Scene Graph Parsing by Attention Graph

Martin Andrews
martin@RedDragon.ai

Yew Ken Chia
ken@RedDragon.ai

Sam Witteveen
sam@RedDragon.ai

Summary

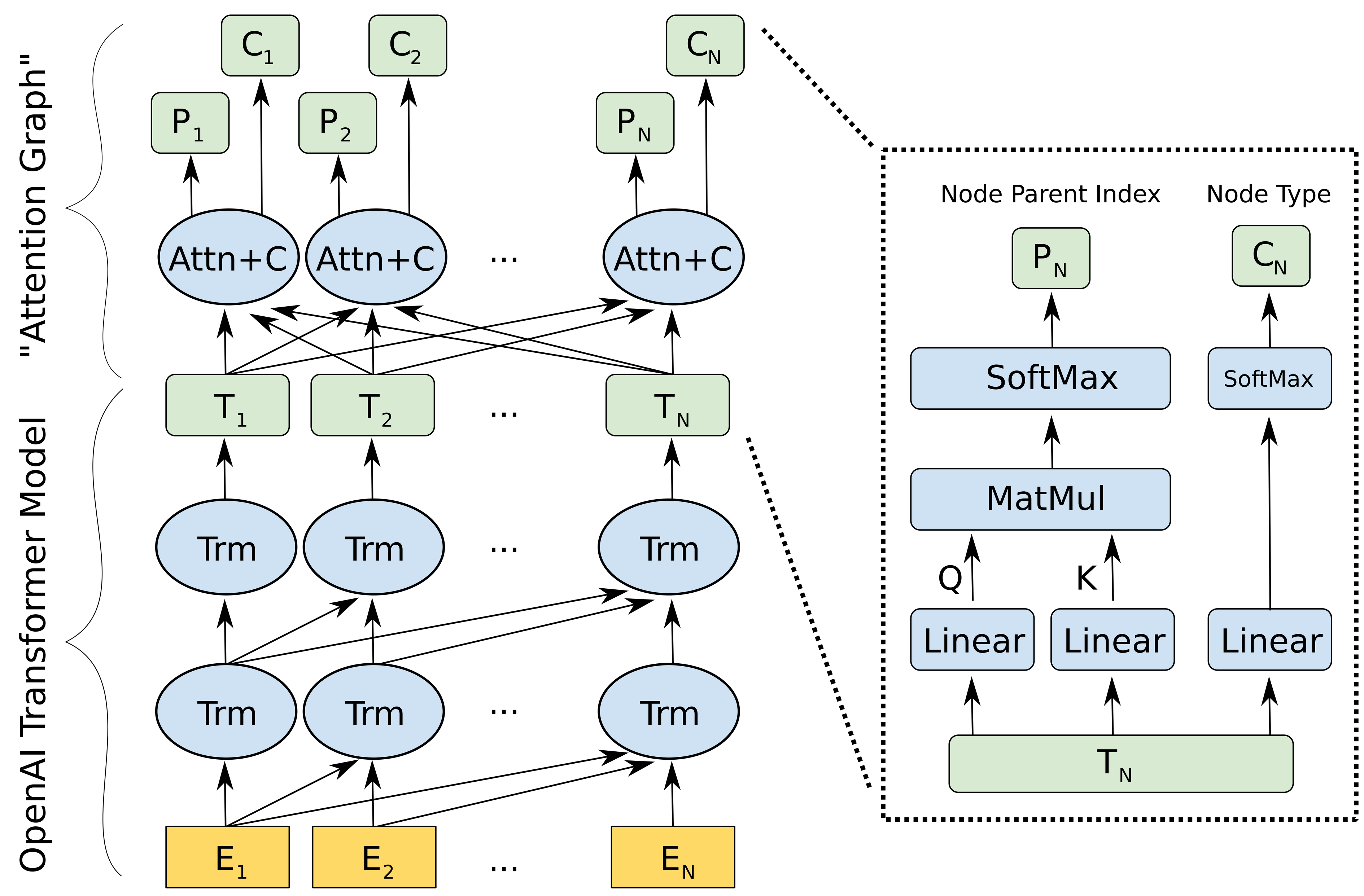
Task :
- Convert text to graph representation

Builds upon :
- Visual Genome dataset : text & graphs
- Transformer architecture

Ideas :
- OpenAI Transformer model as base
- Train additional attention layer
- "Parent" links *defined by* attention
- Graph linkages directly from top layer

Results :
- Higher scores than transition-based parsers
- Early results encouraging...

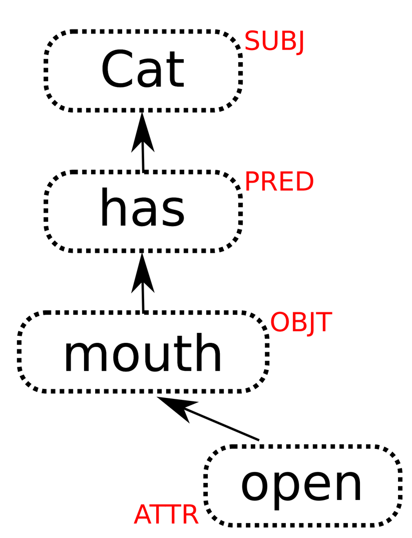
Attention Graph : Model Architecture



Dataset - Visual Genome (with MS COCO splits)

