Understanding Language through Visual Imagination

by

Cheahuychou Mao

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2019

Author . . . .
Department of Electrical Engineering and Computer Science
August 23, 2019

Certified by . . . .
Boris Katz
Principal Research Scientist
Thesis Supervisor

Accepted by . . . .
Katrina LaCurts
Masters of Engineering Thesis Committee

Understanding Language through Visual Imagination

by

Cheahuychou Mao

Submitted to the Department of Electrical Engineering and Computer Science
on August 23, 2019, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

This thesis introduces a multimodal approach to natural language understanding by presenting a generative language–vision model that can generate videos for sentences and a comprehensive approach for using this capability to solve natural language inference, video captioning and video completion without task-specific training. The only training required is for acquiring a lexicon from captioned videos similar to the way children learn language through exposure to perceptual cues. The model generates videos by sampling the visual features of objects described in the target sentences over time. The evaluation results show that the model can reliably generate videos for sentences describing multiple concurrent and sequential actions, and that the ability to reason about language using visual scenes enables language tasks to be reduced to vision tasks and be solved more robustly using information obtained via vision.

Thesis Supervisor: Boris Katz
Title: Principal Research Scientist

Acknowledgments

I am truly thankful for the support and guidance that I have received from Boris Katz and Andrei Barbu. Boris, thank you for your investment in me throughout my MIT career. Andrei, I could not have done this thesis without your support and patience. Thank you both for all the kind words and encouragement when I doubted myself. To everyone else in the InfoLab whom I got a chance to work with especially Sue Felshin, I learned a lot from you and I am grateful for the opportunity to work on improving START. To my parents and brothers, thank you for believing in me and for your endless support.

Contents

1 Introduction **13**

1.1 Motivation 13

1.2 Research Problem 15

1.3 Thesis Roadmap 16

2 Related Work **17**

2.1 Video Generation 17

2.2 Natural Language Inference 18

2.3 Grounded Language Inference 20

2.4 Compositional Language–Vision Inference 21

3 Background **23**

3.1 Sentence Tracker 23

3.2 Natural Language Inference 25

4 Approach **29**

4.1 Generative Language–Vision Model 29

4.2 Training Procedure 32

4.2.1 LAVA Corpus 32

4.2.2 Lexicon 33

4.2.3 Track Generation 34

4.2.4 Training 35

4.3 Video Generation 35

4.4 Video Completion 37

4.5 Caption Generation 37

4.6 Paraphrases and Natural Language Inference 38

5 Evaluation 41

5.1 Video Generation 41

5.2 Video Completion 43

5.3 Caption Generation 44

5.4 Paraphrases and Natural Language Inference 45

5.4.1 Grounded NLI Corpus 45

5.4.2 Baselines 46

5.4.3 Results 47

6 Discussion 49

6.1 Summary 49

6.2 Challenges 49

6.3 Future Work 50

6.4 Contributions 52

A Tables 53

List of Figures

4-1 Two example videos from the LAVA corpus. 33

4-2 Example visualizations of extracted tracks. 35

4-3 An example video generated by the generative language–vision model. Each video has 12 frames in total but for readability only every other frame is shown here. This video is for the sentence “The person carried the telescope toward the chair”. The movement of the legs and the shrinking distance between the person and the chair illustrate that the person is walking toward the chair. 37

4-4 Three pairs of sentences and each of their six pairwise relationships. When finding the relationship between a pair of sentences, videos are sampled from the premise and then are scored on both the premise and hypothesis. The scores here are sorted in descending order by the premise score. 39

5-1 Videos generated by the generative language–vision model. 42

5-2 Two different generated videos for the exact same sentence. 43

5-3 An example original and completed video pair. The top video is the original video, and the bottom video is completed video. The frames in red are the prefix frames given to the model. The look of the person is randomized every time so that difference here can be ignored. . . . 43

List of Tables

3.1 Three example sentence pairs taken from the SNLI corpus. 26

3.2 NLI predictions by AllenNLP [13] implementation of Parikh's Decomposable Attention model with ELMO for tricky sentence pairs. 27

4.1 Characteristics of word models. POS stands for part of speech, I is the arity, K is the number of states, and M is the number of components for a Gaussian mixture model. 31

4.2 The lexicon of our model. 34

5.1 The descriptions for the templates used to generate the dataset. . . . 46

5.2 The accuracy of models on the SNLI test set and our grounded NLI corpus broken down by category. Note that the SNLI accuracies are not included in the overall mean and standard deviation calculation. . 48

A.1 The templates used to generate sentences in our corpus. Here P stands for spatial-relation preposition, whereas $\mathrm{P_M}$ stands for motion preposition. For simplicity, we hide the optional PP in the NP in the templates above, and only expand the NP when we need to show the difference between sentences. 53

Chapter 1

Introduction

1.1 Motivation

Vision plays a fundamental role in language understanding and acquisition. When we acquire language as children, we depend on visual contexts to infer the meanings of sentences that we hear. We later learn to map words to internal meaning representations that refer to concepts in the physical world [28]. For example, we learn that the word "ball" refers to a solid or hollow spherical object that can roll. We also learn that the word "push" refers to the act of exerting force against something to cause it to move forward, and that it is opposite to the word "pull" which refers to the act of exerting force on something to cause it to move toward oneself. The connection between language and vision enables us to *ground* language in perception and have a non-linguistic way to reason about language.

In the realm of machine learning, language understanding tasks are traditionally regarded as linguistic tasks and are performed using advanced linguistic approaches. There exist very few approaches that exploit the intrinsic connection between language and vision. For instance, consider the task of natural language inference (NLI) [22] whose goal is to classify the relationship between two sentences as *entailment*, *contradiction* or *neutral*. The standard approaches generally involve learning to capture and classify the syntactic and semantic relationship between sentences through deep representations. Such approaches require large training data with tens of thousands of examples, yet are shown to make systematic errors for sentences with large lexical overlap, similar syntactic structures and subtle meaning differences [24, 8]. These errors include misclassifying sentences such as “The girl is wearing a black shirt and blue pants” as implying “The girl is wearing black pants” and “The man is staring at the clear sky” as not contradicting “The sky is cloudy”. Capturing such subtle differences and similarities remains a challenge in representation learning.

