Investigating the Arithmetic of Visual Embeddings: Comparisons between Model families and Extraction Techniques

Shwetali Shimangaud
Student
McGill University
shwetali.shimangaud@mail.mcgill.ca

Rishabh Thaney
Student
McGill University
rishabh.thaney@mail.mcgill.ca

Abstract

This research delves into the influence of various visual pre-training paradigms, including self-supervised and language supervision, on the properties of image embeddings. Drawing inspiration from the well-established NLP example of semantic arithmetic (e.g., "King - Male + Woman = Queen"), we aim to assess how different pre-training methodologies affect the geometric properties of image embeddings. Given the inherent diversity across model architectures, it is anticipated that the resulting embeddings will exhibit distinct properties. Such an investigation is of paramount importance in multi-modal spaces where both textual and visual data are integrated, as it contributes to a deeper understanding of geometric relationships and facilitates vector-oriented reasoning. Through a comprehensive examination, this study endeavors to shed light on the intricate interplay between pre-training paradigms and image embedding properties, thus paving the way for enhanced multi-modal learning techniques.

[1]

1 Introduction

In the realm of computer vision and natural language processing(NLP), the generation of effective visual embeddings plays a crucial role in various tasks such as image classification, object detection, and semantic similarity assessment. This study aims to delve into the impact of different training approaches on visual embeddings, specifically focusing on classification-based pre-trained vision encoder models (ResNet) (He et al., 2015), and language supervision models (CLIP) (Radford et al., 2021).

While these models share the commonality of being neural network-based, they exhibit distinct architectures and training methodologies. Classification-based pretrained models utilize large supervised datasets, producing embeddings optimized for differentiating between different classes within the dataset. Language supervision models, on the other hand, combine vision and NLP components and are trained using paired data, associating images with textual descriptions or captions to learn embeddings capturing semantic similarity between images and text.

The diverse nature of these training approaches implies that the embeddings they produce may exhibit different arithmetic properties. For example, one would intuitively expect NLP-like embedding properties to hold more when there is strong supervision, i.e. in CLIP. Additionally, the techniques used for extracting embeddings also influence their arithmetic characteristics, especially when dealing with multiple images corresponding to a single text (e.g., multiple images of "King"). In this study, we will explore average pooling extraction technique.

The primary contributions of this research can be summarized as follows:

1. Investigating the arithmetic properties of embeddings generated by different families of vision models.

2. Creation of an image dataset with analogical pairs, akin to the SemEval dataset for Task 2, to facilitate further research in this domain.

2 Related work

This study explores the properties of embeddings generated by two distinct models: Classification models (ResNet), and language supervision models(CLIP). The efficacy and versatility of pre-trained classification-based models, such as ResNet, have been widely recognized within the vision domain. These models have demonstrated remarkable efficiency in tasks such as image classification, object detection, and feature extraction. By

[1]Github: http://surl.li/suzzb

leveraging large-scale datasets and sophisticated architectures, ResNet and its variants have set benchmarks in image understanding tasks, prompting extensive exploration into their underlying mechanisms and representations.

The emergence of language supervision-based CLIP (Radford et al., 2021) embeddings has proven advantageous across various tasks, including classification and detection. Extensive research in the natural language domain has laid the groundwork for understanding embedding spaces. For instance, prior investigations, as demonstrated in (Mikolov et al., 2013b), have showcased the arithmetic properties of word embeddings generated by traditional models like Word2Vec (Mikolov et al., 2013a). This line of inquiry has revealed both syntactic (e.g., cars:car::cats:cat) and semantic (e.g., King:Queen::Man:Woman) analogies, facilitating vector-oriented reasoning for downstream applications.

Recent endeavors in the multi-modal domain (Couairon et al., 2022), which aims to bridge the gap between textual and visual modalities, have attracted significant interest. Notably, researchers have examined the geometric properties of multi-modal embedding spaces using methods like delta vector transformation. Their findings suggest that while embeddings built on pre-trained sentence embeddings exhibit arithmetic properties, they may not necessarily demonstrate superior linear characteristics. Moreover, investigations from paper (Liang et al., 2022) have uncovered a modality gap in representations of text and images, attributed to model initialization and contrastive learning optimization.

