we analyze only the attention received by tokens from the objects in the composition. In this step objects that are close to each other might share the same token when analyzing the original area of the objects. Because of this, we order the objects by size and tokens assigned to the main object won't be used in the second. We then have the attention each object received, in each layer of the encoder, where we can use it to extract the corresponding internal hidden state. The hidden state at token i represents a state where the input was processed by a Transformer block, at the selected token. This is the internal representation we probe to test compositionality.

For training the supervised learning model we use as a probe, different tokens from the object are selected to compare their performance:

- Minimum attention the object received based on *CLS* token, referred as *Min. object* for simplicity;
- maximum attention the object received based on *CLS* token, referred as *Max. object* for simplicity;
- Random token selected from the object, referred as *Random object* for simplicity;
- Token which received maximum attention in the whole **image** (based on *CLS*), referred as *Max. image* for simplicity.

Each one of these tokens is extracted from different layers and later used in the decoding task.

We extract features vector from images using three different datasets (as defined in 4.1.1 and 4.1.2), with and without caption filter, for each one we selected an object from the composition pair (main, secondary), applied different object instance selection techniques (Non-attentive, Attention based) and four different methods to select the internal representation of the transformer encoder (hidden state) that will be used to represent the compositionality. We visually represent the steps in Figure 7. We run it for all layers in the encoder to test whether the compositionality emerges from different layers. This generated 3x2x2x2x4x12=1152 combinations to be tested, each one producing one representation (feature vector) that we will probe. We train 1152 neural networks to decode the compositions of objects from extracted feature vectors.

4.4 Decoding ANN/Probes

To test whether tokens from objects bind concepts, we use both single and multiple objects datasets explained in section 4.1. The single main object with multiple second objects is a simpler task where tokens from the main object will be used to decode the second object, i.e the token from person can be used to decode the accessories it relates to. In the multiple main objects and multiple second objects the setup is more complex, now pairing two objects, Person and Dining table, with other two objects, cake and pizza. The hypothesis is that there is one token binding Person plus Accessory, or Person and Pizza, representing the transformer encoder has learned the composition between these two objects. The images we analyze contain the representation of a scene, for example, “A woman eats pizza in a restaurant” (Person eats Pizza), “A man with an umbrella is walking in the street” (Person with Umbrella) and we hypothesise whether the transformer encoder is learning the composition of these two nouns internally.

Inspired by [HL19] we try to validate our hypothesis by training supervised models to predict a property (like compositionality) from the internal representations, a method called probing. Our neural network is trained with an internal representation of the objects (feature vectors extracted from tokens) and a label depicting the relation of the main and secondary object (i.e.: Person-Pizza)). As described earlier in section 4.2, we run the feature extraction pipeline on the COCO dataset to extract internal representations of each object pair tested, with different configurations. We later use them as training data for our decoding neural network.

The architecture used for probing is a Feed-Forward Neural Network with one hidden layer with 256 neurons followed by a dropout layer with $p = 0.5$. A shallow FFNN is used to ensure that the desciminative power of the probe originated from the token's representation. A shallow network cannot model complex relationships between the token and class while training the probe. All models are trained using SGD with Nesterov momentum, with $learningrate = 0.0001$, $momentum = 0.9$ and categorical crossentropy loss. The network takes as input a feature vector with $D = 768$, a token from the encoder hidden states at $layerN$, and outputs the object pair using softmax activation. All models trained are a 4-class classifier, since all tasks, Person-Accessory, Dining table-Kitchenware and Person/Dining table-Food contain four different object

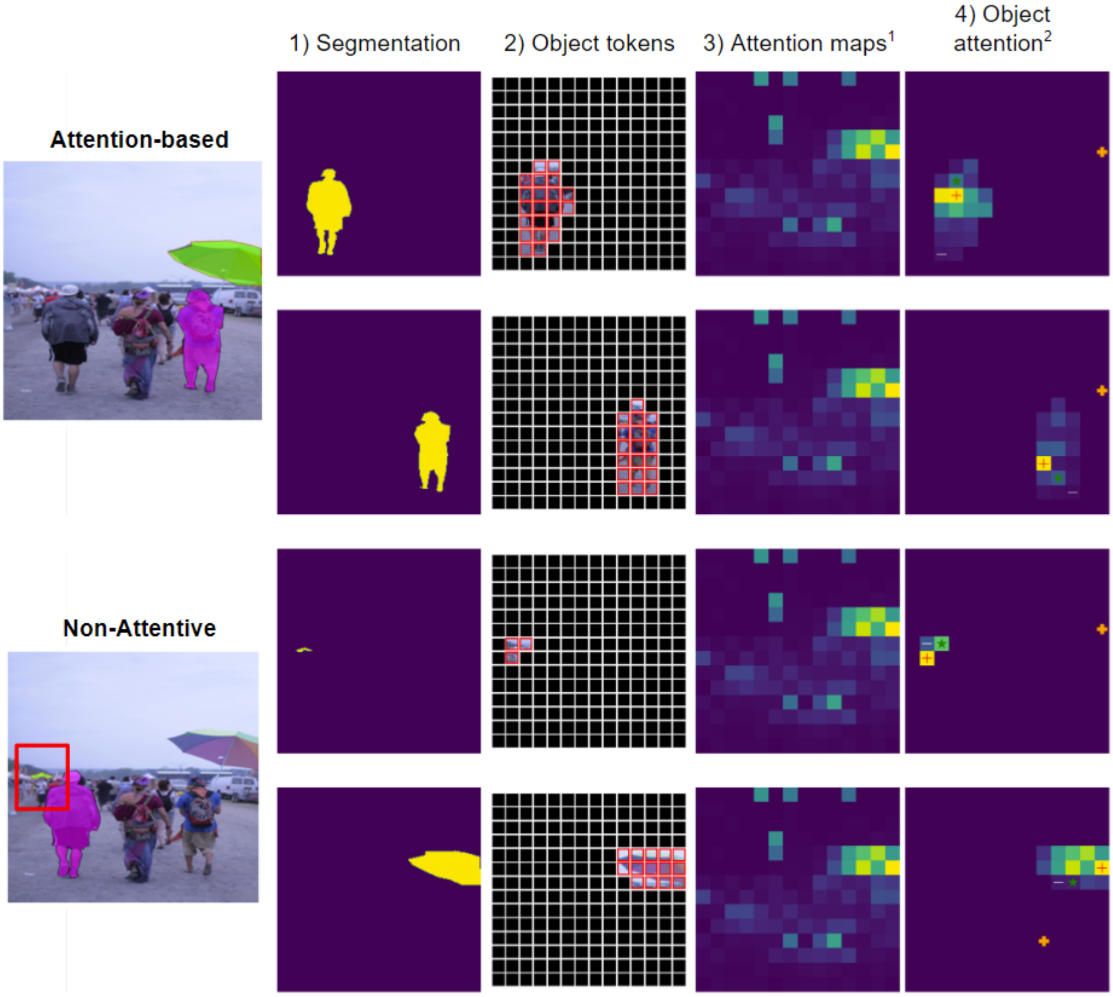


Figure 7. Feature extraction pipeline for an image containing the composition (person, umbrella), using different methods for object instance selection (Non-Attentive, Attention-based). The box in red highlights the object selected in the background. On the last column, signs denote maximum image attention (+, orange), maximum object attention (+, red), minimum attention ($-$, white) and random object token ($*$, green). Better visualized with color. [1]Attention map of layer 12 of the encoder. [2]Object attention is normalized and colors won't match column 3.

pairs. We trained all models for 200 epochs with early stopping, allowing the training process to halt if the validation loss has not improved after 20 consecutive epochs.

Data to train the supervised models to act as probes are sourced from COCO dataset, and split into three subsets: Person-Accessory, Dining table-Kitchenware and Person-/Dining table-Food. Each subset also contains a variant where we filter images where caption and segmented objects match, as described in section 4.1.3. We use all these datasets to train separate decoding neural networks and report results obtained in section 5.

