# Multi-Modal Information Extraction (Question-Answer Framework)

Vamsi Chitters, Fabian Frank, Lars Jebe
{vamsikc, fabfrank, larsjebe}@stanford.edu

CS229
Machine Learning
Fall 2017

## MOTIVATION

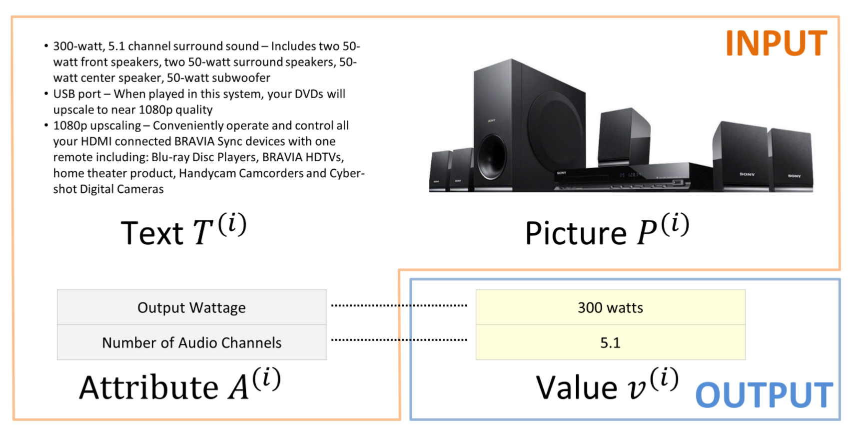
- Humans rely on various modes (e.g., visual, auditory, textual) to understand context in order to make informed decisions
- Multi-modal information extraction similarly aspires to extract domain knowledge using different modes
- Potential applications:
  - Answer questions such as "what is the best bar in the city?" (using reviews and images)
  - Structure info (e.g., *Knowledge Graph*), NER, Q&A

## PROBLEM DEFINITION

- Leverage multi-modal source information to answer attribute queries efficiently using supervised learning

| | |
|---|---|
| # of attribute-value pairs | 8,306,549 |
| # of unique attributes | 2,185 |
| # of unique values | 14,787 |
| # of products | 2,517,709 |
| # of colored images | 4,891,284 |
| dataset size | 206 GB |

**MuMIE dataset (provided by *Diffbot*)**




**Example task**

- Evaluation metric:

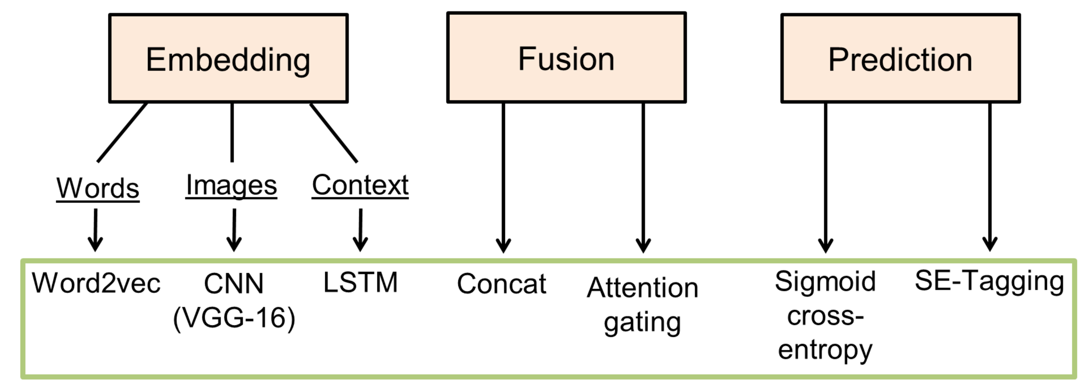
$$Acc@k = \frac{1}{N}\sum_{i=1}^{N} I(v^{(i)} \in \hat{v}_k^{(i)})$$

