Spezifikationsvortrag
NL Generation from structured inputs

Software Project

Prof. Dr. Anette Frank

Kim Almasan, Tatjana Chernenko, Siting Liang, Bente Nittka

12th June, 2018

Institute for Computational Linguistics of University Heidelberg
Im Neuenheimer Feld 325, 69120 Heidelberg, Germany

Table of contents

Specification

Task Definition

- Automatic image description generation from attributes
- Explore the impact of the features extracted from the textual descriptions and attributes
- Focus on spatial relationships between the objects and sufficient attributes for the objects
- Consider only the images with 2-5 objects in one image (see subsection "Data set" for more details).
- Based on Dong et al., 2017 (generate product reviews from attributes)

System Development

Main Idea

- Use the overlap of the V-COCO, MS COCO and COCO-a data sets (see Data set for more information) to get the attributes and image descriptions
- Incrementally add more attributes to the input

Development steps

1. Only information about objects in the image (MS COCO)
2. Attributes describing actions in the image (V-COCO)
3. Additional annotations for interactions (emotions and spatial relations) (COCO-a)
4. Including the image itself (image vector)

Approach

Encoder-decoder architecture

- **Encoder:** the feed-forward neural network encoder.

  Input: the normalized vector representations of all the attributes.
  Output: a rich fixed-length vector representation.

- **Decoder:** several LSTMs.

  Input: output of the encoder + the words from the sequences from the description sentences.
  Output: the target sentence (By inference - for example BeamSearch).

Approach

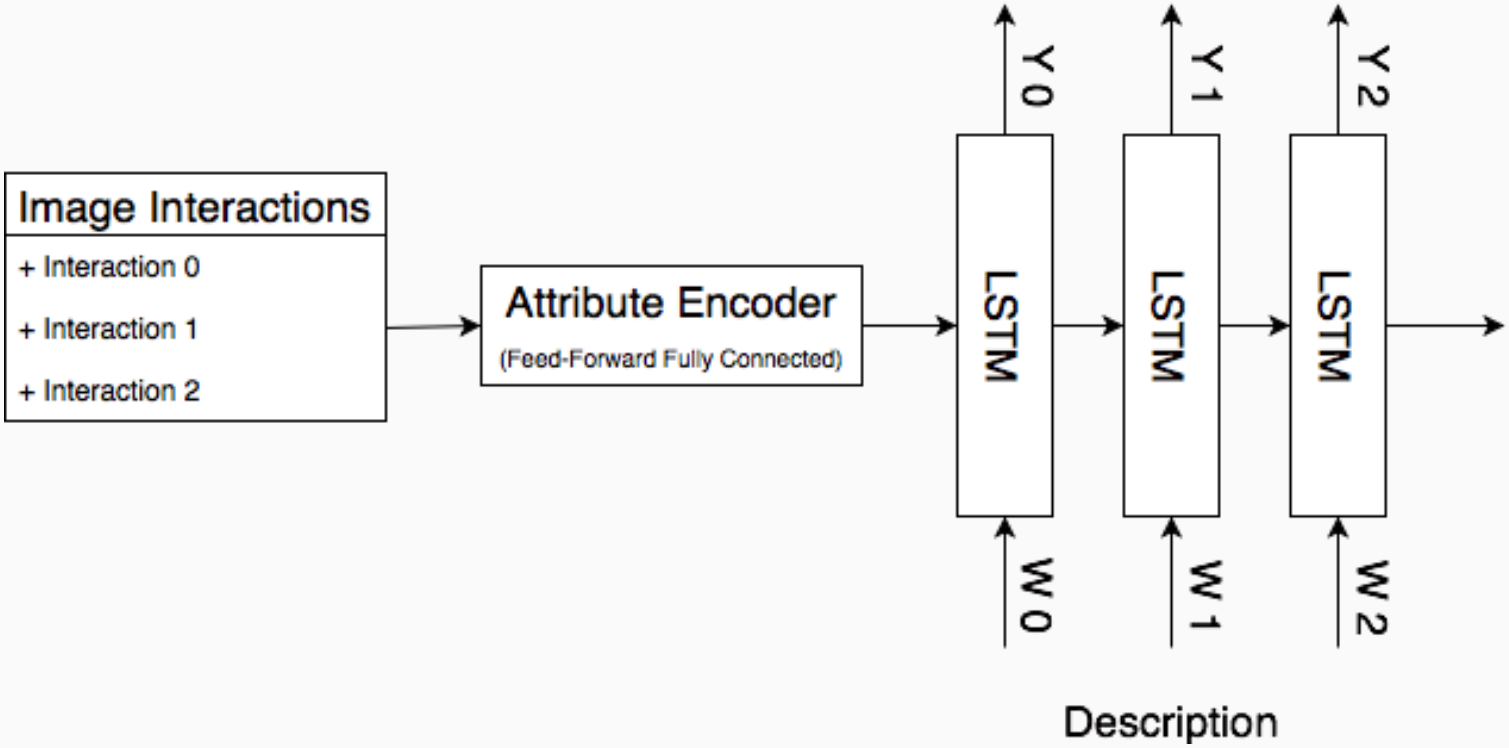

Figure 1: Model Diagram

Special Problems

1. **The attributes** from the COCO-a data set have the structure of a directed graph
   - *Encoder input:* try a multidimensional structure of the input that saves the information about the hierarchical structure of the interaction groups
   - *Architecture of the encoder:* the feed-forward neural network
   - *Convert encoder output* to a suitable for the decoder form: concatenation of the vectors
   - *Other possible solutions:* the one-dimensional representation of the input, BiRNN as the encoder and averaging the output vectors
2. Focus on spatial relations, but MS COCO **descriptions** are not always reach enough (e.g., descriptions only for one subject). Explore the effect of attributes.