Train and test split are based on COCO's train and validation data, where the train split from COCO is used for train and validation (10%), and the whole original validation split from COCO is used as test split. We test different settings to find evidence there is one single token (token from object with maximum, minimum attention, random token from object and maximum attention token from the entire image) that encodes the relation of two objects. More details about all different configurations tested is described in sections 4.2.1 and 4.2.2. The result is 1152 neural networks for each one of the configurations we tested. Next, we analyze the accuracy of all probes and report the differences seen in all tested configurations.

5 Results

This section explores the result of the analysis on the vision Transformer-based captioning model and provides more details about the series of probes used to verify if they can encode compositionality internally. We first report results for each subset of the COCO dataset used (section 4.1), since they depict different objects and compositions. Then we show the impact of different feature selection strategies: non-attentive and attention-based object selection (section 4.2.1) and different tokens (section 4.2.2). We also report the impact of the caption based dataset filter (section 4.1.3) in the results.

5.1 Single object datasets

In this setting two objects are paired together and the main goal is to decode the pair (main object, secondary object) from different tokens obtained from objects through the Transformer encoder. As described in section 4.1, after obtaining the subset of data portraying different compositions, two different datasets are created:

- **Person-Accessory**: Can we decode the compositions formed by Person and Accessory with a single token extracted from the Transformer encoder? Where $Accessory = \{backpack, handbag, tie, umbrella\}$ and $Person = \{person\}$. Images in this subset portray scenes that can be described as “Person [action] object”, “Person’s [object]” or any other relation that involves a “person” with objects in the super category “accessory”. For example, the following description (caption) would be part of the dataset: “a little girl holds an umbrella”.

- **Dinner table-Kitchenware**: Can we decode the compositions formed by Dining table and Kitchenware with a single token extracted from the Transformer encoder? Where $Kitchenware = \{bottle, bowl, cup, knife\}$ and $Diningtable = \{diningtable\}$. Images in this subset portray scenes that can be described as “object [on] Dinner table” or any other relation that involves a “Dinner table” with objects in the subset “Kitchenware”. For example, the following description (caption) would be part of the dataset: “a birthday cake in the shape of a car on top of a table”.

Since there is only one main object per dataset, decoding the secondary object using information from the main one could potentially show signs of compositionality. Table 1 presents the best results achieved by probes in both datasets, grouped by token selection strategy and the object source of the token. With tokens extracted from the main object (Person or Dining table), we can decode the pairs in Person-Accessory with 60.6%, 66.1% and 62.1% accuracy using, respectively, minimum attention, maximum attention and random token. With the objects pair Dining table-Kitchenware the accuracy dropped to 39%, 46.9% and 42.8% under the same settings.

The decoding FFNN in both datasets performed a 4-class classification problem, therefore the random guess would yield 25% accuracy. Despite results being above the random guess baseline, it shows almost 20% in decoding accuracy between both Single object datasets. In section 6 we discuss the main differences observed and highlight potential reasons for this disagreement between results in these datasets. Figure 8 highlights average decoding accuracy for all three datasets, in regards to the object source (main or secondary object), hidden layer and token selected for representation. Individual scores for each probe are reported in the appendix Figures 19 and 18. In general, accuracy increases using representations from higher layers of the Vision Transformer encoder, and all best results reported in Table 1 are obtained from Encoder layers 9 to 11.

Results from decoding the composition from the secondary object might represent the decoding of the object only, since the information is taken from one of the four objects depicted in the 4 classes. If we were to set up a decoding task of the objects in dataset Person-Accessory ({backpack, handbag, tie, umbrella}) it would yield the same results. In the next section we present the results of the Multiple objects dataset that should present a more complex scenario for the decoding task.

Table 1. Best decoding accuracy by dataset and token used for representation. Number in parenthesis denotes the layer where the best score was found.

Dataset	**Max. image**	**Max. object**	**Min. object**	**Random obj.**
Person-Acc.-main	0.512 (11)	0.606 (11)	**0.661 (11)**	0.621 (11)
Person-Acc.-second	0.507 (11)	0.693 (10)	0.698 (9)	**0.725 (10)**
Person-Acc. (mean)	0.51	0.65	**0.679**	0.673
D. table-Kitchen-main	0.352 (11)	0.39 (11)	**0.469 (11)**	0.428 (11)
D. table-Kitchen-second	0.352 (11)	0.585 (10)	0.601 (9)	**0.67 (10)**
D. table-Kitchen (mean)	0.352	0.487	0.535	**0.549**
Person/table-Food-main	0.523 (11)	0.594 (11)	**0.822 (11)**	0.803 (11)
Person/table-Food-second	0.485 (11)	0.67 (9)	**0.707 (11)**	0.697 (11)
Person/table-Food (mean)	0.504	0.632	**0.765**	0.75

5.2 Multiple objects dataset

In the Single objects dataset we present one main object interacting with four secondary objects, and we decode different tokens to test compositionality. The two datasets used, Person-Accessory and Dining table-Kitchenware represent two disjoint sets of data, therefore results are shown separately in the previous section. In the Multiple objects dataset we arrange two main objects interacting with two secondary objects, as depicted in Figure 5(Section 4.1.2). In this configuration, the result from decoding the main object from the second, and vice versa, is more meaningful since objects “Person” and “Dining table" appear with the same objects "Cake" and "Pizza". The dataset Person/Dining table-Food represents the following compositions:

- **Person-Pizza, Person-Cake, Dining table-Pizza, Dining table-Cake**: Can we decode the compositions from a single token extracted from the Transformer encoder? In this dataset we have examples of the same secondary object (i.e.: pizza, cake) appearing with "person" and "dining table".

Results obtained by decoding the compositions using the token from the main object (Person/Dining table) are higher than the second object, however the difference between best scores is close to 10%. Table 1 reports best decoding accuracies achieved at 81.7% and 69.7%, for the main object (Person/Dining table) and secondary (Pizza/Cake) respectively. Figure 8 highlights average decoding accuracy for this dataset, in regards to the object source (main or secondary object), hidden layer and token selected for representation. Individual scores for each probe are reported in the appendix in Figures 19 and 18. It is possible to see the influence of different layers, where deeper layers lead to better decoding accuracy. The performance of the probes in the Multiple objects dataset is significantly higher than the random guess baseline (25%).

We can make the assumption that by achieving high evaluation accuracy in predicting the composition (e.g.: "Person with backpack") from the representation (i.e.: token from the captioning model) implies the composition was encoded in the representation, and the probe found it. In both types of compositionality we tested, with single and multiple objects, it is noticeable that deeper layers are reporting better decoding accuracy. This is consistent with the findings in the original ViT work[DBK+20] that says deeper layers attend to most of the image and might contain more information than lower layers. In this study all tokens we used to test compositionality constantly report better decoding accuracy as we go deeper in the layers of the Transformer encoder. The effect of using different layers for selecting different outputs of Transformers block (i.e.: internal representation) can be seen in Figure 9. In Figure 10 we see the individual decoding accuracy of all 1152 Neural Networks is spread over a wide range of values. In the next section we explain the impact of different features used for probing.

5.3 Impact of object selection strategy and caption filter

Images in the COCO dataset will often have several occurrences of the same object and one of our premises is to use a single token to represent the combination of objects. In section 4.2.1 we explain the different object selection strategies implemented in this work to choose the most likely one the captioning model used to generate the image description. We also have alternative datasets filtered by caption to test whether this filter results in images that better represent the composition.