## CHALLENGES

- Predicting values of varying length (from 0 to 14 words)
- Extracting semantic meaning from images
- Dataset characteristics:
  - Data imbalance impacts effectiveness of LSTM context: $len(description)\ \epsilon\ [20, 2783]$
  - Previously unseen vocabulary words in test set
- Web-scraped dataset:
  - Non-intuitive and nonsensical attribute-value pairs (e.g., (1) color: "no" (2) m: "xs, s, l")
  - Duplicate values (e.g., bluetooth: "4", "4.0", "v4")
- Balancing model expressiveness with performance

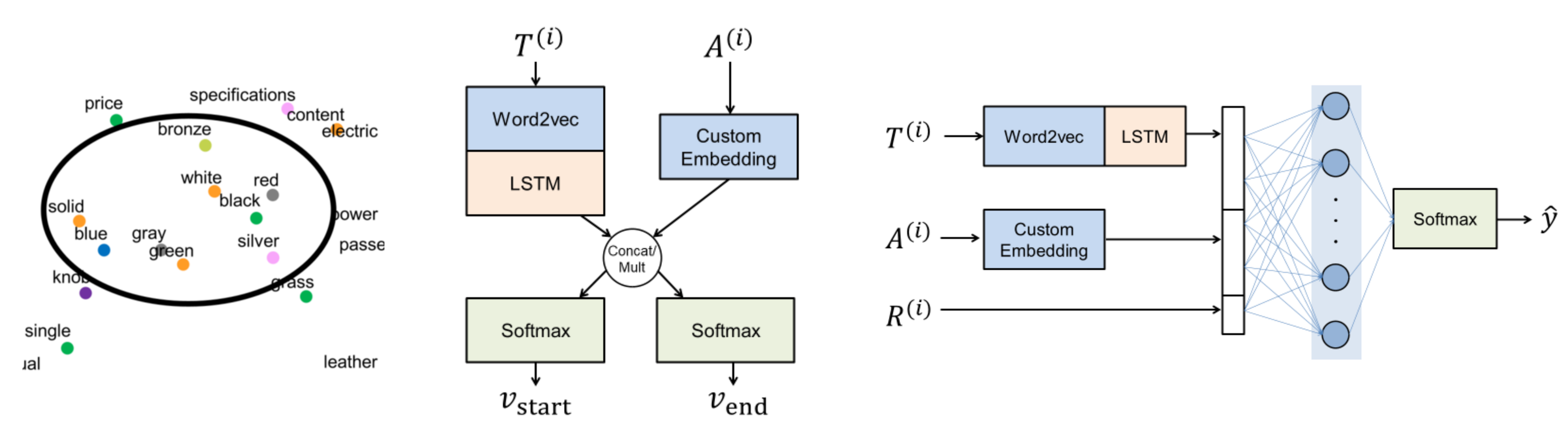
## METHODS

### Applied Techniques




### Textual Mode

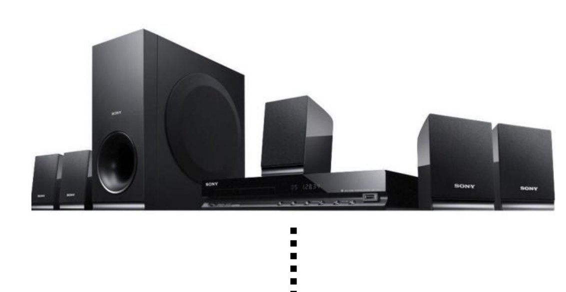
1) Baseline: self-trained Word2vec with single-word value predictions using cosine similarity
2) LSTM-Pointer Network: LSTM layer for description, custom embedding layer for attributes, model outputs value start and end positions within description
   - LSTM represents text as context vector
   - Attribute embedding captures query intention
   - Predicts start and end of value independently
3) Recommendation-based Pointer Network: Construct recommendations (*hand-crafted features*) based on prior distr. of values, capture dependence of value start and end positions
   - Predicts span instead of independent start and end positions
   - Additional hidden layer to learn weighting recommendations