Unlike machines, we humans have multimodal ways of reasoning about language that enable us to robustly handle all kinds of tasks. One way is to use visual imagination. That is, we can imagine visual scenes or *possible worlds* for sentences and use those scenes to make inferences about the sentences. Through visual information, subtle differences and similarities such as the ones described above are often directly inferable. For example, given the sentence “The bear chases after the man with the honey”, we can easily visualize it with a scene where the man runs away from the bear with a jar of honey in his hand, or a scene where the man rows a boat loaded with honey as the bear swims after him. Using these visual scenes, we can easily infer that sentences such as “The man moves away from the bear with the honey” and “The man escapes from the bear” are paraphrases of the original sentence; despite the syntactic differences these are just other valid ways of describing the imagined scenes. Likewise, we can infer that sentences such as “The man chases the bear” and “The bear looks after the man with the honey” contradict the original sentence; despite the lexical overlap these sentences are visibly not true of the imagined scenes. Through visualizing scenes, we can also make other useful inferences that are only inferable through vision such as “The man and the bear move in the same direction” and “The man is in front of the bear”; we can also predict the upcoming scenes and sentences such as “The bear catches up with the man” and “The bear rushes toward the man and grabs the honey from him”. Such inferences can be quite difficult to make without information obtained via vision and knowledge of the physical implications of actions.

The tasks that visual imagination enables us to solve, as described above, are analogous to the machine learning tasks of video generation, video completion, video captioning, paraphrase recognition and natural language inference. These are popular tasks that have applications ranging from commonsense reasoning to question answering. Contrary to machines, we do not need to be presented with thousands of examples to be able to solve the tasks. That is, we only need to know the visual implications of the words in the sentence, visualize scenes for it, and make inferences through vision. In this work, we emulate humans' ability to visualize scenes for visually descriptive sentences in order to solve language–vision tasks. We envision this as a critical step toward a multimodal language understanding approach that can enhance the performance of machine learning in language and vision tasks and eliminate the need for large task-specific training data.

1.2 Research Problem

We formalize the notion of visual imagination by presenting a generative language–vision model that can synthesize videos conditioned on sentences. We show that visual imagination enables this grounded model to solve a set of vision and language tasks (namely video completion, video captioning and natural language inference) without task-specific training and still achieve performance comparable to models that are trained on large datasets. We argue that this provides a cognitively plausible explanation for why humans are capable of forming generalizations and performing tasks that they are never specifically trained to solve.

To do this, we make use of Sentence Tracker introduced by Siddharth et al. [34] and Yu et al. [39] to perform joint inference on sentences and generated videos. During training, the model acquires a lexicon from captioned videos similar to the way children learn language from visual cues. The video generation model is a Markov Chain Monte Carlo-based model that generates videos by sampling visual features (see Table 4.1) for the participants in the event described in the sentence. We evaluate the tasks above both qualitatively and quantitatively, as applicable. For the task of NLI, we created our own evaluation corpus of sentence pairs that can be grounded in vision and that are difficult for standard approaches. We show that our approach can not only achieve results comparable to state-of-the-art models but also handle sentence pairs in some categories more robustly due to the information obtained via vision.

To summarize, the contributions of this work are:

- a novel application of vision to language where language tasks are reduced to vision tasks
- a generative language–vision model that can synthesize videos for visually descriptive sentences
- an approach for completing the missing segments of videos
- an approach for video caption generation for complete and incomplete videos
- a visual approach to NLI and paraphrase recognition
- a grounded NLI evaluation corpus that is difficult for existing approaches

1.3 Thesis Roadmap

The remainder of this thesis is organized as follows. Chapter 2 describes existing video generation, natural language inference and grounded inference approaches. Chapter 3 provides background on the Sentence Tracker approach and the natural language inference task. Chapters 4 and 5 describe in detail the approach for video generation, video completion, caption generation and natural language inference along with the evaluation procedure and results. Chapter 6 discusses the challenges and future directions for the work.

Chapter 2

Related Work

2.1 Video Generation

Recent advances in deep generative approaches such as Generative Adversarial Networks (GANs) [15] have led to significant progress in both image and video generation. Yet, there are still notably fewer successful approaches for video generation than image generation. In addition, existing approaches are not yet able to generate videos with multiple participants interacting with one another. This is primarily due to the complexity of motions and the lack of standardized captioned video datasets [20]. In this section, we describe a few of the standard approaches.

One common approach for video generation is to decompose each video into a static background and a dynamic foreground [35, 20]. For example, Li et al. [20] introduced an approach for generating videos conditioned on captions by training a conditional generative model to extract both static and dynamic information from text. To generate a video, a conditional variational autoencoder [19] model is used to first generate an image that gives the background color and object layout of the target video. After that, the content and motion of the video is generated by a GAN generator by conditioning on both the text and generated image. This approach requires a large training dataset of captioned videos and only produces low resolution videos without human poses.

There are several existing approaches for human pose video generation; however, those approaches can only generate one single pose sequence so cannot produce realistic videos with multiple participants. For instance, Yang et al. [37] presented a GAN-based approach for human pose and facial expression video completion. The method also consists of two steps. First, the pose is extracted from the input image and then a Pose Sequence GAN is used to generate a temporally smooth pose sequence conditioned on the extracted pose and the target action class. Next, a Semantic Consistent GAN is used to generate realistic and coherent video frames conditioned on the input image and generated pose sequence. Similarly, Cai et al. [5] proposed a GAN-based approach that first generates a pose sequence for the target action, then uses a supervised reconstruction network with feature matching loss to transfer the pose sequence to the pixel space in order to generate a complete video. These approaches also require large training datasets.

Unlike the approaches above, our video generation approach is grounded in vision and can robustly handle concurrent and sequential actions involving multiple participants. It generates videos by directly sampling the visual features (instead of the individual pixels) for the participants in each frame using Markov Chain Monte Carlo (MCMC) methods. It relies on the compositionality of language and events to perform video generation. Therefore, the parameters of the model are not the weights of deep convolutional networks but the parameters of Hidden Markov Models (HMMs) [30] that encode the meaning and physical implications of each word. As a result, our model requires less training data and is easier to train.

2.2 Natural Language Inference

Natural Language Inference (NLI) [22] is a popular language understanding task. The goal is to classify the relationship between a pair of sentences as *entailment*, *contradiction*, or *neutral*. This task is effectively a variant of the paraphrase recognition task where *paraphrases* corresponds to entailment, and *not paraphrases* corresponds to neutral or contradiction. It is a linguistic task and is primarily solved using language models. The standard approach is to use deep representation learning with attention and memory. Such NLI approaches are very sophisticated in architecture and have achieved impressive performance on benchmark datasets. Here we describe the approaches that we used as the baselines for our approach.