In light of these developments, our study investigates whether pure vision embeddings exhibit arithmetic properties and explores the influence of different pre-training and supervision methods on embedding arithmetic.

3 Dataset : SemEval-2012 Task 2

The SemEval-2012 Task 2 dataset (Jurgens et al., 2012) (see Figure 2) focuses on measuring the relational similarity between word pairs, assessing how well computational models can identify and quantify the similarity of relationships rather than the words themselves. It's important for testing algorithms that understand the types of relationships words can have, enhancing tasks like analogy solving and semantic search. This dataset was instrumental in our research as it provided a standardized benchmark for measuring relational similarity between word pairs, facilitating the evaluation of algorithms designed to understand and quantify semantic relationships, thereby enhancing tasks such as analogy solving and semantic search

Our dataset was assembled by selecting 50 word pairs from the SemEval-2012 Gold dataset. Additionally, we devised a Python script to systematically retrieve 30 stock images for each word pair. Then the images underwent manual verification to ensure alignment with the contextual semantic meaning of each word under study. For instance, in the case of the word relation: "fruit:apple::furniture:chair" images of apple were sought. However, it was observed that the automated script retrieved some images of Apple iPhone and other products from Apple Inc. Figure 1 shows the final compilation of images for the word "apple" in our dataset.

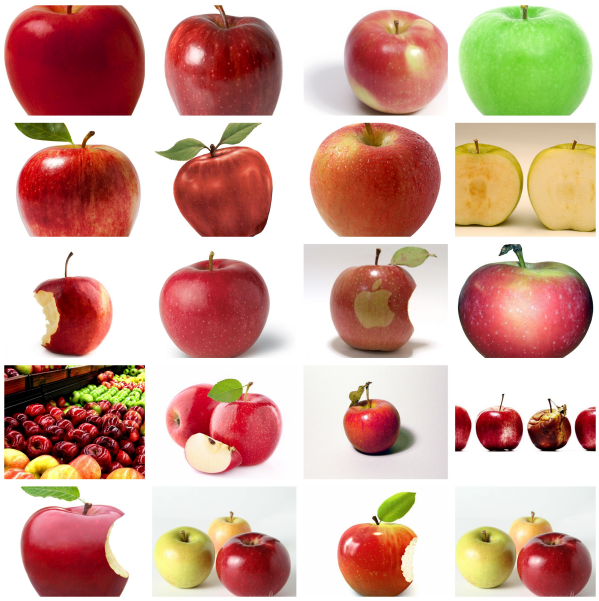

Figure 1: Example images for the word 'apple' in our dataset

4 Evaluation metrics

4.1 Cosine similarity

Cosine similarity can be applied as an important metric to measure the degree of similarity between embeddings. By computing the cosine angle between two embedding vectors, this metric offers a way to quantify how closely related two images are.

84.2 "animal:pig"
74.1 "fruit:grape"
74.0 "sweater:knit"