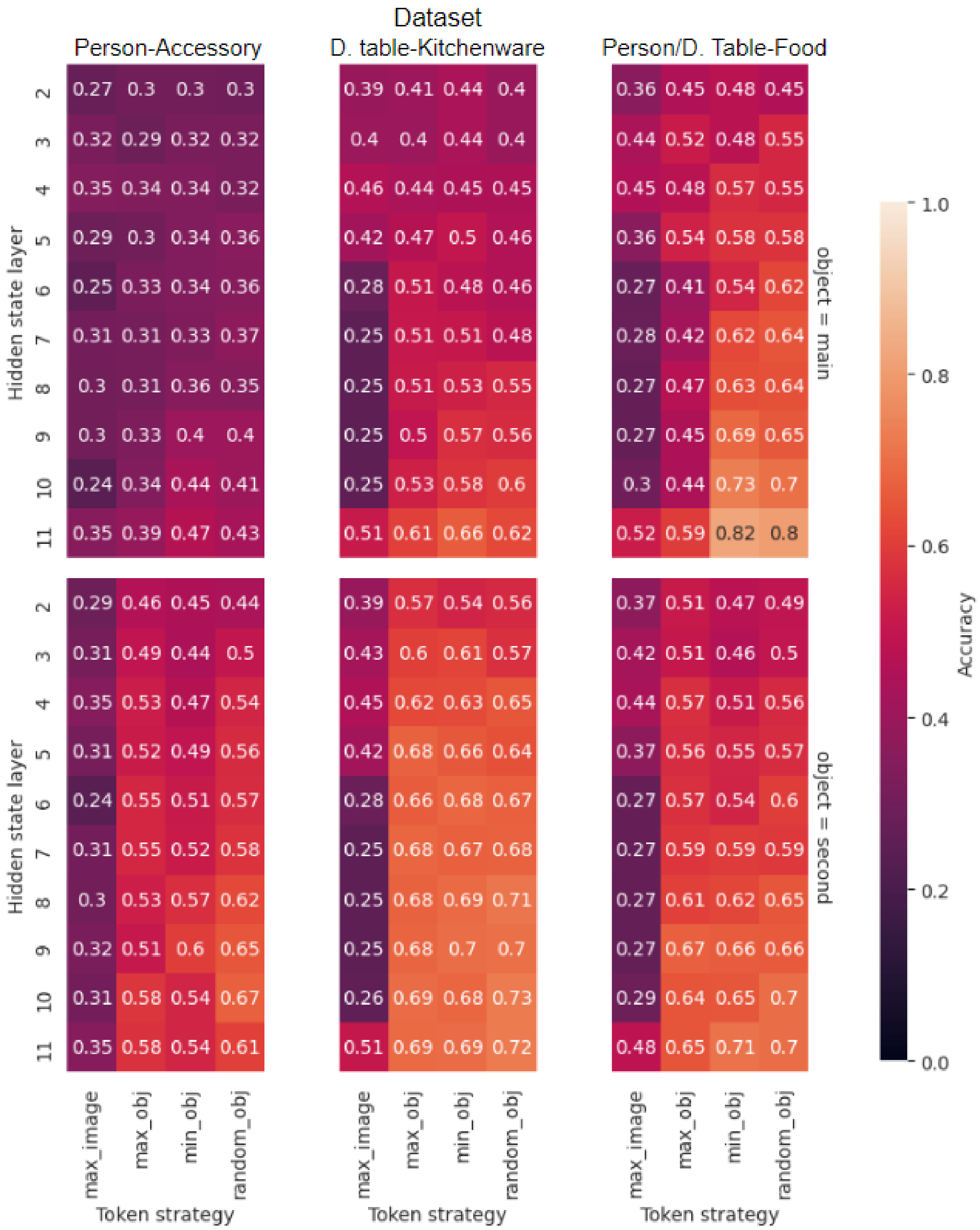


Figure 8. Mean decoding accuracy scores for all datasets, grouped by main and secondary object, transfomer hidden state layer and token selection strategy.

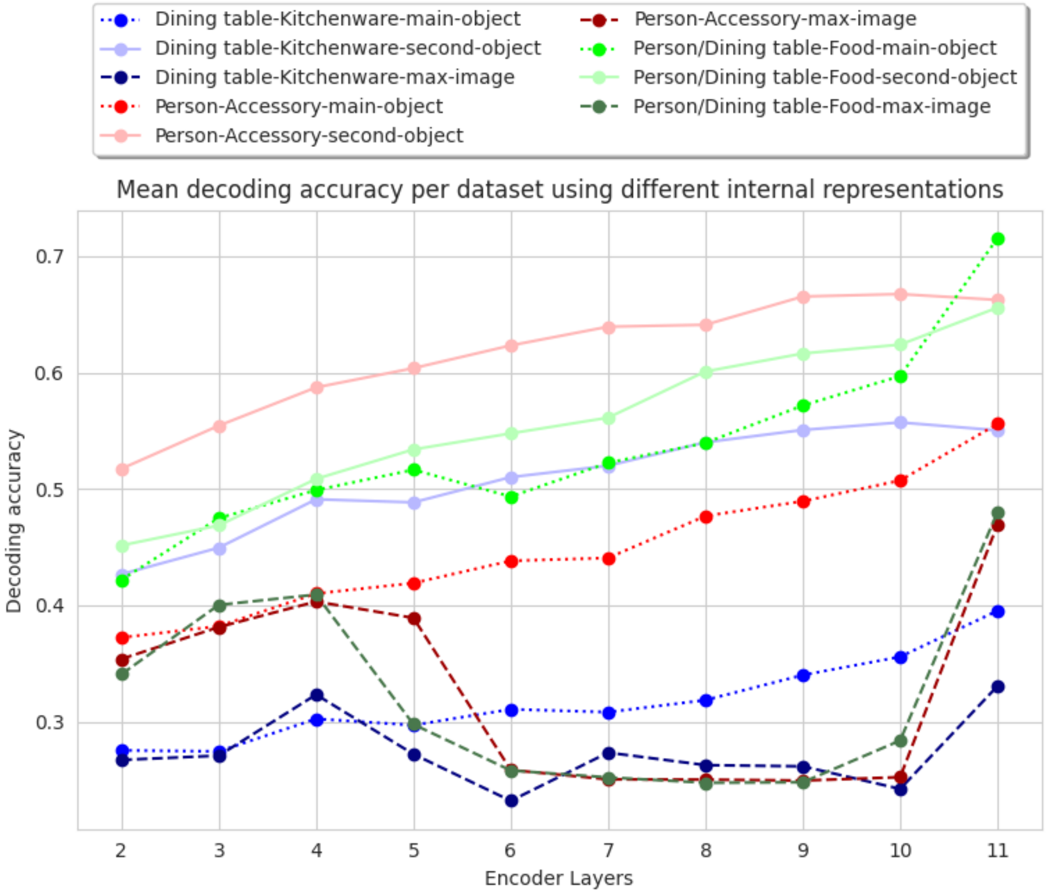


Figure 9. Mean decoding accuracy per layer across all experiments.

5.3.1 Caption filtered datasets

In section 4.1 we explain how the subset of COCO was created by selecting images with pairs of objects based on their *category_id*. This method ensures both objects in the pair are present in the image, however it does not guarantee they are relevant for image captioning, as they might be small, partially ocluded, or totally irrelevant for describing the image. We apply the caption filter in each one of the three datasets we use and test whether it impacts the results. Section 4.1.3 and Figure 6 explains in more detail how the caption filter was developed and applied. In Table 2 we show the number of images sourced from the COCO dataset and how many can be filtered by caption.