The decomposable attention model by Parikh et al. [25] is a representative alignment and attention-based approach that solves the task by creating a soft alignment matrix for the sentence pair using neural attention and using it to decompose the task into subproblems that are solved separately. The approach does not incorporate word order information but outperformed more complex neural architectures when it was introduced. ESIM by Chen et al. [10] is an encoder-based approach that uses enhanced bidirectional LSTMs [17] to encode sentences and perform local and compositional inference. This approach also achieves very promising performance on the task. In addition, Peters et al. [27] demonstrated that this performance can be further enhanced through the use of deep contextualized word embeddings (ELMO) in place of GloVe embeddings [26].

Infersent, introduced by Conneau et al. [11], is a universal sentence representation approach that uses a bidirectional LSTM with max pooling trained on a benchmark NLI corpus (SNLI, described in Section 3.2). It generalizes to many different transfer tasks ranging from sentiment analysis to caption-image retrieval. Similarly, the Bidirectional Encoder Representations from Transformers (BERT) [12] is a self-attention transformer model for pretraining deep bidirectional text representations by jointly conditioning on both left and right context. BERT is designed to easily fine-tune to a broad range of language understanding tasks including NLI. It is considered one of the leading approaches for language representation.

To date, all benchmark NLI datasets have best test accuracy of over 85%. The rapid increase in performance has led to a lot of interest in doing controlled evaluation of the datasets. Those works have uncovered various fundamental weaknesses in the datasets and the existing approaches. Notably, Gururangan et al. [16] and Poliak et al. [29] both showed that popular NLI datasets contain statistical irregularities that allow hypothesis-only models—ones that completely ignore the premise—to significantly outperform the majority class baseline. Similarly, Cer et al. [8] of SemEval

Semantic Textual Similarity, featuring a task much similar to NLI, observed that participating models shared a set of systematic errors on sentence pairs that involve negation, agency, spatial relations and more. These results highlight the limitations of language models in solving inference tasks, and in this work we argue that reducing language tasks to vision tasks can help overcome this obstacle.

2.3 Grounded Language Inference

There exist some prior works that combine language and vision in solving inference tasks similar to NLI. However, they require task-specific training and can only reason about static scenes. The denotational similarity metric introduced by Young et al. [38] is one such example. The core of this approach is the notion of a denotation defined to be the set of possible worlds represented by images in which a visually descriptive sentence is true. The similarity between a premise and hypothesis is computed using a hierarchical denotation graph constructed with training images and sentences based on the partial ordering induced by denotations. The approach can correctly capture the relationships between concepts such as sitting and eating lunch, and walking up stairs and walking down stairs. However, it does not do inference by generating static or dynamic scenes by itself.

Other works that combine language and vision are in the realm of visual question answering [1] and grounded commonsense inference [40]. These tasks are considered very challenging for existing approaches. The closest work to our own is that of Lin and Parikh [21] who pointed out that visual paraphrasing can be solved using visual commonsense learned from captioned images. This work requires task-specific training data and does not incorporate motion, so it cannot model consequences or sequences of actions extended in time. In essence, the model solves paraphrase recognition by imagining a scene in the form of an image for each sentence, and leveraging visual information from the imagined scenes to compare the sentences. This approach is different from the one presented in this work both mathematically and practically. Our model solves paraphrase recognition or NLI by imagining and reasoning about all possible dynamic scenes for the sentence pair. In addition, it generates visual scenes by using Markov Chain Monte Carlo methods with the scoring function given by a compositional event-tracking model, as opposed to the Conditional Random Field method with the scoring function used by Lin and Parikh.

2.4 Compositional Language–Vision Inference

The grounded language–vision model presented in this work is a variant of the Sentence Tracker developed by Siddharth et al. [34] and Yu et al. [39]. The approach utilizes the compositional structure of language and events to drive grounded language inference. It is nominally generative although the inference algorithm used in prior work turns it into a classifier. Section 3.1 provides in depth description of the approach. In short, at training time the model acquires a lexicon from captioned videos similar to the way children acquire language through exposure to perceptual cues. At inference time, each sentence is represented as compositions of words mediated by a grammar and inference is performed using a function that computes the likelihood that a video depicts a sentence. The Sentence Tracker has been applied to a wide range of language and vision tasks such as tracking, video retrieval, ambiguity resolution, and grounded language acquisition [34, 39, 3, 32]; however, the approach presented in this work is its first generative extension.

Chapter 3

Background

In this chapter, we provide background on the Sentence Tracker, an approach that our work is based on. We also describe the task of natural language inference and its existing benchmark datasets.

3.1 Sentence Tracker

The Sentence Tracker [34, 39] is a general-purpose approach for performing multi-object tracking and event recognition through simultaneously reasoning about a video clip and a natural language sentence. The core of this approach is the scoring function $\mathcal{S} : (\mathbf{B}, \mathbf{s}, \Lambda) \to (\tau, \mathbf{J})$. This function takes in a video (in the form of an overgenerated set of detections $\mathbf{B}$) along with a sentence $\mathbf{s}$ and a learned lexicon Λ, and outputs the likelihood τ that the video depicts the sentence and the sequences of detections $\mathbf{J}$ (henceforth referred to as tracks) that satisfy the sentence while maximizing the aggregate detection score and temporal coherence.

To do that, given a sentence, the Sentence Tracker parses it using an off-the-shelf dependency parser and uses the parse to structure a sentence-specific graphical model consisting of a hierarchical Factorial Hidden Markov Model (FHMM) [14] in which one layer physically locates objects in the video frames while the other observes the resulting tracks and enforces sentential meaning. In essence, the object-tracking layer consists of HMM-based trackers for all the event participants, each with tracks as the latent states, and the detection score and the temporal-coherence score acting analogously as the output probability and the state-transition probability. The other layer consists of word models for all the words in the sentence; each model observes a time series of video features extracted from at least one track in the other layer. Each word model is a discrete HMM with a small number of learned states, a banded-diagonal state transition matrix, and an output model that recognizes quantized features relevant to that word. The structure of each word model depends on the part of speech and the meaning of the word it represents. For example, a verb such as "approach" or "carry" is represented as a multi-state two-participant HMM whose outputs include the velocity and orientation of the participants, whereas an adjective such as "large" or "short" is represented as a single-state one-participant HMM whose outputs include the dimensions and aspect ratio.