Evaluation

Baseline: the first development step

Evaluation metrics

1. **Word-based metrics** (better correlations to human ratings of informativeness):

- BLEU (measure the word overlap)
- METEOR (measure the word overlap)
- CIDEr (measure the word overlap)
- Semantic Similarity (SIM - distributional similarity and Latent Semantic Analysis, complemented with semantic relations extracted from WordNet)

Evaluation

Evaluation metrics

2. Grammar-based metrics (better correlations to quality and naturalness):

- Readability Flesch Reading Ease score (RE)

3. Human evaluation:

- 6-point Likert scale for informativeness, naturalness and quality

Data Sets

Base data set: MS-COCO

- 328k images
- objects labeled as one of 91 object types
- 5 single-sentence descriptions per image
- average: 3.5 categories and 7.7 object instances per image

Example image

Figure 2: Example image from MS COCO

Extension: V-COCO

- 10346 images from the MS COCO data set
- created for visual semantic role labeling
- action labels (26 different actions) + semantic roles (which object fills which semantic role)

Extension: COCO-a

- 10,000 images from MS COCO
- rich annotation of human actions and interactions
- on average 5.8 interactions per image

Interaction

- one subject
- one object
- several visual actions
- several visual adverbs (location, distance, emotion)

Example Attributes from COCO-a

Table 1: First Interaction from the example image

First Interaction	
Subject category:	person
Object category:	baseball bat
Visual actions:	hold, touch, wear
Visual adverbs:	full contact, in front

Implementation

Software Architecture

- Class Based
- Prepossessed Data is Saved
- Training Batches Saved Periodically
- Automated Evaluation
- Visualization of Evaluated Data
- Incremental Increase:
  - Input Complexity
  - Model Complexity