In Figure 11 we observe that employing the caption filter in the dataset leads to better decoding accuracy in the majority of the datasets on average. This might indicate

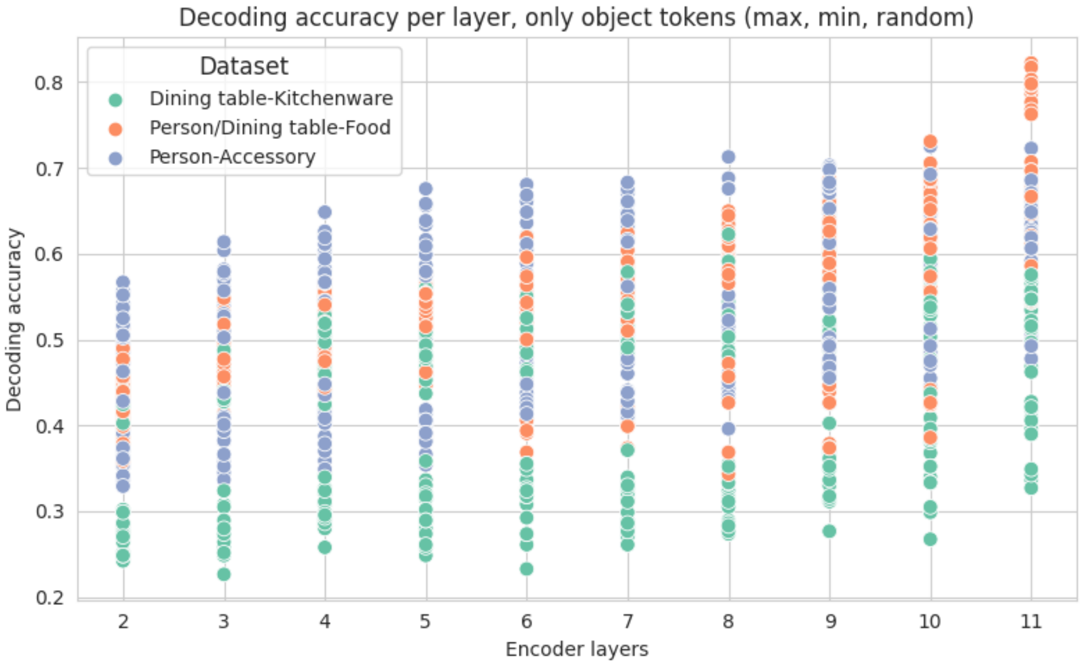


Figure 10. Decoding accuracy per layer, only object tokens.

the quality of the dataset influenced the probe's results. We present the data split by object selection strategy to observe whether it impacts the results differently. We discuss different aspects of the data in section 6.

5.3.2 Object selection strategy

In Figure 12 we see the decoding accuracy difference between the attention-based and non-attentive object selection methods. A positive difference means the attention-based selection led to higher decoding accuracy in the probe. The median differences are close to 0 in all tests. In fact, most of the values in the interquartile range are negative, indicating using the non-attentive selection is better in these settings. We present results split by dataset and caption filter (with or without filter) to investigate whether the quality of the data may impact the results.

5.4 Additional Results

In this section we report additional findings related to our analysis, but are not fundamental.

Table 2. Number of images available for each dataset and the share of images where caption matches objects in the scene

Dataset	Total images in the subset	Images filtered by caption
Person-Acessory	14829	4492 (30%)
Dining table-Kitchenware	7933	2252 (28%)
Person/Dining table-Food	4490	3131 (69%)

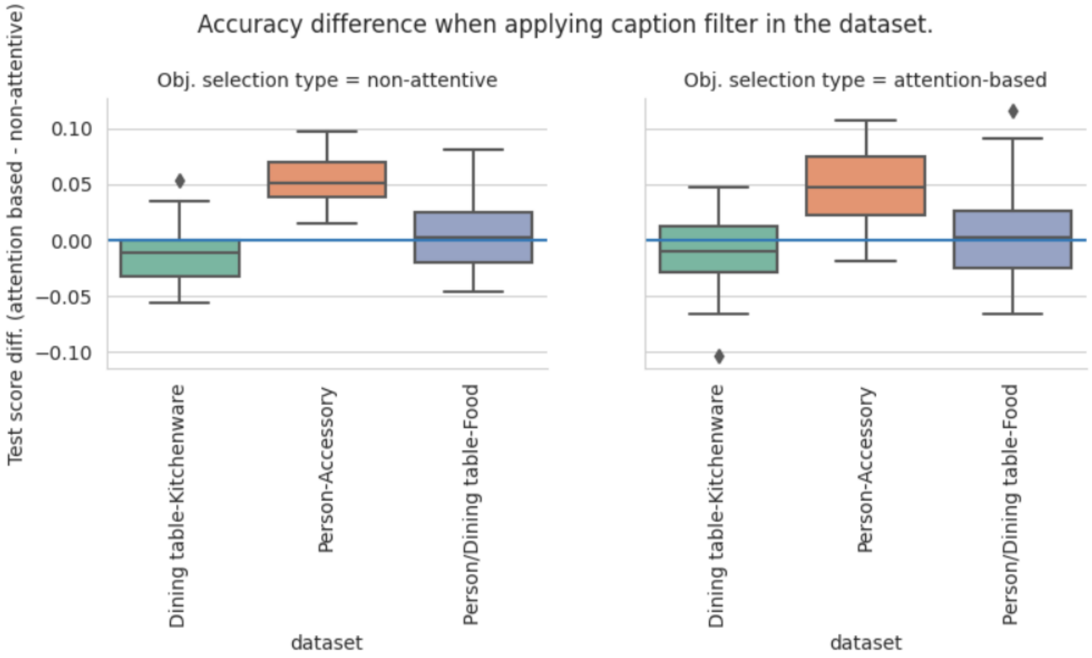


Figure 11. Accuracy difference using caption filter in the dataset

5.4.1 Decoding Network learning curves

Learning curves from a neural network can offer valuable insights into the performance and behavior of the network during training. One of the insights we can obtain is how quickly the network is able to learn patterns in the data. A fast convergence rate indicates that the network is able to learn quickly and the data contain useful features, while a slow convergence rate may imply the model is not expressive enough or that the data contains noise or irrelevant features to describe data.

Our work tested a total of 1152 neural networks to decode compositionality from selected parts of the hidden state of a transformer encoder. All these models, trained with settings as described in section 4.3, produced different learning curves yet they are similar in their shape if we aggregate the information by token strategy ([max image, max

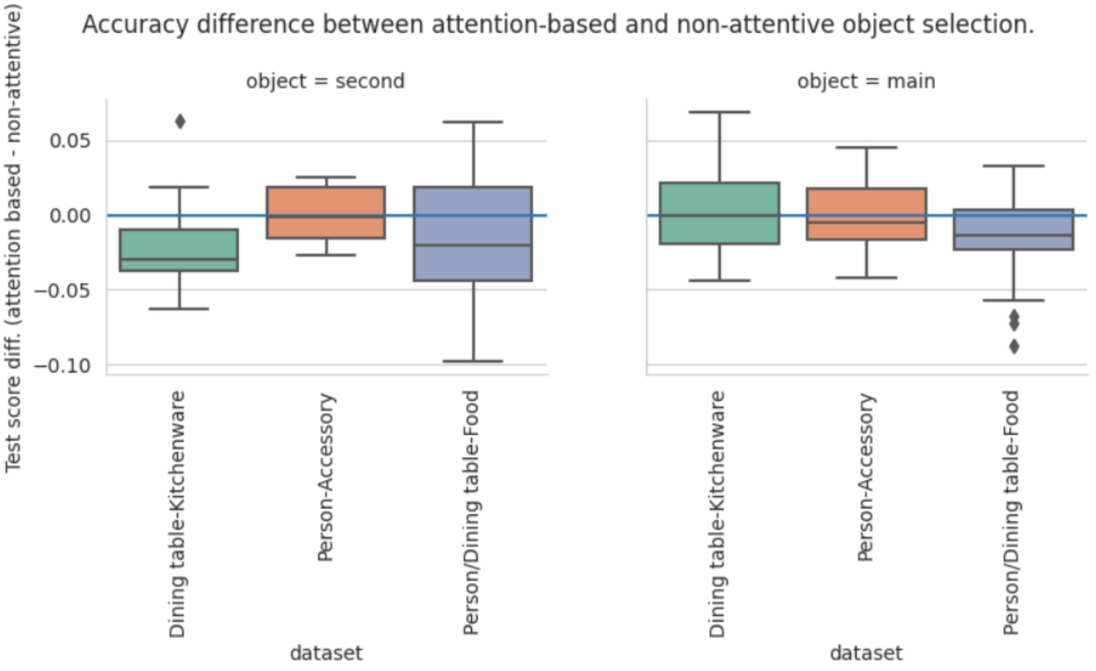


Figure 12. Accuracy difference using different strategies for object instance selection

object, min object, random object]) and encoder layer used. The internal representation extracted from the token which received maximum attention in the whole image shows a steep loss curve in most of the layers, and validation accuracy over the epochs varies significantly, resulting in a noisy chart (Figure 13). Layers 7 to 11, using the same token selection strategy, also report similar behaviour. However, training stops after 20-50 epochs due to the early stopping used to halt the training if no improvement in loss is achieved after 20 epochs. For feature vectors extracted from object regions the behavior of loss curves are similar to an exponential distribution curve, which is expected from the training regime applied. This resonates with accuracies reported in Figure 8 where darker regions are seen in maximum image token from layers 7 to 11.