The two layers are connected based on a linking function as given by the dependency parse. Together these individual trackers and word models constitute a factorial HMM that encodes the static and dynamic properties of the participants of the event described by the sentence, and that is very sensitive to subtle changes in sentential meaning. For example, consider the sentence "The person approached the green chair to the left of the table". The factorial HMM for this sentence consists of the trackers for the participants *person*, *chair* and *table*, and the word models for the nouns "person", "chair" and "table", the verb "approach", the adjective "green" and the preposition "to the left of". Each word has one or more arguments mapped to tracks by the linking function. Namely, the noun models for "person", "chair" and "table" observe the tracks for the participants they represent. The adjective model for "green" observes the track for *chair*. The preposition model for "to the left of" observes the tracks for *chair* and *table*. The verb model for "approach" observes the tracks for *person* and *chair*. These relationships result in constraints on the interactions between objects in the video, and a graphical model that recognizes the occurrence of the events described by the sentence.

Let L be the number of participants, T be the number of frames in the video, W be the number of words in the sentence, and $J = \langle j_1, \cdots, j_L \rangle$ and $K = \langle k_1, \cdots, k_W \rangle$

be the candidate tracks and states respectively. The log-likelihood that the sentence is true of the video is computed using the Viterbi algorithm as

$$\begin{aligned}\max_{J,K}\Bigg[&\sum_{l=1}^{L}\left(\sum_{t=1}^{T} f(b^{t}_{j^t_l}) + \sum_{t=1}^{T} g(b^{t-1}_{j^{t-1}_l}, b^{t}_{j^t_l})\right)\\ &+\sum_{w=1}^{W}\left(\sum_{t=1}^{T} h_{s_w}(k^t_w, b^t_{j^t_{\theta^1_w}}, \cdots, b^t_{j^t_{\theta^{I_{s_w}}_w}}) + \sum_{t=1}^{T} a_{s_w}(k^{t-1}_w, k^t_w)\right)\Bigg]\end{aligned} \tag{3.1}$$

where b represents a candidate detection, θ represents the linking function, and I_e represents the arity (number of arguments) of the lexical entry e. The first term corresponds to a detection-based tracker where $f(b)$ is the detection score for detection b in log space, and $g(b^{t-1}, b^t)$ is the temporal-coherence score between adjacent-frame detections b^{t-1} and b^t in log space. In the original Sentence Tracker [34, 39], f is defined as the logarithm of the normalized detector score output by the object detector, whereas g is defined as the logarithm of the normalized Euclidean distance between adjacent frames found using optical flow. The second term corresponds to an event-recognition model where $h_e(k, b_1, \cdots, b_{I_e})$ is the log probability of observing a set of detections when the lexical entry e is in state k, and $a_e(k^{t-1}, k^t)$ is the log probability that the lexical entry e transitions from state k^{t-1} to k^t. These HMM parameters are learned from captioned videos using the Baum Welch algorithm [2] and are stored in the lexicon Λ. The time complexity of this is exponential in the number of participants L and the sentence length W; however, these numbers are bounded in practice.

3.2 Natural Language Inference

The task of Natural Language Inference (NLI), also known as Recognizing Textual Entailment (RTE), is a fundamental task in natural language understanding. The goal is to classify the relationship between a sentence pair, referred to as premise and hypothesis, into one of three classes: *entailment*, *contradiction* or *neutral.* Fundamentally, *entailment* means that the hypothesis is true given the premise, *contradiction*

Premise	:	A group of kids are splashing in deep water nearby a rock formation.
Hypothesis	:	The kids are in deep water.
Label	:	Entailment
Premise	:	Four women are taking a walk down an icy road.
Hypothesis	:	Four women are walking near the dry highway.
Label	:	Contradiction
Premise	:	An elderly woman is preparing food in the kitchen.
Hypothesis	:	A person makes dinner.
Label	:	Neutral

Table 3.1: Three example sentence pairs taken from the SNLI corpus.

means that the hypothesis is false given the premise, whereas *neutral* means that the trueness of the hypothesis cannot be determined given the premise. Table 3.1 contains some example sentence pairs taken from a benchmark NLI corpus.

NLI is useful for a wide range of applications such as paraphrase recognition, commonsense reasoning, and question answering. However, it is a challenging language understanding task as it involves dealing with the complexity of compositional semantics, quantification, coreference, and lexical and syntactic ambiguity. Nonetheless, in recent years there has been a significant increase in the performance on NLI, thanks to the release of large benchmark datasets such as the Sentences Involving Compositional Knowledge (SICK-E) corpus [23], the Stanford Natural Language Inference (SNLI) corpus [4], and the Multi-Genre Natural Language Inference (MultiNLI) corpus [36]. SICK-E contains 10k sentence pairs that are partly automatically constructed from image captions and video descriptions. SNLI contains 570k sentence pairs that are manually annotated by humans based solely on images captions. MultiNLI contains 433k sentence pairs compiled similarly to SNLI but covers both written and spoken speech in a wide range of styles and topics instead of descriptions of visual scenes.

As described in Section 2.2, existing NLI models vary widely in architecture but generally make use of alignment, encoder and attention-based approaches to capture the semantic and syntactic relationship between each premise–hypothesis pair. Despite impressive performance on the benchmark datasets, those models share some

Premise	:	The man in blue is standing behind the man in red.
Hypothesis	:	The man in blue is standing in front of the man in red.
Predicted Label	:	Entailment
Actual Label	:	Contradiction
Premise	:	The girl with blue goggles is swimming without a swimming cap.
Hypothesis	:	The girl is wearing goggles and swimming cap.
Predicted Label	:	Entailment
Actual Label	:	Contradiction
Premise	:	The man is staring at the clear sky.
Hypothesis	:	The sky is cloudy.
Predicted Label	:	Neutral
Actual Label	:	Contradiction
Premise	:	The man is carrying a canoe with a dog.
Hypothesis	:	The dog is carrying the man in a canoe.
Predicted Label	:	Entailment
Actual Label	:	Contradiction
Premise	:	The man placed the book by his backpack.
Hypothesis	:	The man put the book in his backpack.
Predicted Label	:	Entailment
Actual Label	:	Contradiction
Premise	:	The dog is sleeping under the table.
Hypothesis	:	The dog is on the table.
Predicted Label	:	Entailment
Actual Label	:	Contradiction
Premise	:	The girl is wearing a black shirt and blue pants.
Hypothesis	:	The girl is wearing black pants and a blue shirt.
Predicted Label	:	Neutral
Actual Label	:	Contradiction

Table 3.2: NLI predictions by AllenNLP [13] implementation of Parikh's Decomposable Attention model with ELMO for tricky sentence pairs.

major limitations. Table 3.2 illustrates some incorrect predictions made by a state-of-the-art NLI system for a set of manually constructed sentence pairs. Such sentence pairs are however very rare in the large NLI corpora noted above. To address this, McCoy et al. [24] compiled a dataset, named Heuristic Analysis for NLI Systems (HANS), containing 30k sentence pairs with three fallible syntactic heuristics: a lexical overlap heuristic, a subsequence heuristic, and a constituent heuristic. As expected, they found that state-of-the-art models performed poorly on the dataset. In this work, we focus on sentence pairs with similar characteristics but that are grounded in vision.

