A ViT-based Siamese Network for Visual Reasoning in the KiVA Challenge

October 7, 2025

Joan Salvà Soler* [1]

[1]H2O.ai

Abstract

We present a novel Vision Transformer (ViT)-based Siamese network for visual analogical reasoning developed for the Kid-inspired Visual Analogies (KiVA) Challenge. This architecture achieves 95.9% accuracy on the benchmark, demonstrating strong performance across all difficulty levels and establishing the effectiveness of the architecture for visual reasoning tasks. Code is available at https://github.com/jsalvasoler/kiva-iccv.

1 Introduction

Visual analogical reasoning, the ability to infer and apply abstract rules from visual examples, is a hallmark of human intelligence and a critical component of flexible, general-purpose problem-solving [1]. The KiVA benchmark provides a framework for evaluating this capability in AI systems, grounding the task in developmental psychology by using simple transformations of everyday objects that are solvable even by young children [2]. The challenge frames this task in the classic A:B :: C:? format, where a model must identify the transformation that turns A into B and apply it to C to find the correct outcome among several choices. The following is an example of a task:

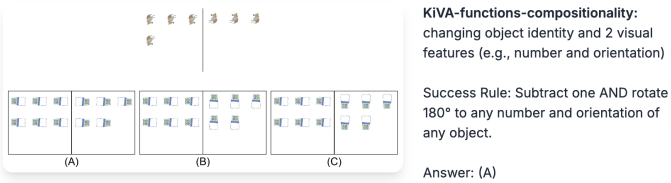

Figure 1: Example of a KiVA task.

The KiVA benchmark comprises three difficulty levels: **KiVA (easy)**, where only the object identity changes between the example and test transformations; **KiVA-functions (moderate)**, in which both the object identity and one visual feature (such as orientation, size, or number) change; and **KiVA-functions-compositionality (difficult)**, where the object identity and two visual features (for example, orientation, size, or number) change between the example and test. This progression allows for a systematic evaluation of analogical reasoning as the complexity of the required transformation increases.

2 Method

Our approach employs a Siamese Network [3] designed specifically for visual analogical reasoning. Given an example transformation pair (A, B) and a test scenario C with multiple candidate choices $(D_1, \ldots, D_n)$, the model must identify which choice correctly completes the analogy. The model operates as follows:

1. The example transformation (A, B) is encoded using a transformation encoder to produce a transformation vector $\mathbf{t}_{AB}$.
2. Each candidate transformation $(C, D_1), \ldots, (C, D_n)$ is similarly encoded to produce choice vectors $\mathbf{t}_{CD_i}$.
3. The goal is to identify the choice i where $\mathbf{t}_{CD_i}$ is most similar to $\mathbf{t}_{AB}$, typically measured using cosine similarity.

An important detail of our approach is that we assume that the eight images of the KiVA task can be extracted from the "stitched" image. It is a fair assumption, as the complexity of the task is the same when the images are given separately. Moreover, it would be simple to train a model to extract the images from the stitched image, if necessary. In our KiVA task, with $n = 3$, we will refer to these images as A, B, C, D_1, D_2, D_3, where D_1, D_2, D_3 are the three choices.

Transformation Encoder A traditional approach to this task would be to encode each image independently and compute the transformation as $t = f(B) - f(A)$, following the analogy-making strategy used in models like Word2Vec [4]. However, we found that this method was insufficient for capturing the nuanced visual relationships between images in the analogy task. Instead, we model the transformation directly as $t = f(A, B)$, where f is a transformer-based image encoder that leverages cross-image attention to jointly process both images. Through experiments with various vision transformer architectures (ViT and DINOv3), we determined that a ViT-based approach was the most effective for this purpose.

Architecture of the ViT-based Transformation Encoder We adapt the ViT architecture, which is pretrained on 224×224 images, to process two images as a unified sequence. Both input images (A and B) are independently passed through the pretrained ViT patch embedding layer. For 224×224 images with 16×16 patches, this produces two sequences of 196 patch embeddings each, with dimension d (e.g., $d = 384$ for ViT-Small). These patch embeddings are concatenated and a learnable [CLS] token is prepended to form a unified sequence that will aggregate the transformation representation. Positional embeddings are then added to provide spatial location information, where the embedding matrix is extended by duplicating the original patch positional embeddings for both images. Learned segment embeddings distinguish patches from the "before" image (Segment 1) versus the "after" image (Segment 2), with Segment 0 for the [CLS] token. The sequence is then passed through the pretrained

ViT transformer blocks (12 layers for ViT-Small). Critically, the self-attention mechanism allows patches from image A to attend to patches from image B, enabling direct comparison of corresponding spatial regions. Finally, after layer normalization, the [CLS] token is extracted and passed through a learnable projection head (not part of pretrained ViT) consisting of linear layers, ReLU, dropout, and layer normalization to produce the final transformation embedding $\mathbf{t}_{AB}^{\text{final}} \in \mathbb{R}^e$. The complete forward pass can be summarized as:

$$\text{patches}_A,\ \text{patches}_B \in \mathbb{R}^{196\times d}$$
$$x = [[\text{CLS}]\,;\ \text{patches}_A\,;\ \text{patches}_B] \in \mathbb{R}^{393\times d}$$
$$x \leftarrow x + [\text{pos}_{\text{CLS}}\,;\ \text{pos}_{\text{patches}}\,;\ \text{pos}_{\text{patches}}]$$
$$x \leftarrow x + \text{segment\_embed}(\text{seg\_ids})$$
$$x \leftarrow \text{TransformerBlocks}(x)$$
$$\mathbf{t}_{AB}^{\text{final}} = \text{ProjectionHead}(\text{LayerNorm}(x)[0]) \in \mathbb{R}^e$$