Software Architecture II

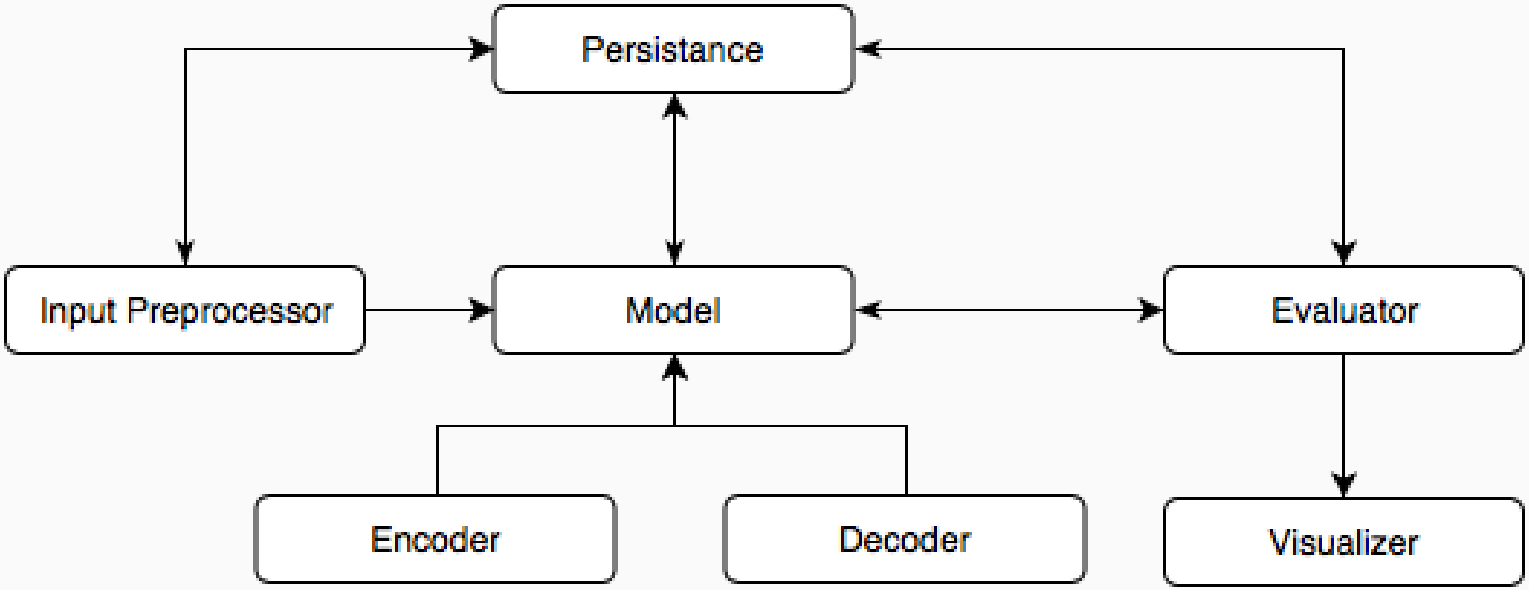

Figure 3: Architecture

Data Structure & Interfaces

Data Structure

- SQL Database for Persistence
- Converting Input Graph to One Hot Vectors

Interfaces

- Predefined
- Seperated Modules/Packages

Distribution of Tasks

Distribution of Tasks

- Preprocessing: Kim
- Basic model (Encoder and Decoder): Siting & Tatjana
- MS COCO Encoder: Tatjana
- V-COCO Encoder: Bente
- COCO-a Encoder: Kim & Siting
- Evaluation: Tatjana
- Visualization and statistics: Bente
- Hyperparameter: Tatjana, Siting, Bente & Kim

Timeline

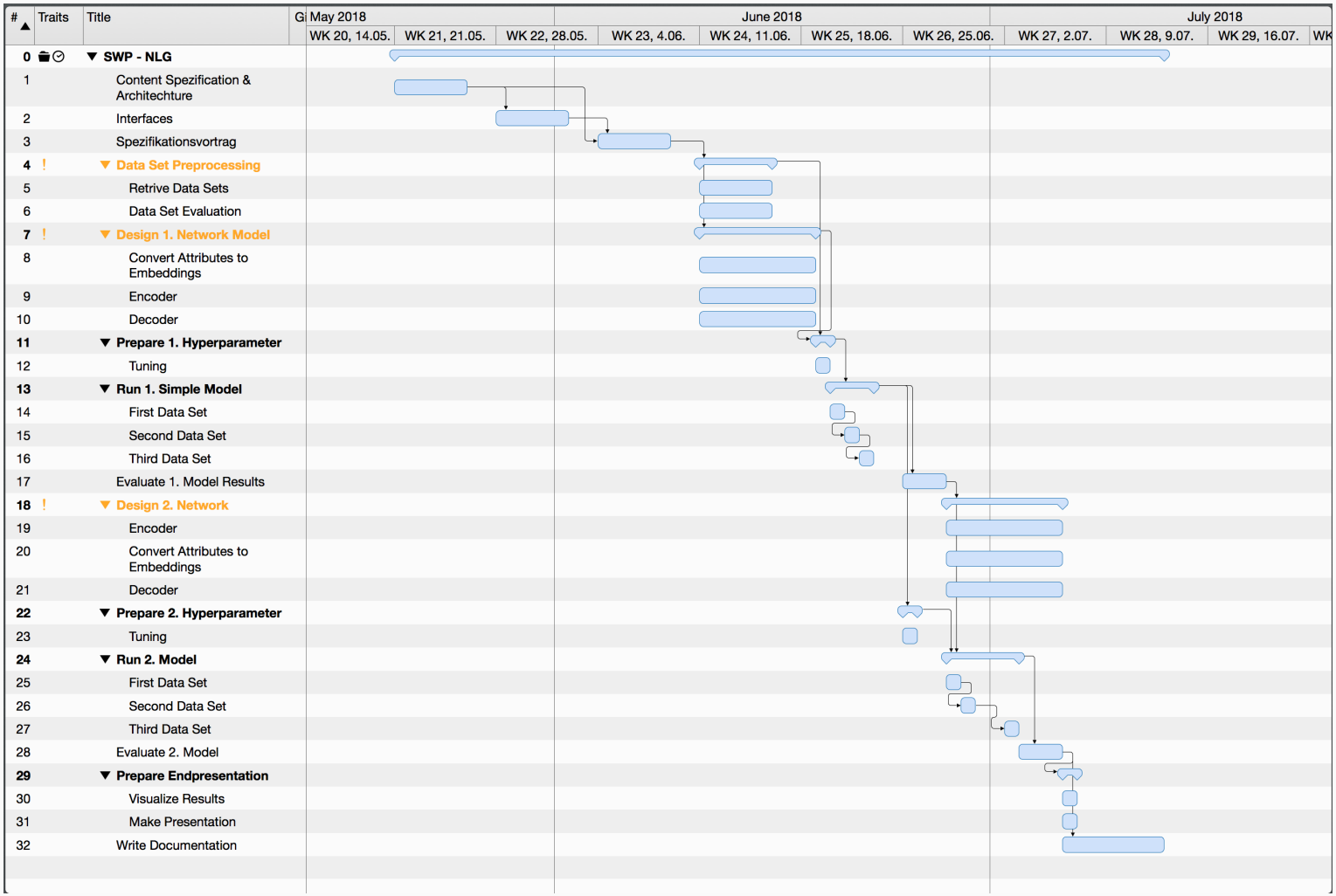

Figure 4: Gantt Diagramm

Ende

Fragen?