1) Word cluster with colors
2) LSTM-Pointer Network
3) Recommendation-based Pointer Network

### Visual mode (WIP)

- Transfer learning: pre-trained on ImageNet VGG-16 architecture
- Two approaches:
  - Caption images and combine with text descriptions
  - Extract image features and combine with word features



[u'loudspeaker', u'home_theater', u'desktop_computer', u'radio']

**Image caption from our model**

## RESULTS AND ANALYSIS

| Experiment | Specifications[1] (Train dataset size: 2150) (Test dataset size: 544) | Accuracy @ k = 1 Total | | Accuracy @ k = 1 Multi-Word | |
|---|---|---|---|---|---|
| | | Train | Test | Train | Test |
| Baseline | Word2vec cosine similarity | 1.5% | 0.5% | 0.0% | 0.0% |
| LSTM-Pointer Network | Independent pointers | 86.5% | 16.3% | 14.1% | 12.1% |
| Recommendation-Ptr-Net | Additional hidden layer | 90.7% | 23.0% | 90.1% | 22.0% |
| Rec-Ptr-Net + Images | With image captioning | 88.6% | 22.4% | WIP | WIP |
| **Recommendation-Ptr-Net** | **Trained on 342,099 examples** | **77.7%** | **72.7%** | **65.3%** | **58.0%** |

[1] Experimented with various hyperparameters, including # of LSTM memory units, max sequence length, extent of data sanitization, vocabulary composition, learning rate and choice of optimization algorithm

| Error Type[2] | Attribute | Prediction | Actual Value |
|---|---|---|---|
| Single Word (close) | batteries included | included | yes |
| Multi-Word (close) | measurements | height 1.17 in | heel height 1.17 in |
| Complete Miss | finish | price | semi-gloss |

[2] Other errors included predicting e.g., $v_{end} < v_{start}$ (SE-Tagging). Improvements for all error types include: (1) pre-train value-categories for attributes (2) impose mutual dependence between $v_{end}$ and $v_{start}$ (3) add domain-aware constraints such as [value for "dimensions" must include numbers]

## DISCUSSION

- Data quality really matters (e.g., web-scraped data contains unintuitive ground truth values)
- Difficult to replicate human-like multi-modal information extraction
- Expected: using two modes would improve task accuracy
  Reality: images don't help much (hard to extract semantic meaning)
- LSTMs proved effective for QA due to context (well-researched)
- Image captions predicts classes (nouns), not fine-grained predictions (adjectives)
- Domain knowledge (manually selected features) improves accuracy
- Pointer network addresses multi-word value prediction effectively - captures dependency between start and end index well

## FUTURE WORK

**Text**

- Gain intuition from more complex models (e.g. R-NET)
- Explore other multi-word value approaches (e.g. BIO-Tagging)

**Image**

- Train ImageNet on own images (re-train last few hidden layers)
- Learn relevance of image feature vectors relative to text

- Enhance attention model to focus on relevant text and image parts (i.e., look at car body if query is "car brand")

REFER-ENCES

- Rajpurkar, Pranav, et al. "Squad: 100,000+ questions for machine comprehension of text." *arXiv preprint arXiv:1606.05250* (2016).
- Ma, Lin, Zhengdong Lu, and Hang Li. "Learning to Answer Questions from Image Using Convolutional Neural Network." *AAAI*. Vol. 3. No. 7. 2016.
- Antol, Stanislaw, et al. "Vqa: Visual question answering." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.
- Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal Machine Learning: A Survey and Taxonomy." *arXiv preprint arXiv:1705.09406* (2017).
- Wang, Shuohang, and Jing Jiang. "Machine comprehension using match-lstm and answer pointer." *arXiv preprint arXiv:1608.07905* (2016).
- Vinyals, Oriol, Meire Fortunato, and Navdeep Jaitly. "Pointer networks." *Advances in Neural Information Processing Systems*. 2015.