Loss Function After comparing the standard triplet loss and softmax cross-entropy loss, we found that a simple **contrastive analogy loss** [5] performed best. Given a training example consisting of an example transformation $\mathbf{t}_{\text{ex}}$, one correct choice transformation $\mathbf{t}_{\text{pos}}$, and n incorrect choice transformations $\{\mathbf{t}_{\text{neg}_i}\}_{i=1}^n$, the loss aims to maximize the similarity between $\mathbf{t}_{\text{ex}}$ and $\mathbf{t}_{\text{pos}}$, minimize the similarity between $\mathbf{t}_{\text{ex}}$ and each $\mathbf{t}_{\text{neg}_i}$, and enforce a margin m between positive and negative similarities:

$$\mathcal{L} = \frac{1}{n-1}\sum_{i=1}^{n-1} \max\left(0, m - \left(s_{\text{pos}} - s_{\text{neg}_i}\right)\right) \tag{1}$$

where $s_{\text{pos}} = \text{sim}(\mathbf{t}_{\text{ex}}, \mathbf{t}_{\text{pos}})$ and $s_{\text{neg}_i} = \text{sim}(\mathbf{t}_{\text{ex}}, \mathbf{t}_{\text{neg}_i})$ are the cosine similarities. In the KiVA challenge with $n = 3$ choices per example, we have 2 negative examples.

3 Experiments & Results

Implementation and Hardware The implementation uses PyTorch for modeling and Neptune for logging. We perform the experiments on a single NVIDIA L40S GPU with 40GB of memory.

Model Configuration We use `vit_small_patch16_224` as our encoder backbone, initializing its weights with pretrained checkpoints from `timm`. This sets a resolution of 224×224 pixels. We set the embedding dimension to $e = 512$, and the projection head consists of two linear layers with a ReLU activation, dropout, and layer normalization, mapping the ViT output to the final embedding space. This gives the model a total of 22.7M learnable parameters.

Training Configuration The optimizer AdamW with weight decay 1×10^{-4} is used to train the model. A cosine annealing schedule is used to decay the learning rate over 20 total epochs, starting from 3×10^{-5} for the encoder parameters and 3×10^{-4} for the projection parameters. The batch size is 64 with a gradient accumulation of 4 for an effective batch size of 256. The contrastive margin is $m = 0.05$. We use mixed precision training to speed up the training process.

Training Data We use the code from the official KiVA dataset [2] to augment the provided training data. Our pipeline generated random training samples on-the-fly that follow a similar distribution as the official training set. We included transformations with parameter combinations not present in the original training set to improve model generalization. We set the epoch length to 65536 examples, 2752 of which come from the official training set, and the rest are generated on-the-fly.

Results Table 1 presents our model's Top-1 accuracy on both validation and test sets, broken down by difficulty level. The results demonstrate the effectiveness of our approach, and the model's generalization is excellent, as the gap between validation and test accuracy is minimal. Figure 2 provides a granular breakdown of test set performance, showing that while the model performance decreases as task complexity increases, particularly on compositionality tasks involving counting. The final configuration was identified after extensive experimentation (152 runs), as illustrated by the sample of training curves in Figure 3.

Table 1: Final accuracy (%) on the KiVA benchmark. Results are shown for validation and test sets across three difficulty levels.

Split	KiVA (Easy)	KiVA-func (Moderate)	KiVA-comp (Difficult)	Overall
Validation	100%	100%	94%	95.4%
Test	100%	100%	95%	95.9%

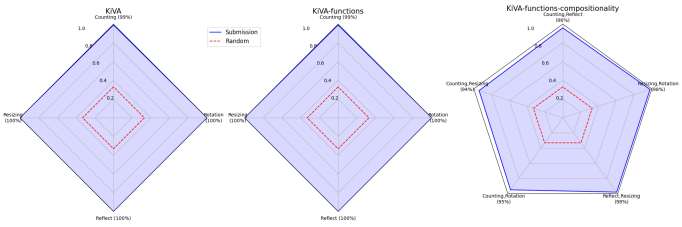

Figure 2: Radar plot of accuracy on the test set by transformation type.

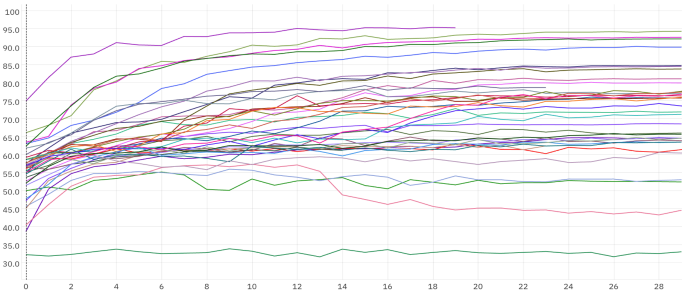

Figure 3: Validation accuracy during KiVA training runs.

Conclusion We presented a ViT-based Siamese network that processes image transformation pairs as unified sequences, enabling the direct modeling of visual changes via cross-attention. Our approach achieves 95.9% accuracy on the KiVA benchmark, demonstrating that joint encoding of transformation pairs is a highly effective strategy for visual analogical reasoning. This architecture provides a strong foundation for future work on visual reasoning tasks.

References

[1] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170, 1983.

[2] Eslie Yee, Gabriel Grand, Judy E Fan, Dan Gutfreund, and Joshua B Tenenbaum. KiVA: Kid-inspired Visual Analogies. *arXiv preprint arXiv:2407.17773*, 2024. Published as a conference paper at ICLR 2025.

[3] Davide Chicco. Siamese neural networks: An overview. *Methods in Molecular Biology*, 2190:73–94, 2021. doi: 10.1007/978-1-0716-0826-5_3.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[5] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.