Chapter 4

Approach

We present a generative language–vision model that can synthesize videos for sentences describing actions, and use this capability to perform a range of language–vision tasks without seeing training examples. In the sections below, we describe in detail the model architecture, the training procedure, and the approach used to solve the tasks of video generation, video completion, caption generation and natural language inference.

4.1 Generative Language–Vision Model

Our language–vision model is a variant of the Sentence Tracker described in Section 3.1. Our approach utilizes the compositional structure of language and events to drive grounded language understanding, and is generative in nature. We seek to reformulate the scoring function and the structure of the sentence-specific graphical model to work with generated videos so that the scoring function can be used as the likelihood function in video sampling. To this end, we modified the Sentence Tracker in several ways. We removed track finding from the original algorithm since here tracks are generated by the model itself. For similar reasons, trackers no longer need to enforce the detection score constraint and for simplicity use Gaussian distributions for the temporal-coherence constraint instead. To provide more fine-grained and realistic understanding of actions, we expanded the latent space of trackers to include not only the center coordinates and the dimensions of the participants but also the coordinates of the hands and feet if applicable. We also removed state sequence finding from the algorithm since we can just directly specify the states for the sampled video at the start. Let C^t_l be the coordinates and dimensions of the participant l in frame t, the likelihood of a video–sentence pair is now:

$$\begin{aligned}\left[\sum_{l=1}^{L}\left(\sum_{t=1}^{T} g(C_l^{t-1}, C_l^t)\right) + \sum_{w=1}^{W}\left(\sum_{t=1}^{T} h_{s_w}(k_w^t, C^t_{j^t_{\theta^1_w}}, \cdots, C^t_{j^t_{\theta^{I_{sw}}_w}})\right.\right.\\ \left.\left. + \sum_{t=1}^{T} a_{s_w}(k_w^{t-1}, k_w^t)\right)\right] \end{aligned} \tag{4.1}$$

where the set of state sequences $K = \langle k_1, \cdots, k_W \rangle$ is prespecified.

In addition, instead of using a hand-built grammar, we use the Stanford dependency parser [9] to get the government relations directly and handcraft a set of rules to handle ambiguities. This enables us to handle a larger range of linguistic phenomena such as conjunction and passive voice, which is crucial for all of our target tasks. We still use a lexicon with a moderate number of nouns, verbs, adjectives, spatial-relation prepositions and motion prepositions. However, we introduced a procedure for mapping unknown words to synonyms in our lexicon by finding the word in the lexicon whose GloVe embedding has the closest cosine distance to the embedding of the unknown word. We seek to recognize more complicated events described in multiple sentences. Thus, we generalized the scoring function to handle multiple sentences with multiple interpretations (linking functions) as follows:

$$\begin{aligned}\sum_{s \in S}\left\{\sum_{l=1}^{L_s}\left(\sum_{t=1}^{T} g(C_l^{t-1}, C_l^t)\right) + \max_{\theta \in \Theta_s}\left[\sum_{w=1}^{W_s}\left(\sum_{t=1}^{T} h_{s_w}(k_w^t, C^t_{j^t_{\theta^1_w}}, \cdots, C^t_{j^t_{\theta^{I_{sw}}_w}})\right.\right.\right.\\ \left.\left.\left. + \sum_{t=1}^{T} a_{s_w}(k_w^{t-1}, k_w^t)\right)\right]\right\} \end{aligned} \tag{4.2}$$

where S is the set of all sentences and Θ_s is the set of all linking functions for all interpretations of the sentence s.

Here each word model is a discrete and/or multivariate Gaussian HMM with a

POS	I	K	M	Features
Noun	1	1	1	object class (discrete), width, height, aspect ratio, x distance between hands/feet and center to width ratio, y distance between hands/feet and center to height ratio
Adjective	1	1	-	color index (discrete)
Transitive Verb	2	3	2	velocity and orientation of participants, relative velocity and orientation of participants, distance between participants, distance from hands/feet of agent to center of agent, distance from hands/feet of agent to center of patient, distance from hands/feet of agent to edges of patient, change in width and height of agent
Intransitive Verb	1	3	2	velocity and orientation, distance from hands/feet of agent to center of agent, change in width and height
Preposition	1	1	2	x, y and z distance between participants change in width and height
Motion Preposition	2	3	2	velocity and orientation of participants, relative velocity and orientation of participants, distance between participants

Table 4.1: Characteristics of word models. POS stands for part of speech, I is the arity, K is the number of states, and M is the number of components for a Gaussian mixture model.

banded-diagonal state transition matrix (no state skipping), and an output model that recognizes video features representing the meaning of the word. In other words, we do not quantize continuous features as done in the original Sentence Tracker because having discrete features could make video sampling very difficult. Table 4.1 shows the characteristics of the word models in our lexicon. The features are carefully chosen to capture the static and dynamic properties of the participants as implied by the meanings of the words. The noun and adjective models have discrete features over the set of object classes and colors respectively, whereas the other models have only continuous features. To handle directions, we use Gaussian mixture models where appropriate, von Mises distributions for orientation features, and truncated normal distributions for the distance features in the spatial-relation preposition models.

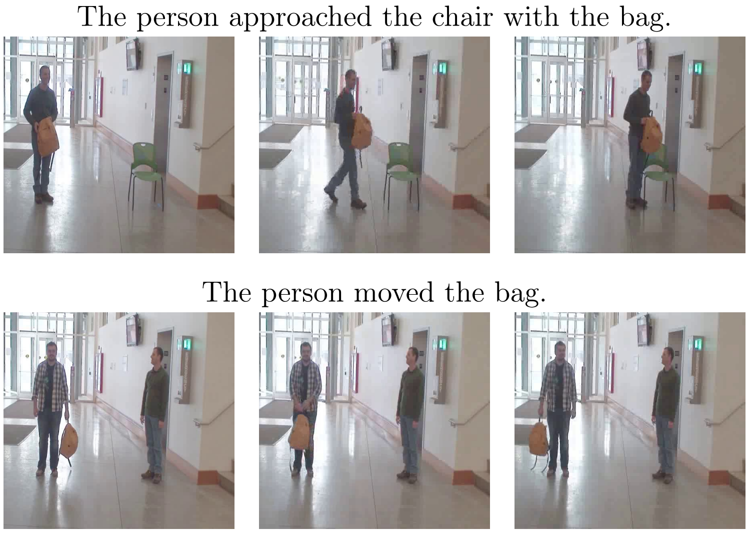
The parameters of the word models are learned from captioned videos using the Baum Welch algorithm with priors on the variance of the key features. The learned lexicon represents the model's grounded understanding of language similar to that of children. That is, the model has knowledge of the perceptual implications of the words in the physical world. For instance, it knows that "carry" implies that there is an agent supporting and moving a patient from one place to another, and that "toward" implies that the distance shrinks over time, etc. This provides it with the ability to recognize actions and do perceptual reasoning.