5.4.2 Datasets filtered by caption

As previously mentioned in section 4.1, the captioning model has been fine tuned on COCO captions dataset and we expect it to be grounded, attending to the right objects to generate the caption words. Later, in section 6, we discuss the effect of it considering the fact captions in the COCO dataset are highlighting a few salient objects when describing the scene, therefore having a pair Person-object in the image does not necessarily mean these two objects received high attention. In Figure 15 and 14 we show the differences in

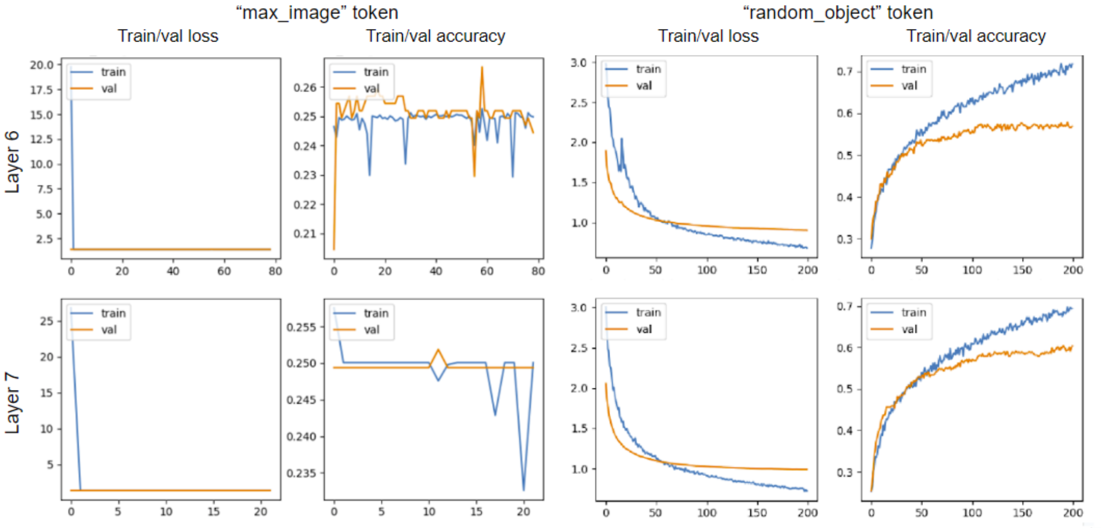
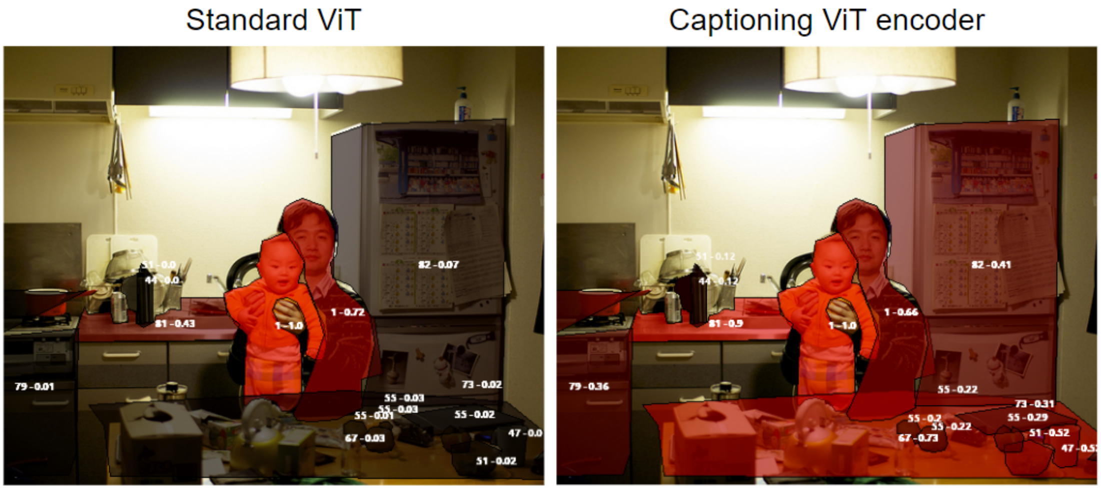


Figure 13. Learning curves for selected layers and token selection strategy. We see the network is not able to learn from "max_image" token.

attention of a standard ViT encoder, and the ViT encoder finetunned in the captioning model, where can see table is a large object, however the normalized attention it gets from the captioning model is low (0.03) compared to the attention received in the standard ViT (0.73). Ground truth captions from COCO mentions only “kitchen”, “man” and “baby”. The caption filtering process described in 4.1.3 tries to clean the dataset and prevent images where the selected pair of objects does not contribute to the scene to be used in the probe.

The object instance selection based on the largest object yield decent results compared to the attention based selection due to the same fact that often the largest object in the image is the one mentioned in the caption, and therefore more likely the same object has received high attention.



Caption: 'a man is holding up a baby while he is in the kitchen'

Figure 14. Image highlighting different attention patterns using different encoders. On the standard ViT the attention is more evenly distributed among objects, such as fridge, table, sink (intensity of red denotes more attention). In the captioning model the attention goes mainly to the objects mentioned in the caption (person).

6 Discussion

One of our main hypotheses was whether the Vision encoder unsupervisedly developed a notion of compositionality and encoded this notion in its internal representations in order to perform the task it was trained on. Our work explores how Transformers-based vision and language models, particularly a captioning model, represents concepts internally and tries to answer whether it is possible to decode compositionality of two objects from a single token. Results presented in section 5 show decoding accuracy is better than chance, hinting that it is possible to weakly decode a composition of visual features extracted from the Transformer encoder. One important factor is that the process we devise to select images to train the decoding neural network might produce noisy data, where the pair of objects is not visible, or not relevant. This can be seen in our results as different datasets report different numbers. Our results expose a series of different factors that can influence our probe, such as dataset, object instance selection, and choice of token to represent the composition. This analysis shows there are signs of compositionality, but it does not confirm the hypothesis we can use a single token to decode it; it merely reopens the question.

In our experiments, feature vectors extracted from deeper layers of the encoder can be used to decode pairs of objects better than random guess, at higher than 70% accuracy

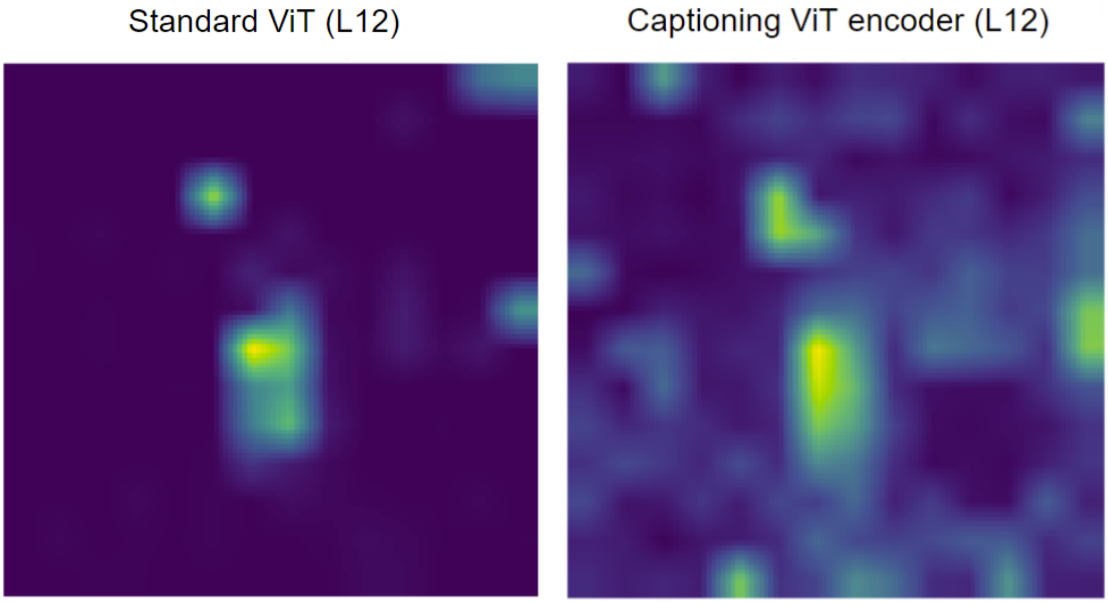


Figure 15. Attention maps (layer 12) comparing a standard ViT without fine tuning and the ViT encoder trained on image captioning task. The attention is based on the same image as in 14