Figure 2: Example of the SemEval-2012 Task 2 dataset

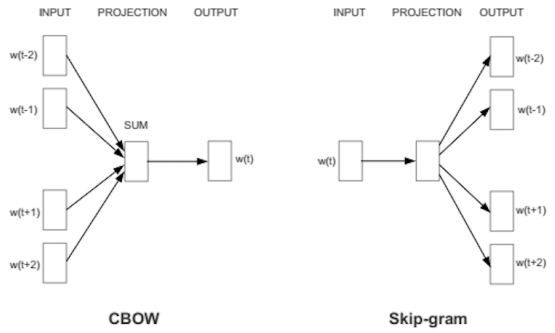

Figure 3: Word2Vec architecture

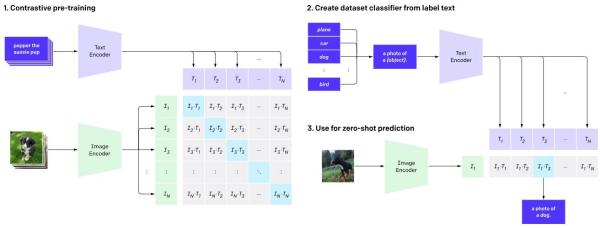

Figure 4: CLIP Training Architecture

4.2 Spearman's rank correlation coefficient

Spearman's rank correlation coefficient, (Xiao et al., 2015), serves as a statistical tool for assessing the magnitude and orientation of the relationship between two variables ranked in order. By calculating the Spearman's coefficient between the Offset vector and the fourth word embedding, we can ascertain the degree of alignment between these two vectors. For instance, in the analogy "fruit:apple::furniture:chair," the Spearman's coefficient will gauge the strength and direction between 'fruit - apple + chair' and 'furniture'.

5 Models

We hypothesize that the Contrastive Language-Image Pre-training (CLIP) model, trained on a fusion of language and image data, will demonstrate the 'Embedding Arithmetic' property, allowing for operations on embedded representations of concepts. Conversely, we anticipate that the ResNet model, trained exclusively on visual data, may not exhibit this property to the same degree.

5.1 Baseline model : Word2Vec

The Word2Vec model generates word embeddings by training a neural network to predict a word based on its context within a corpus, capturing semantic similarities through distributed representations. It can be implemented using two architectures: the Continuous Bag of Words (CBOW) Model and the Continuous Skip-Gram Model as shown in Fig. 3. In the paper (Mikolov et al., 2013b), it is shown that the word representations learned by RNNLM do an especially good job in capturing linguistic regularities with the vector offset method. Also, for the semantic generalization on the SemEval 2012 task, this method outperformed the previous state-of-the-art models.

5.2 Proposed Models

5.2.1 CLIP

The CLIP model generates embeddings by learning to associate images and corresponding textual descriptions, such as captions, through a contrastive learning framework. The contrastive learning is a method where the AI model is taught to recognize similarities and differences of a large number of data points. For instance, imagine you have a main item (the "anchor sample"), a similar item ("positive sample"), and a different item ("negative sample"). The goal is to make the model understand that the anchor and the positive item are alike, so it brings them closer together in its mind, while recognizing that the negative item is different and pushing it away. CLIP model encodes both images and texts into a shared latent space where similarity between them is maximized, enabling the model to understand complex visual and textual concepts and their relationships. The training process for CLIP is shown in Fig. 4. As CLIP encodes text and images together, we hypothesize that it will follow "Embedding Arithemtic" property.

5.2.2 ResNet

ResNet (Residual Network) employs a deep neural network architecture featuring residual connections, or skip connections, which enable the network to efficiently train very deep models. These

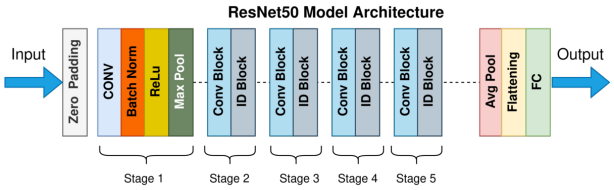

Figure 5: ResNet50 Architecture

skip connections allow gradients to flow directly through the network, mitigating the vanishing gradient problem and enabling the training of much deeper architectures. ResNet generates embeddings by processing input images through a series of convolutional layers with residual blocks. The final layer before the classification stage typically consists of a global average pooling layer followed by a fully connected layer, which transforms the feature maps into a fixed-size vector representation, or embedding, capturing the semantic content of the input image. For this research, ResNet50 model is used. The architecture is shown in Fig. 5.

6 Experiments

6.1 The Vector offset method

As we have seen, the tasks have been formulated as analogy questions. As mentioned in the paper (Mikolov et al., 2013b), a simple vector offset method based on cosine distance is remarkably effective in solving these questions. In this method, to answer the analogy question a:b c:d where d is unknown, we find the embedding vectors x_a, x_b, x_c, and compute y = $x_b - x_a + x_c$. Note that, here the embeddings are calculated by averaging the embeddings for all 30 images. y is the continuous space representation of the image we expect to be the best answer. Of course, there might not be any image in the dataset which has exact same embedding, so we then search for the word whose image embedding vector has the greatest cosine similarity to y and output it:

$$w^* = \arg\max_w \left(\frac{x_w \cdot y}{\|x_w\| \cdot \|y\|} \right) \quad (1)$$

When d is given, as in our dataset, we simply use cos($x_b - x_a + x_c$, x_d) for the words provided. We note that this measure is qualitatively similar to relational similarity model of (Turney, 2012), which predicts similarity between members of the word pairs (x_b, x_d),(x_c, x_d) and dis-similarity for (x_a, x_d).