4.2 Training Procedure

The training of this generative language–vision model refers to the learning of the parameters of the word models which include the discrete output probability B for each discrete HMM and the mean μ, variance Σ, and kappa κ for each multivariate Gaussian HMM (the state transition matrix A and the initial and final state distributions π_i and π_f are predefined). The model learns the meanings of words from captioned videos similar to the way children acquire language through exposure to perceptual contexts [34, 39]. We make use of the Baum Welch algorithm computed in log space for this training. Additionally, we place priors on the variance of the key features of each Gaussian word model to tackle the noise in the videos and object detectors. For training, the model needs as input, for each video–sentence pair, the dependency parse and the tracks for all the participants. The next subsections detail how we went from captioned videos to tracks, and finally did the training.

4.2.1 LAVA Corpus

Our approach can work with any captioned-video dataset. In this work, we make use of the LAVA (Language and Vision Ambiguities) corpus [3]. The corpus consists of 1,679 video–caption pairs depicting a range of actions performed by humans, namely *approach*, *leave*, *hold*, *move*, *pick up*, *put down* and *look at*. The objects present in the videos include *person*, *chair*, *bag*, and *telescope* in *yellow* or *green*. The covered spatial relations include *on* and *with*. Figure 4-1 shows two example video–caption pairs from the LAVA corpus. The captions are created based on a fixed lexicon and



The person approached the chair with the bag.

The person moved the bag.

Figure 4-1: Two example videos from the LAVA corpus.

grammar but cover a wide range of syntactic, semantic and discourse ambiguities. Each caption comes with a parse tree along with a manually-annotated visual setup description. Each video is about 3 seconds in length, containing about 90 frames. For simplicity and efficiency, we aim to sample videos with only 3 frames per second so we downsampled the frames in each video accordingly at training time. The actions are shot with multiple actors and from different angles and directions where applicable. To handle more actions and spatial relations, we expanded the corpus by annotating the videos with additional captions to cover concepts such as *carry*, *run*, *bend down*, *to the left of*, *to the right of*, *toward*, *away from*, *below*, etc.

4.2.2 Lexicon

Our lexicon consists of all the words in the LAVA corpus and the words introduced through the additional captions. To handle a larger set of sentences, we added synonyms of the existing words to the lexicon. We also handcrafted the parameters for a handful of additional adjectives and nouns as they are either just categorical distributions (color index and object class) or easily derivable from the parameters (width, height, and aspect ratio) of the other trained words. Table 4.2 shows the words in the resulting lexicon.

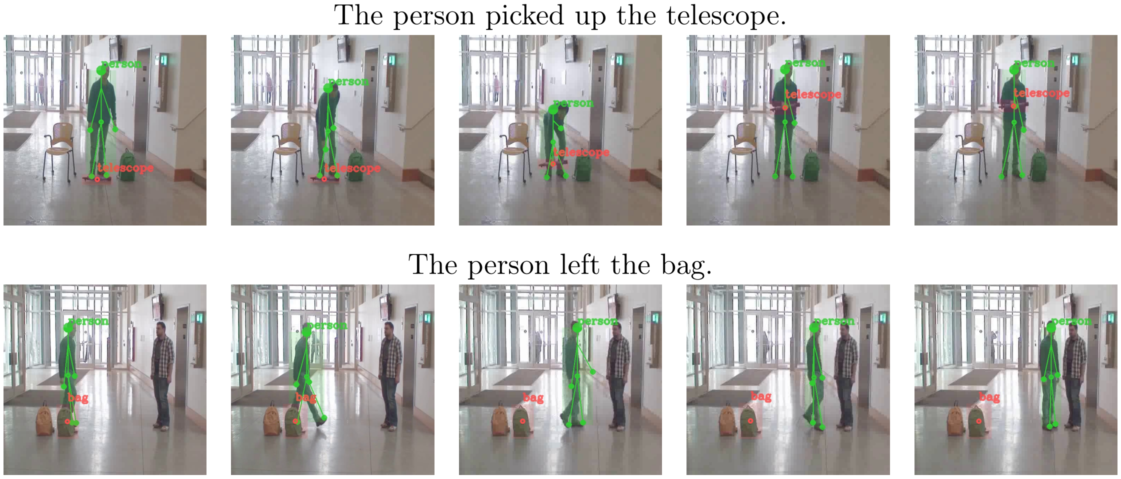
Part of Speech	Words
Noun	person, dog, cat, monkey, bag, chair, telescope, car, table, bike, pizza
Transitive Verb	approach, pass by, go near, walk past, leave hold, clasp, clutch, grasp, grip move, displace, pick up, raise, lift, put down look at, glance at, gazed at, stare at carry, transport
Intransitive Verb	run, sprint, race, jog, walk, march, stand, bend down
Preposition	to the left of, to the right of, behind, in front of, above, on, below, under, near, next to, far away from
Motion Preposition	toward, into, away from
Adjective	yellow, green, blue, red, brown, black, white

Table 4.2: The lexicon of our model.

4.2.3 Track Generation

To extract tracks from videos, we started by training object detectors to get bounding boxes for objects in the LAVA corpus. For that, we used YOLOv3 [31] and its default pretrained model. We sampled around 450 random images from the LAVA videos and annotated all the objects using the LabelMe annotation tool [33]. To extract the positions of hands and feet, we made use of OpenPose [6] to get the positions of the keypoints of all humans that appear in the videos. The LAVA corpus contains some videos with multiple objects in the same object class. So to enable partly automatic track extraction, we manually annotated the first frames of those videos. The last step is to combine all the detections to generate coherent tracks.