in a few probes in Person-Accessory dataset and the multi pairing of Person/Dining table-food. However, the setup with a Single object performed worse when compared to the Multiple objects dataset. In theory, the Single one is a simpler task where we decode a series of 4 different objects connected with the same main object. For example, decoding the class person-umbrella from a token selected from the object person should indicate some information is being aggregated outside of the object umbrella to represent both. Our results show better decoding accuracy when using datasets with Multiple objects correlated. We achieve 80% average accuracy when decoding Person-Food and Dining table-food from the main object (Person or Dining table), However, these two results are not comparable and only help us understand that under different circumstances we are still able to partially decode information. We further discuss the decoding capabilities and limitations in section 6.1.

It also appears decoding accuracy improves progressively as the information reaches deeper layers, indicating information from two objects is binding in latter stages of Transformers (Figure 9). We also find that multiple tokens from the transformer encoder are reporting similar decoding accuracies. In our experiments, a token randomly sampled from the object region performs similarly to the token which received maximum attention as we can see in Figure 8. However, it is not possible to accurately state such networks are capable of binding concepts under the setting used alone. Further investigation including larger models, such as [RKH$^+$21], might enrich the analysis conducted in our work and provide significant evidence Vision Transformers can represent complex concepts internally.

Our work differs from other related analysis reported in section 3 by looking deeper at internal representation of Vision transformers, layer by layer, under different settings. Nonetheless, our work shed some light on the understanding of compositionality from a limited and controlled scope, and might be useful for developing new experiments in this area.

6.1 Limitations and future work

In the next section we analyze limitations in our work in regards to the data used, the network architecture chosen and the effect of confounders in the probes.

6.1.1 Datasets for compositionality

One of the limitations of our work was accurate data portraying complex scenes with segmentation annotation and captions. In the COCO dataset annotated captions tend to focus on a few salient parts of the image, ignoring other regions and small objects. In Figure 16 we see a few examples where the image is rich in terms of number of objects, however their captions limit its description to salient parts of the image. As a result of the model being fine tuned on this data, we expect the model to pay less attention to objects not mentioned in the caption. This makes the object selection more relevant for our task since we prioritize selecting objects with higher attention for feature extraction. If the self-attention mechanism considered an object important, probably the same object will contain more information to be decoded.

Initially our work filtered images from COCO dataset using segmentation annotations, however they might include composition of two objects which are irrelevant to the scene, since they might be too small, partially ocluded, or just irrelevant compared to other objects. By applying a filter by caption, as described in section 4.2.3, we try to capture contextual information of the scene portrayed in the image and select images where the objects in the composition are also mentioned in the caption. This process aligns with the expectation of the model to be grounded and pay more attention to objects mentioned in the caption, therefore the data resulting from it should be more suitable for our analysis. [CBC22] explore transformer-based image captioning models and devise metrics to test if they are grounded while similarly, [ID21], analyzes attention links between words in the caption and objects. In both works results show transformers-based image captioning models are grounded to some extent.

The Natural Language Processing pipeline developed to identify whether the objects' names being tested are present in the caption might not find all words that can be used to denote the object. We used lemmatization and an analysis of the most frequent nouns in the text to manually identify synonyms. In our results, the caption filter improves decoding accuracy on average, and the composition (person, accessory) shows 5% improvement in decoding accuracy. The reason for higher impact in only one of the tasks is unknown, however the filtering process and the usage of different words might have affected datasets unevenly.

The use of a pre-trained word embeddings from common models used for NLP tasks, such as BERT[DCLT18], might improve dataset filtering. In addition to NLP, [SKC+15] proposes a graph-based semantic representation of images based on textual descriptions that can be used to filter the relation between objects. Visual genome[KZG+17] tries to address the relationship of objects in a scene by adding different region descriptors in the image, focusing on small and large details of each object. In their work, objects and groups of objects are annotated with rectangular bounding boxes, making it unsuitable for pixel-wise understanding of the scene as our work proposes. In future works we can explore further COCO dataset to devise better ways to improve data selection to test compositionality. To this date, datasets used for this task, such as Winoground, CREPE and visual genome lack segmentation annotations.

For the most part, the data available on COCO and the filtering process limited the scope of the analysis. The object pairs tested for compositionality were selected among a few others with a number of images higher than 1000. However, when filtering the data by caption, this number was significantly reduced making it difficult to compare two different decoding tasks with datasets heavily unbalanced. Future experiments using different datasets might provide cleaner data and shed more light on the methodology we used to test compositionality.

6.1.2 Network architecture

The neural network architecture we chose, as described in section 4.1, directly maps the relation of patches in the input image with internal layers of the encoder. In the future we can include the analysis of the decoder's attention to understand which words are aligned to the visual concepts represented by the ViT encoder. Furthermore, repeating our work on different captioning models would contribute to a better understanding of compositionality in different architectures.

We also notice from our results that the random token extracted from the object pair consistently reports similar decoding accuracy as the token which received maximum attention. This suggests the information might be equally distributed among more than one token within the object, therefore combining different feature vectors might provide better representation of the scene. In our results we also validate that it is not any token that can contain information about objects in the image. When we utilize the "max image" token (Table 1, Figure 8), a token with high attention but not from the object tested, the decoding accuracy decreases significantly and also the training process (Figure