Text :
- Cat has his mouth open

To Graph :



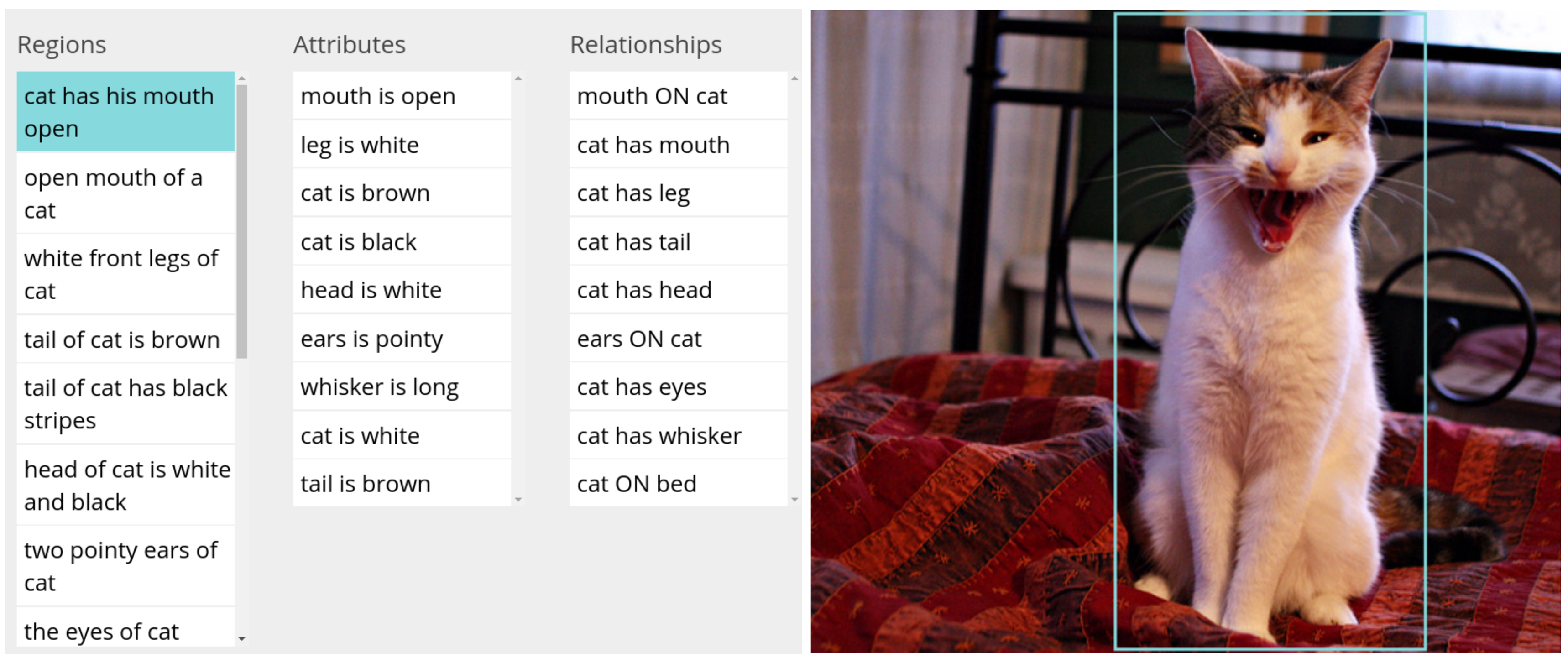


Figure 1: Example from data exploration site for [20]. For this region, possible graph objects would be {cat, mouth}, attributes {brown←cat, black←cat, white←cat open←mouth}, and relationships {cat←has←mouth, mouth←ON←cat}.

Results

Table 1: SPICE metric scores for the Oracle (using code released by [13]) and our method, under the base assumptions, and also where the number of tuples is bounded above by the number of potentially useful words in the region description

Parser	F-score reported in [13]	F-score (our tests)	F-score (limited tuples)
Attn. Graph (ours)		0.5221	0.5750
Oracle	0.6985	0.6630	0.7256

Table 2: SPICE metric scores between scene graphs parsed from region descriptions and ground truth region graphs on the intersection of Visual Genome [20] and MS COCO [22] validation set.

Parser	F-score
Stanford [23]	0.3549
SPICE [14]	0.4469
Custom Dependency Parsing [13]	0.4967
Attention Graph (ours)	0.5221
Oracle (as reported in [13])	0.6985
Oracle (as used herein)	0.6630

Discussion

Motivation :
- Benefits of creating KB from text
- Transition-based parsers seem limited
- Elegance of lifting attention to graph

Dataset issues :
- Ground truth results < 100%
- Heuristics can improve a little
- Looking for alternatives

Model Architecture :
- Training builds on pretrained LM
- Simplest attention mechanism used
- No hyperparameter search done

Future directions :
- Apply to bigger graph chunks
- Adapt to more general graphs
- Encoder/decoder Transformers for sequence-to-graph

Source code available:
- http://RedDragon.ai/research

Key References

"Scene graph parsing as dependency parsing" - Wang et al. (2018)
"Visual genome: Connecting language and vision using crowdsourced dense image annotations" - Krishna et al. (2016)
"Improving language understanding with unsupervised learning" - Radford et at. (2018)

Contact

martin@RedDragon.ai
+65 8585 1750
http://RedDragon.ai