To construct tracks for a video, we parsed its caption using the Stanford parser to identify the event participants. Then, we initialized the tracks by starting with the detections in the first frame for all the participants. For participants with multiple detections, we relied on the manual annotations mentioned earlier to select the right detection. Next, we extended the tracks by either selecting the nearest detection in the same object class in the next frame or doing forward projection if there is no detection. We found that the object detectors and OpenPose were not always reliable and led to a notable number of bad tracks. Hence, we visualized the tracks one by one and manually excluded the videos with incorrect tracks from training. We ended up



The person picked up the telescope.

The person left the bag.

Figure 4-2: Example visualizations of extracted tracks.

training with only 250 videos in the corpus. This was sufficient as our model does not require a large amount of training data. Figure 4-2 shows some example visualized tracks that were generated by our script.

4.2.4 Training

For each video–caption pair in the corpus, we parsed the caption using the Stanford parser and transformed its dependency parse to our custom representation of its linking function, and then we cached the parse along with the tracks extracted from the downsampled video as detailed above. We trained the model on this processed data using the Baum Welch algorithm and stored the best model for later use. In other words, we did not train the model for each task separately. Each task can be performed solely because of the learned lexicon, the scoring function and the video generation capability which is described in the next section.

4.3 Video Generation

The primary novel contribution of this work is the ability to synthesize videos conditioned on sentences. It is analogous to humans' ability to imagine scenes for sentences based on their knowledge of the perceptual implications of the words in those sentences. To generate a video for a sentence, we parse the sentence using the Stanford parser and create a hierarchical graphical model based on its linking function as usual. Multi-sentence stories are handled similarly. Next, we perform joint inference using Markov Chain Monte Carlo methods to sample a 12-frame video for each interpretation of the sentence. That is 4 seconds in length at 3 frames per second. We do not sample individual video pixels; instead we sample the width, height, color, and coordinates of the center (also the hands and feet if applicable) of all the participants in each frame. The model was implemented in the probabilistic programming language Stan [7] and inference is performed using a No-U-Turn sampler (NUTS) [18] with the Sentence Tracker scoring function as the log-likelihood function. We initially experimented with Gibbs sampling and Metropolis-Hastings but we found that Hamiltonian Monte Carlo (HMC) is necessary for robust and efficient inference for such a complex model.

The `parameters` of the Stan model include the initial coordinates of the center (also the hands and feet if applicable), the initial width and height, the color, the displacement of the center (and the hands and feet) and the change in the width and height in each frame. The `transformed parameters` include the coordinates of the center (and the hands and feet), the width and height, and the color of all participants in all frames. The coordinates have the frame width and height as the upper bounds, whereas the displacement and the change in dimensions have small chosen lower and upper bounds. All initial coordinates have uniform interval priors. The log-likelihood in each iteration is the score that the sampled video depicts the input sentence. Empirically, sampling does not require a lot of iterations to converge—the setup that we use is 4 chains and 150 warm-up and sampling iterations. Videos can contain static and dynamic objects, static and dynamic relationships and properties, and concurrent actions. This setup also works for short multi-sentence stories extended in time.

After the sampling step, we combine each set of sampled coordinates, width, height and color of all participants into tracks, then we compute the scores for all the sampled videos and return the best one. These tracks can be rendered as videos using

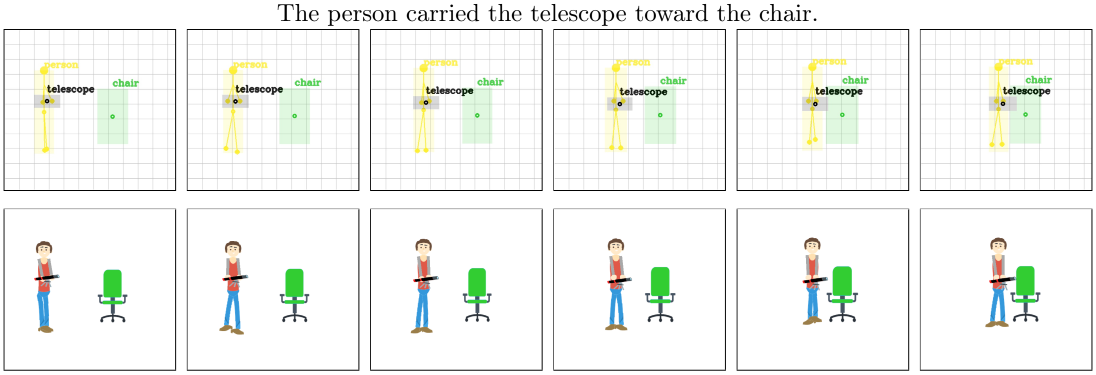

The person carried the telescope toward the chair.

Figure 4-3: An example video generated by the generative language–vision model. Each video has 12 frames in total but for readability only every other frame is shown here. This video is for the sentence "The person carried the telescope toward the chair". The movement of the legs and the shrinking distance between the person and the chair illustrate that the person is walking toward the chair.

OpenCV[1] and clipart[2] for qualitative evaluation purposes. Figure 4-3 illustrates an example sampled video rendered in two formats.

4.4 Video Completion

Using the ability to imagine scenes as described above, we can do video completion by synthesizing the missing frames conditioned on a sentence and prefix and/or suffix frames in a similar way as synthesizing a video from scratch. To do this, we extended the Stan model in Section 4.3 to take in prefix and suffix coordinates of the center (and the hands and feet), the width and height, and the color of all the participants, and only sample the parameters for the missing frames. We found that this approach also works for multi-sentence stories. Section 5.2 discusses the evaluation results.

4.5 Caption Generation

The scoring function enables us to compute the score or likelihood that a sentence describes a video. So given a video, we can generate a caption for it by systematically

[1] OpenCV library; https://opencv.org.

[2] https://www.kenney.nl/assets/modular-characters.

searching through the space of all possible sentences to get the one with the highest score. Those sentences can be generated using a context-free grammar and the lexicon. This was already done in the original Sentence Tracker papers [34, 39]. Here we focus on generating a caption for a video with missing frames by making use of the video completion capability. In other words, given an incomplete video, we seek to recover the video and generate a caption for it by finding the highest scoring video–sentence pair. Just as Yu et al. [39] pointed out, scores decrease with the number of words in the sentence. So we need to use beam search. We start with the top-scoring single-word sequences and then repeatedly expand the top-scoring sequences by one word and stop the search when the ratio between the score of the original sequence and the score of the expanded sequence falls below a *contradiction threshold*. The difference is that, for each candidate sentence, we need to sample the missing frames conditioned on it together with the prefix and/or suffix frames. For this reason, this task is expensive to perform but works quite reliably for simple sentences.