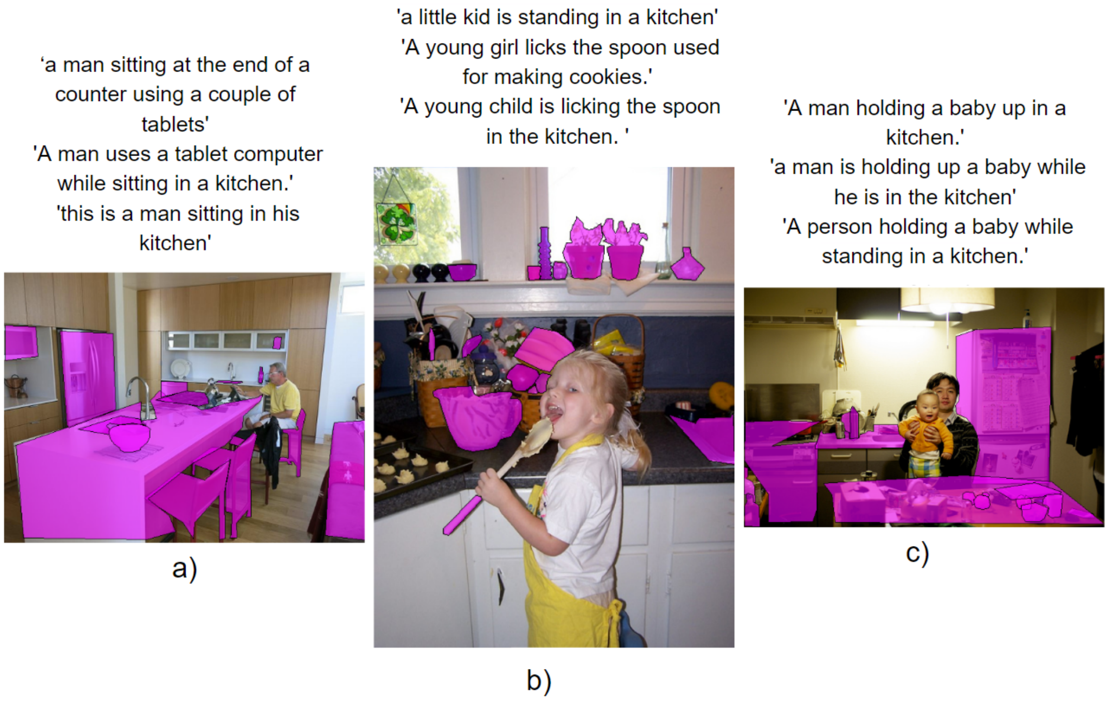

'a man sitting at the end of a counter using a couple of tablets'
'A man uses a tablet computer while sitting in a kitchen.'
'this is a man sitting in his kitchen'
a)
'a little kid is standing in a kitchen'
'A young girl licks the spoon used for making cookies.'
'A young child is licking the spoon in the kitchen. '
b)
'A man holding a baby up in a kitchen.'
'a man is holding up a baby while he is in the kitchen'
'A person holding a baby while standing in a kitchen.'
c)

Figure 16. Limitations of COCO dataset for detailing relationships between objects: Images from COCO dataset where all possible captions focus on describing only a part of the image, leaving all the surrounding objects in pink out of the text. To name a few objects from each image: In (a) there is a bowl on top of a table; in (b) there are plants next to each other and in (c) there is a person behind a table.

13) shows the decoding neural network cannot learn from this token alone.

6.1.3 Probing methods

The setup we created for probing represents a limited scope of compositionality in the dataset. We were able to label a few composite classes of two objects, based on the data we used. A more conclusive study may include more classes to decode and a large scale dataset. However, different from probing properties like part-of-speech in NLP[HL19] there is a large number of combinations that can be formed when parsing the visual scene for objects and relations.

7 Conclusion

This work reviews the literature on compositionality in neural networks and perform analysis on a particular type of architecture, Vision transformer. We try to answer whether a image captioning model, that learns from text and images, is capable of encoding relations of two objects internally. We devise a method to create a dataset for the study of compositionality and explore different forms of probes to validate our hypothesis. Our work expands the reseach on this area by desinging a series of experiments to validate whether the network is able to learn and represent complex concepts. The evidences we found are not conclusive to prove they are indeed learning such concepts, but provide intuition for further research on the matter.

References

[AB16] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[BDD$^+$17] Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. What do neural machine translation models learn about morphology? *arXiv preprint arXiv:1704.03471*, 2017.

[CBC22] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Explaining transformer-based image captioning models: An empirical analysis. *AI Communications*, 35(2):111–129, 2022.

[CSDS21] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[DBK$^+$20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[DCDN22] Bing Du, Xiaomu Cheng, Yiping Duan, and Huansheng Ning. fmri brain decoding and its applications in brain–computer interface: A survey. *Brain Sciences*, 12(2):228, 2022.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[GNQL21] Ruohao Guo, Dantong Niu, Liao Qu, and Zhenbo Li. Sotr: Segmenting objects with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7157–7166, 2021.

[HL19] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*, 2019.

[ID21] Nikolai Ilinykh and Simon Dobnik. What does a language-and-vision transformer see: The impact of semantic information on visual representations. *Frontiers in Artificial Intelligence*, page 182, 2021.

[KSH17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[Kum22] Ankur Kumar. The illustrated image captioning using transformers. *ankur3107.github.io*, 2022.

[KZG$^+$17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[LBPL19] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[LMB$^+$14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[LXT$^+$22] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022.

[LYMP22] Martha Lewis, Qinan Yu, Jack Merullo, and Ellie Pavlick. Does clip bind concepts? probing compositionality in large image models. *arXiv preprint arXiv:2212.10537*, 2022.

[LYY$^+$19] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[MBZ$^+$22] Jie Ma, Yalong Bai, Bineng Zhong, Wei Zhang, Ting Yao, and Tao Mei. Visualizing and understanding patch interactions in vision transformer. *arXiv preprint arXiv:2203.05922*, 2022.

[MHG$^+$22] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? *arXiv preprint arXiv:2212.07796*, 2022.

[NRK$^+$21] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

[PGFC20] Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks. *arXiv preprint arXiv:2012.12352*, 2020.

[RDS$^+$15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[RHGS15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[RKH$^+$21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[RWC$^+$19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[SHG$^+$22] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[SKC$^+$15] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015.

[SVB$^+$21] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[TJB$^+$22] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

[TSF$^+$16] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[VSP$^+$17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[VTBE15] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

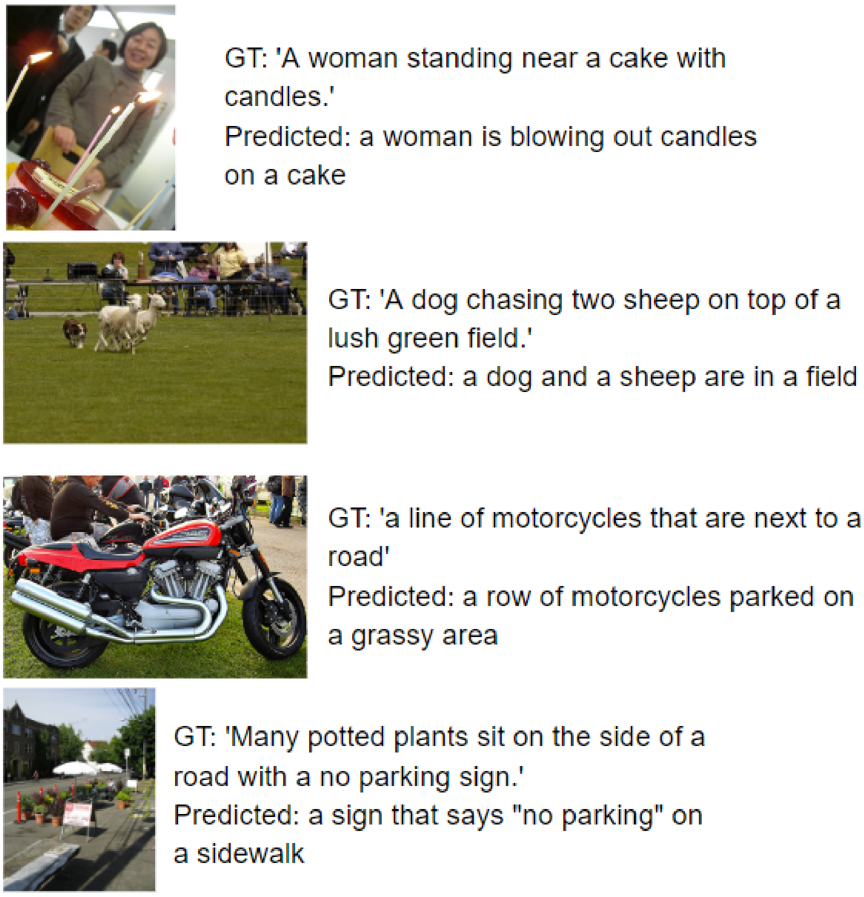
[WYM$^+$22] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.

[XBK$^+$15] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[YBK$^+$22] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv e-prints*, pages arXiv–2210, 2022.

[ZLZ$^+$21] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

Appendix



GT: 'A woman standing near a cake with candles.'
Predicted: a woman is blowing out candles on a cake

GT: 'A dog chasing two sheep on top of a lush green field.'
Predicted: a dog and a sheep are in a field

GT: 'a line of motorcycles that are next to a road'
Predicted: a row of motorcycles parked on a grassy area

GT: 'Many potted plants sit on the side of a road with a no parking sign.'
Predicted: a sign that says "no parking" on a sidewalk

Figure 17. Sample images with Ground truth (GT) from COCO and predictions by the ViT-based image captioning model.

