Beyond myopic inference in Big Data pipelines

Karthik Raman, Adith Swaminathan, Thorsten Joachims, Johannes Gehrke
Cornell University, Ithaca, NY USA

Introduction

- **Setting**: Big Data pipelines constructed using modular components
- **Problem**: Error by a component cascades through the pipeline causing catastrophic failure in the eventual output
- **Key idea**: Establish correspondence between pipelines and Probabilistic Graphical Models that explains pipeline operation theoretically
- **Result**: More robust inference procedures while still using existing components

An illustrative example: A NLP pipeline

- Error detection needs a notion of confidence scores for predictions.
- Error recovery needs a mechanism for alternative predictions

Efficient inference: Beam and Adaptive inference

- **Figure 5.** Top-k inference causes multiplicative blowup of inference cost
- **Observation**: Diminishing returns from more values
- **Idea**: Use beam search to limit list lengths
- **Result**: Given budget m * k, retain top m after each stage

- For robust inference, ideal #outputs required from each component will vary for different inputs
- Unlike Top-k and Beam, Adaptive inference exploits this
- Effect of an output on overall prediction is estimated first
- Propagate if it has a large effect

Create scored list \([x_1^*, ..., x_k^*]\). If \(\text{Score}(x^*) > \tau \cdot \text{Score}(x^{i+1})\), return \([x_1^*, ..., x^*]\).

Discussion

- **Top-K, Beam and Adaptive Inference** are generic algorithms
- No assumptions about components' error models, or the pipeline structure.

Conclusion and Future Work

- Canonical inference with myopic components cause unrecoverable pipeline errors
- Viewing pipelines as graphical models allows reasoning about overall inference
- Proposed different inference procedures to approximate ideal inference problem
- Experiments demonstrate robust pipelines constructed using existing components
- Handling pipelines with feedback
- Incorporating uncertainty of predictions into training

Contact

The full paper is available for personal use at http://www.cs.cornell.edu/~adith/Papers/PipelineInference.pdf
For more information, please e-mail: adith@cs.cornell.edu