6.1.1 Correct Analogy questions

We conducted experiments using three models: Word2Vec, CLIP, and ResNet. For CLIP, experiments were performed separately for text and image embeddings. In initial phase, our primary focus was to assess the accuracy of these models on a smaller dataset. We repeated the experiment by normalizing the embeddings to investigate if normalization affects the Embedding Arithemtic property.

6.1.2 Randomly chosen Pairs

In the next phase, we randomly selected pairs from our dataset to create new analogy questions. For instance, given the analogy "hour:seconds::feet:inches," we substituted random pairs (e.g., "raise:elevate") for "feet:inches" to form new analogies (e.g., "hour:seconds::raise:elevate"). We generated a total of 50 such pairs and conducted the vector offset experiment on each.The objective of this experiment was to observe model behavior when presented with analogies containing dissimilar relationships. We expected to encounter random answers or words more closely associated with "hour," "seconds," or "feet," as the relationship between "hour" and "seconds" differs from that between "raise" and "elevate."

6.1.3 Extended Dataset

In the final phase of our experiment, we augmented our dataset by incorporating all semantic analogy pairs from the SemEval-2012 Task 2 Gold dataset, comprising 3,217 analogy questions. For instance, the pairs "weapon:spear," "bird:robin," and "insect:ant" yielded two analogy questions: "weapon:spear::bird:robin" and "weapon:spear::insect:ant." While the relationship between "weapon" and "spear" mirrors that between "bird" and "robin," the relationship between "insect" and "ant" differs from that of "weapon" and "spear." Consequently, the degree of similarity between "weapon:spear" and "insect:ant" is lower, resulting in a decreased probability of obtaining the correct answer. Note that, for this dataset the vector offset method was employed only with CLIP Text embeddings and Word2Vec embeddings. This nuanced variation in similarity levels enriches the dataset and allows for a more comprehensive assessment of model performance across different analogy types.

6.2 Cosine similarity

In our cosine similarity experiment, we aimed to explore the degree of similarity between the embeddings obtained from the three models that were used: CLIP, Word2Vec, and ResNet. For each analogy question within our dataset, we calculated the offset vector embedding (second word - first word + third word) and compared it with the fourth word embedding to assess how closely related they are. i.e. for analogy "fruit:apple::furniture:chair", cosine angle will be calculated between 'fruit - apple + chair' and 'furniture'. This method allowed us to measure how closely related these embeddings are in meaning across different models, giving us a sense of how well each model captures subtle semantic nuances in our context. Note that, We performed this experiment seperately for CLIP Image encoder and Text encoder.

6.3 Spearman's coefficient

In our Spearman rank coefficient analysis, we delved into the relationships between embeddings derived from the CLIP, Word2Vec, and ResNet models. Spearman's correlation coefficient, calculated before and after normalization, served as a valuable metric for assessing the strength and direction of monotonic relationships between these embeddings, expressed as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where ρ represents the Spearman's rank correlation coefficient, d_i represents the differences in ranks between paired data points and n is the number of data points. We examined the alignment between the offset vector embedding and the embedding of the fourth word to further understand the semantic relationships captured by each model. This analysis offered a nuanced understanding of how well each model captured semantic nuances and provided valuable insights into the robustness of the embeddings in representing semantic information within the context of our study.

7 Results and Discussion

7.1 The Vector offset method

7.1.1 Correct Analogy questions

The Word2Vec model exhibited higher accuracy, correctly answering 6 out of 50 analogies in the Top-1 and 32 out of 50 analogies within the top-5 predictions. In contrast, the CLIP image encoder achieved a top-1 accuracy of 2 out of 50 analogies and a top-5 accuracy of 25 out of 50 analogies. Table 1 presents the performance of all the models in answering the vector offset analogies created during our experiment. Note that, when we conducted the same experiment by normalizing the embeddings, results were same.

