Disentangling and Integrating Relational and Sensory Information in Transformer Architectures

Awni Altabaa, John Lafferty

Statistics & Data Science, Yale University

Scan for project webpage

Motivation & High-level Goals

Relational reasoning is fundamental to intelligence

- Cornerstone of human intelligence, underpins capabilities for *analogy*, *abstraction*, *generalization*
- Standard Transformers are data-inefficient at relational tasks and exhibits brittle OOD generalization
- Prior relational architectures are narrow in domain due to restrictive inductive biases.

Our Goal: Augment Transformers with explicit relational mechanisms while retaining sensory processing capabilities

Key Insight: Two types of information require two types of attention:

- Sensory : features of individual objects
- Relational : relationships between objects

Background: Sensory & Relational Information in Standard Transformers

Strength of Transformers: *attention* ; Versatile *information retrieval* mechanism

The Transformer architecture, essentially:

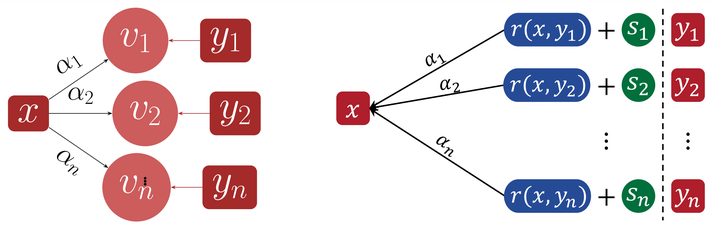
1. *Information Retrieval*: Attention
$$x_i' \leftarrow \sum_j \alpha_{ij}\, \phi_v(x_j)$$
2. *Local Processing*: Token-wise feedforward network
$$x_i' \leftarrow \mathrm{MLP}(x_i)$$
3. Repeat

Here, the attention scores α_{ij} serve as a *selection criterion* for routing information between objects, whereas the information being routed is sensory embeddings of object-level features $\phi_v(x_j)$.

Two key types of information: *sensory* and *relational* .

The standard attention mechanism of Transformers represents a learned information retrieval operation over *sensory information*, but does not explicitly route *relational information*.

Dual Attention Transformer: Disentangling & Integrating Attention over Sensory & Relational Information



(a) $\mathrm{Attention}(x,(y_1,\ldots,y_n))$ (b) $\mathrm{RelationalAttention}(x,(y_1,\ldots,y_n))$

Left: Sensory attention retrieves object features. Right: Relational attention retrieves relation vectors tagged with symbols

Relational Attention : An attention mechanism for explicitly routing *relational* information:

$$\boldsymbol{a}_i \leftarrow \sum_j \alpha_{ij} \cdot (W_r\, \boldsymbol{r}_{ij} + W_s\, s_j)$$

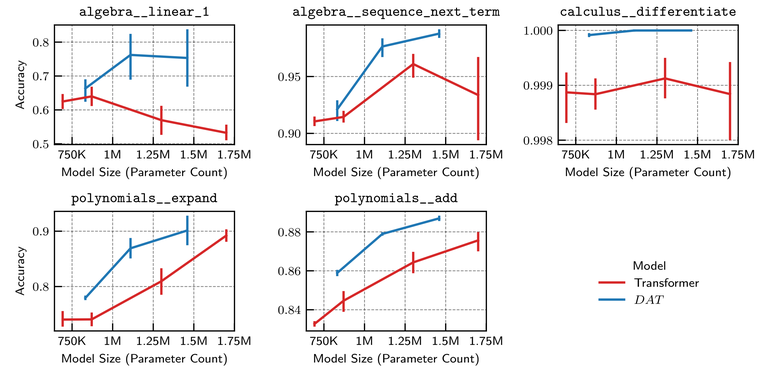
1. *Attend:* Compute *attention scores* to select object(s) to attend to
$$\alpha_{ij} = \mathrm{Softmax}([\langle \phi_q^{\mathrm{attn}}(x_i), \phi_k^{\mathrm{attn}}(x_j)\rangle]_{j=1}^n)_j$$
2. *Relate:* Compute *relation vectors* representing relationship with attended object(s)
$$\boldsymbol{r}_{ij} = \left(\langle \phi_{q,\ell}^{\mathrm{rel}}(x_i), \phi_{k,\ell}^{\mathrm{rel}}(x_j)\rangle\right)_{\ell\in[d_r]} \in \mathbb{R}^{d_r}$$
3. Tag with *symbols*: Abstract identifiers
$$(s_1,\ldots,s_n) = \mathrm{SymbolRetriever}(x_1,\ldots,x_n)$$

Dual Attention: A *multi-head* attention mechanism consisting of a *combination* of ***standard attention*** heads—for attending to *sensory* information—and ***relational attention*** heads—for attending to *relational* information.

Dual Attention Transformer (DAT): A Transformer architecture where each *multi-head attention* module is replaced with *dual attention*, enabling explicit joint routing of both sensory and relational information.

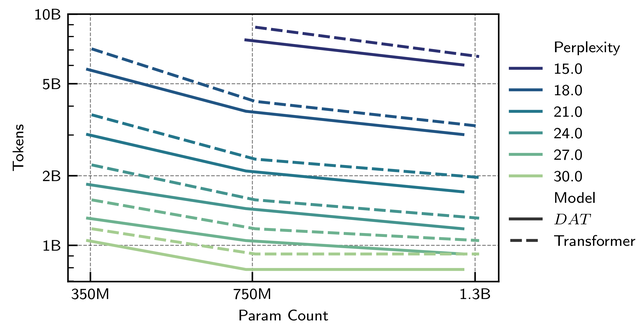
Highlights of Experimental Results

Mathematical Problem Solving (Seq2Seq)



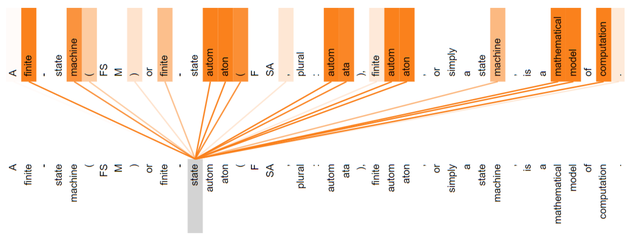
DAT shows superior scaling on mathematical reasoning tasks

Language Modeling Scaling Laws



DAT is more data-efficient and parameter-efficient than standard Transformers

Visualization of Learned Latent Relational Representations



A visualization of the learned relation activations in a *DAT* language model. This depicts one dimension of the relation vector $\boldsymbol{r}_{ij}$ in the twelfth layer, focusing on the token `` `state' `` as the source i, which has high activation with the tokens `` `mathematical' ``, `` `model' ``, and `` `computation' ``.