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Braian Olmiro Dias**,
(author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Content based analysis of compositionality in Vision Transformers,
   (title of thesis)

   supervised by Tarun Khajuria, MSc.
   (supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Braian Olmiro Dias
09/05/2023

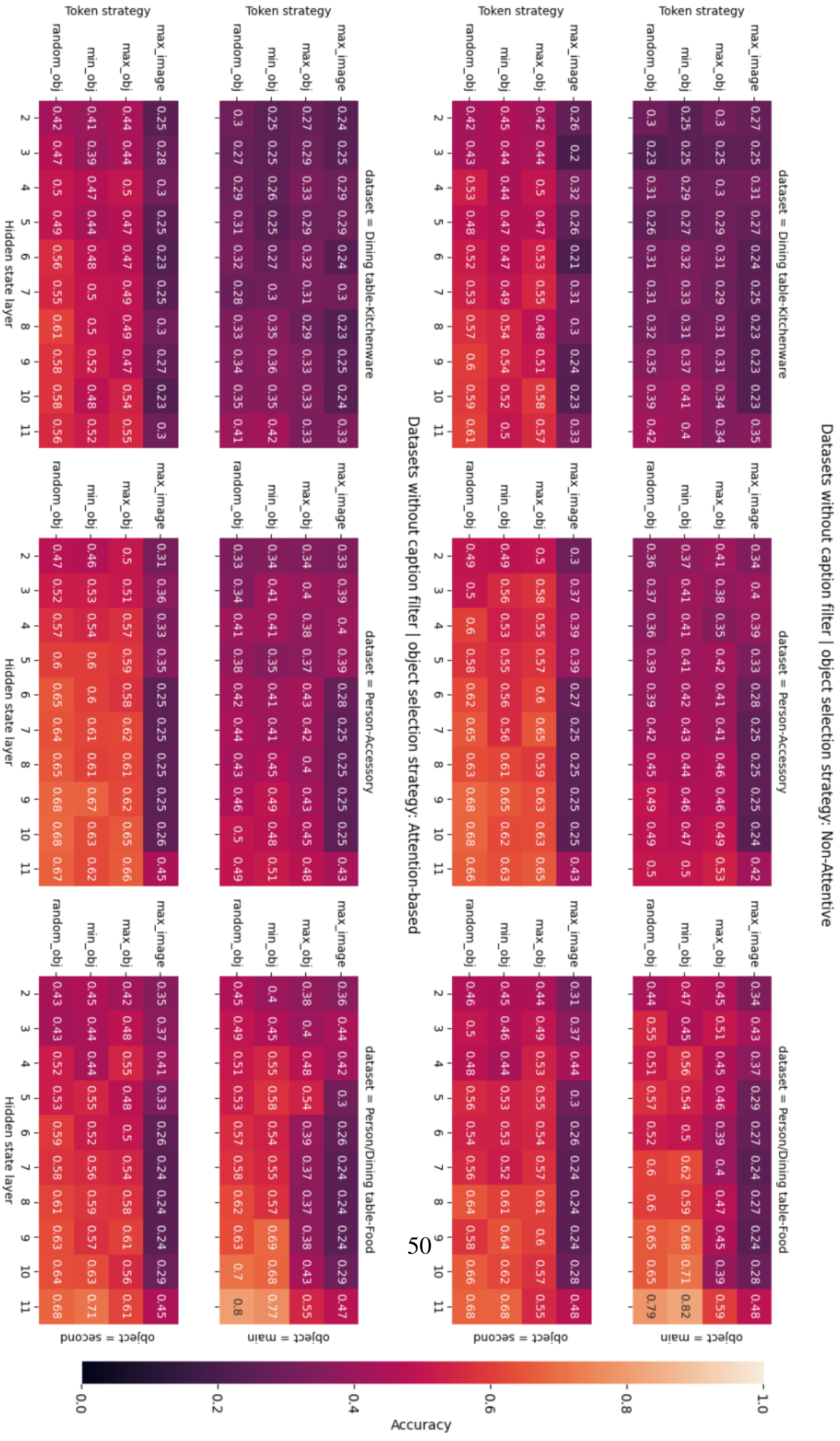


Figure 18. All experiment results

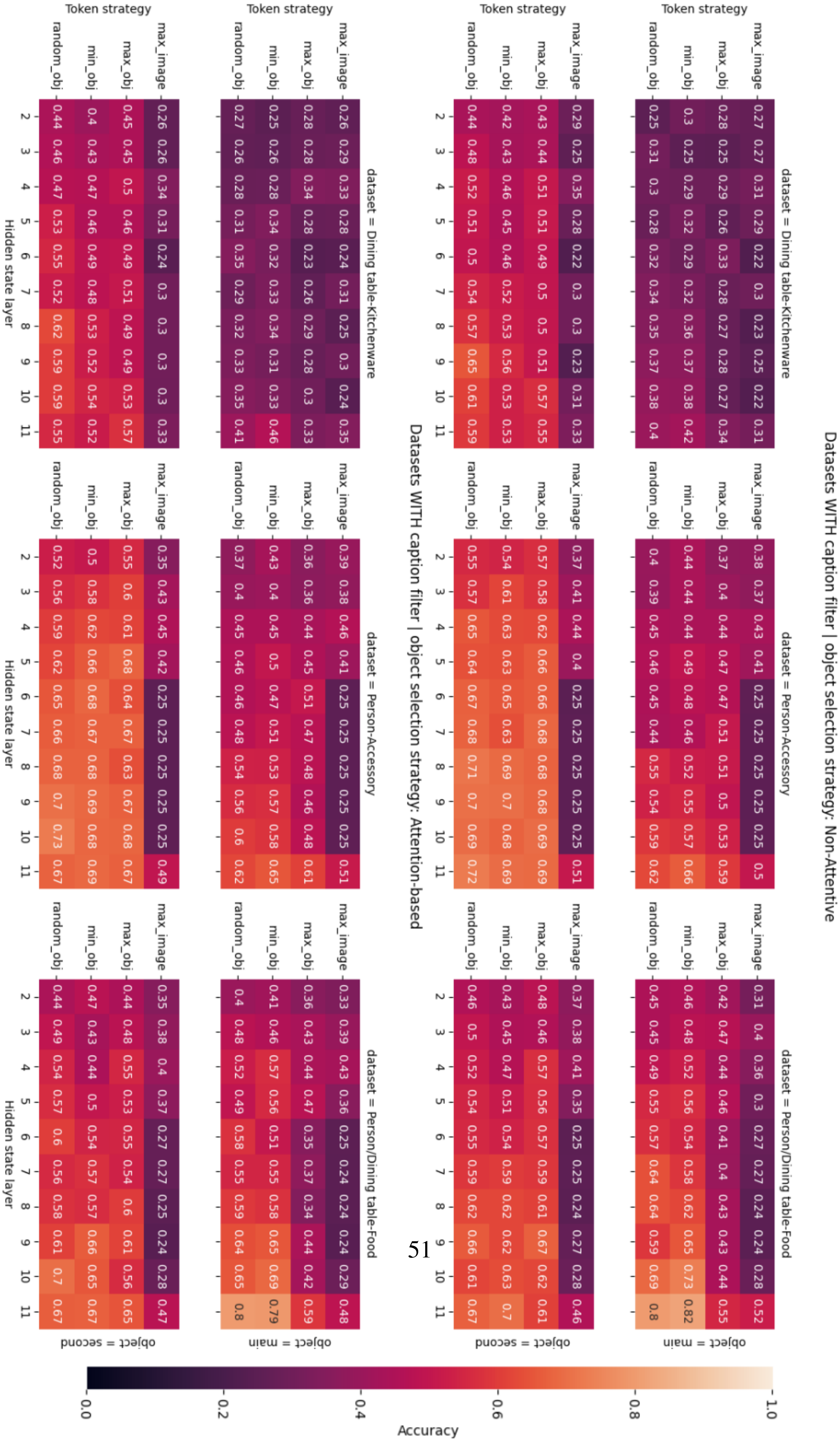


Figure 19. All experiment results