	Word2Vec	CLIP image encoder	CLIP Text encoder	ResNet-50
Top-1	6	2	2	2
Top-5	32	25	25	16

Table 1: Vector Offset method experiment for Analogy questions

	Word2Vec	CLIP image encoder	ResNet-50
Top-1	0	0	0
Top-5	23	23	18

Table 2: Vector Offset method experiment for Randomly chosen pairs

7.1.2 Randomly chosen Pairs

When pairs were randomly selected, none of the models provided the correct answer within the top-1 predictions. This experiment was conducted across Word2Vec, the CLIP image encoder, and ResNet-50 models. Notably, while Word2Vec and the CLIP model exhibited comparable performance, ResNet-50 performed notably worse than the CLIP model. The results are shown in table 2

Our experiment revealed interesting insights into the performance of models when presented with randomly chosen pairs versus correct pairs. Notably, for the CLIP model, the answer appeared in the top-5 predictions 23 out of 50 times when pairs were randomly chosen, compared to 25 out of 50 times for correct pairs. This marginal difference suggests a consistent ability of the model to capture analogical relationships even when pairs are randomly selected. The results are consistent with those of the ResNet model.

We hypothesize that this phenomenon may be attributed to the "island hypothesis." This hypothesis suggests that in embedding spaces, the words of each pair exist in isolated islands. For instance, in the analogy "weapon:spear::bird:robin," the embeddings of "weapon" and "spear" will exist in isolated island spaces, and no other image embeddings will

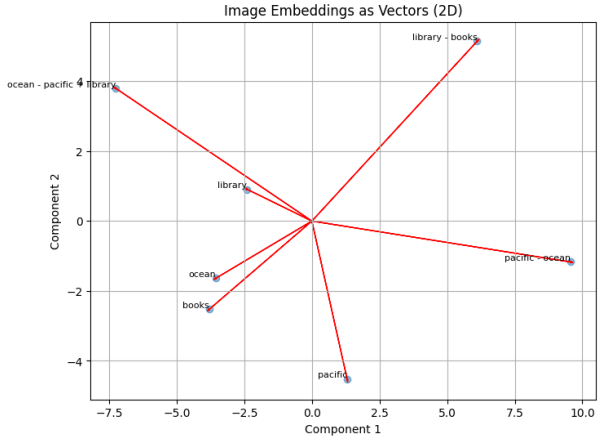

Figure 6: Image Embeddings

	Word2Vec	CLIP Text encoder
Top-1	16	21
Top-5	244	194

Table 3: Vector Offset method experiment for Extended dataset

be closer to them. Consequently, the difference (weapon - spear) between these two word vectors will be very small, and when this small vector offset is added with another word (weapon - spear + robin), the resulting word embedding remains close to the original word, i.e., "robin." This issue is less prominent in larger vocabulary spaces, where word embeddings are more distributed and interconnected minimizing the likelihood of isolated islands. The visualization of these embeddings is shown in Fig. 6 for one pair in dataset.

7.1.3 Extended Dataset

In case of extended dataset, the Word2Vec model outperformed CLIP text encoder model. The answer was present for 244 pairs in Top-5 for Word2Vec model and 194 times for CLIP Text encoder. The detailed results are shown in Table 3

This outcome suggests that with larger datasets, the CLIP text encoder demonstrates performance similar to that of the Word2Vec model. However, verifying this for image embeddings poses challenges due to the difficulty in collecting correct images for all 3217 pairs. Given that CLIP is trained on both text and images simultaneously, it is conceivable that CLIP image embeddings may exhibit similar behavior. Nevertheless, this remains an open question that necessitates a larger dataset for verification.

7.2 Cosine Similarity and Spearman's coefficient

We observed distinct performance characteristics of the CLIP Image Encoder, CLIP Text Encoder, ResNet-50, and Word2Vec models.

7.2.1 Cosine Similarity experiment (Table 5)

In examining the cosine angles across various models, we observe that the CLIP Image Encoder's results span from a minimum of 15.7° to a maximum of 62.9°, averaging at 38.4°, suggesting that the model maintains a consistent embedding space with moderate variation in visual content. The CLIP Text Encoder presents a narrower range with angles from 24.2° to 49° and an average of 34.9°, indicative of a strong and consistent capture of textual semantic similarity. ResNet-50, with its wider angle range from 19.3° to 88° and an average of 50.4°, shows a considerable variation that might be attributed to its diverse visual processing capabilities. Word2Vec, primarily a linguistic model, exhibits the broadest range of cosine angles with an average of 72.4°, suggesting that it excels in capturing linguistic relationships.
Additionally, the experiment revealed that CLIP consistently showed strong cosine similarity between offset vector embedding and fourth word embedding resulting in smaller cosine angles. This performance is particularly highlighted in the context of table 4, which clearly illustrates CLIP's leading scores in the cosine similarity comparisons.

7.2.2 Spearman's Coefficient (Table 6)

Regarding Spearman's coefficient, the CLIP Image Encoder shows a range from 0.13 to 0.89 with an average of 0.46, reflecting a robust ability to correlate ranked variables in a visual context. The CLIP Text Encoder has a more modest correlation range from 0.1 to 0.75 with an average of 0.34, which still demonstrates a reasonable degree of association in the text domain. ResNet-50 exhibits the highest maximum coefficient of 0.93 and an average of 0.61, suggesting that despite its wide range, it can effectively capture strong monotonic relationships in visual data. Word2Vec's Spearman coefficients, ranging from 0.02 to 0.74 with an average of 0.29, reaffirm its linguistic roots.

Pair 1	Pair 2	Offset vector	4th word	CLIP	ResNet	Word2Vec
hour:seconds	feet:inches	seconds - hour + feet	inches	44.09°	56.29°	58.04°
ocean:pacific	planet:earth	pacific - ocean + planet	earth	37.99°	48.78°	70.36°
laugh:happiness	nod:agreement	happiness - laugh + nod	agreement	56.88°	56.85°	85.79°

Table 4: Cosine angles amongst different models

	Min value	Max Value	Avg value
CLIP Image encoder	15.7°	62.9°	38.4°
CLIP Text encoder	24.2°	49°	34.9°
ResNet-50	19.3°	88°	50.4°
Word2Vec	37.88°	86.3°	72.4°

Table 5: Cosine angle between offset vector and fourth word embedding

	Min value	Max Value	Avg value
CLIP Image encoder	0.13	0.89	0.46
CLIP Text encoder	0.1	0.75	0.34
ResNet-50	0.03	0.93	0.61
Word2Vec	0.02	0.74	0.29

Table 6: Spearman's coeffiecient between offset vector and fourth word embedding

8 Conclusion

Based on our findings, we assert that CLIP image encoder showed stronger geometric properties for image embeddings than ResNet and CLIP text encoder was comparable to Word2Vec, aligning with our initial hypothesis. Notably, despite the distinct architectures leading to differing embeddings, all models demonstrated adherence to arithmetic embedding principles to varying degrees. Given the limited scale of our dataset (30 images per pair across 50 pairs), our results, while satisfactory, indicate the potential for further enhancement with a larger and more comprehensive dataset, suggesting the promise of achieving superior performance under more expansive conditions.

9 Contributions

The authors contributed equally to this work.

References

Guillaume Couairon, Matthijs Douze, Matthieu Cord, and Holger Schwenk. 2022. Embedding arithmetic of multimodal queries for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4950–4958.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. SemEval-2012 task 2: Measuring degrees of relational similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364, Montréal, Canada. Association for Computational Linguistics.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

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

Chengwei Xiao, Jiaqi Ye, Rui Esteves, and Chunming Rong. 2015. Using spearman's correlation coefficients for exploratory data analysis on big dataset: Using spearman's correlation coefficients for exploratory data analysis. *Concurrency and Computation: Practice and Experience*, 28:n/a–n/a.