4.6 Paraphrases and Natural Language Inference

Our approach for solving NLI tasks using vision relies on the intuition that if two sentences are related in meaning, their scoring functions must also be related. In other words, with the closed-world assumption, if two sentences have identical meanings, the sets of videos that each sentence can be a caption of should be identical; on the other hand, if two sentences have completely opposite meanings, the sets of videos that each sentence can be a caption of should be disjoint. Consider the sentences below:

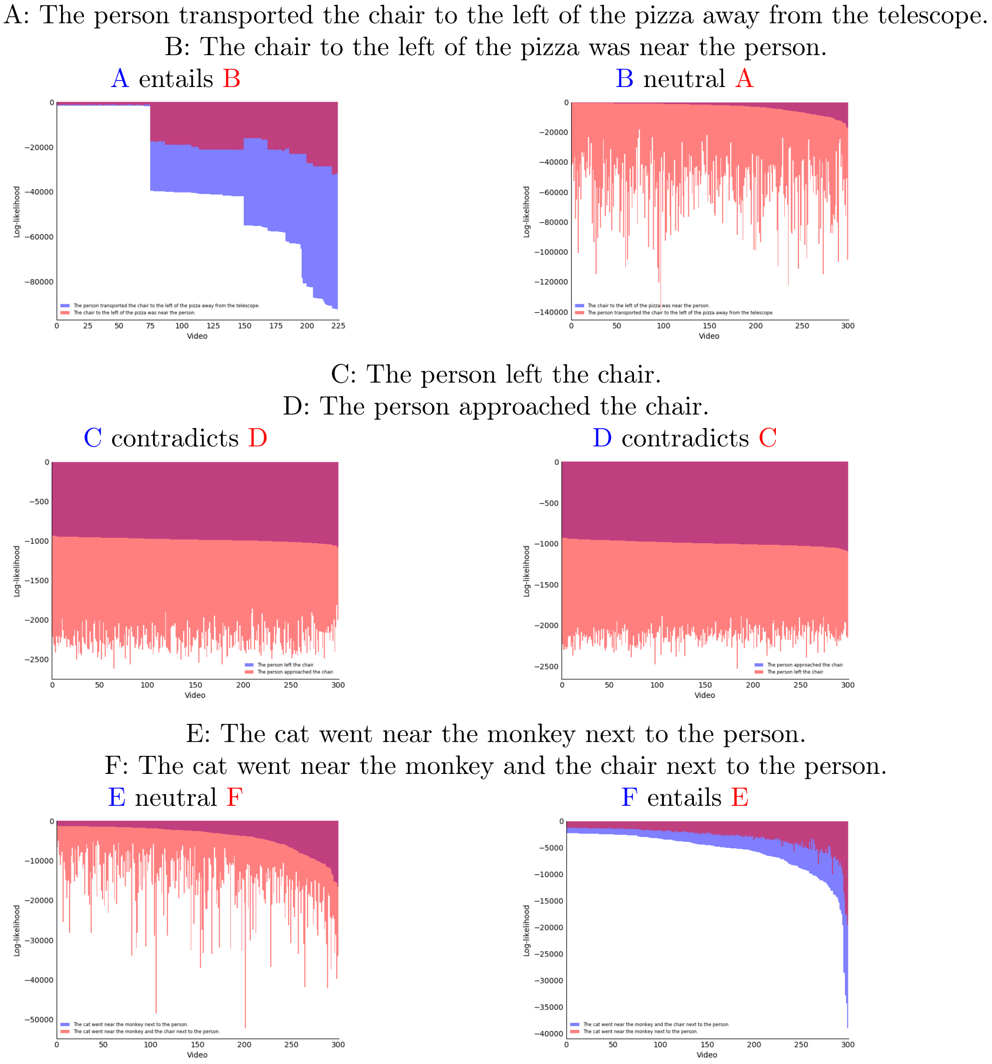
A. The person carried the bag toward the table.

B. The person approached the table.

C. The person left the table.

D. The person walked past a cat.

Taking A as the premise and the other as the hypotheses, we have that A entails B, A contradicts C and A and D are not related. The reasoning is as follows. Any imaginable scene for “The person carried the bag toward the table” can be captioned “The person approached the table with the bag” or just “The person approached the table”. However, it cannot be captioned “The person left the table with the bag” or “The person left the table”. It may though be captioned “The person walked past a cat.” if the scene indeed consists of a cat located somewhere between the person


A: The person transported the chair to the left of the pizza away from the telescope.
B: The chair to the left of the pizza was near the person.

A entails B

B neutral A

C: The person left the chair.
D: The person approached the chair.

C contradicts D

D contradicts C

E: The cat went near the monkey next to the person.
F: The cat went near the monkey and the chair next to the person.

E neutral F

F entails E

Log-likelihood
Video
0
−20000
−40000
−60000
−80000
0
25
50
75
100
125
150
175
200
225
The person transported the chair to the left of the pizza away from the telescope.
The chair to the left of the pizza was near the person.
0
−20000
−40000
−60000
−80000
−100000
−120000
−140000
0
50
100
150
200
250
300
The chair to the left of the pizza was near the person.
The person transported the chair to the left of the pizza away from the telescope.
0
−500
−1000
−1500
−2000
−2500
The person left the chair.
The person approached the chair.
The person approached the chair.
The person left the chair.
0
−10000
−20000
−30000
−40000
−50000
The cat went near the monkey next to the person.
The cat went near the monkey and the chair next to the person.
0
−5000
−10000
−15000
−20000
−25000
−30000
−35000
−40000
The cat went near the monkey and the chair next to the person.
The cat went near the monkey next to the person.

Figure 4-4: Three pairs of sentences and each of their six pairwise relationships. When finding the relationship between a pair of sentences, videos are sampled from the premise and then are scored on both the premise and hypothesis. The scores here are sorted in descending order by the premise score.

and the table. Therefore, in terms of likelihood, entailment (or paraphrase) implies that high-likelihood videos sampled from the premise will have high likelihood for the hypothesis, whereas contradiction implies that high-likelihood videos sampled from the premise will have low likelihood for the hypothesis. Neutral implies that there is no correlation between the likelihoods. So the relationship between two visually grounded sentences can be determined simply by comparing the likelihood histograms for the videos sampled from the premise as shown in Figure 4-4.

As such, to determine the relationship between a premise and hypothesis, we sample a large number of videos for the premise, then compute the likelihoods of those videos conditioned on the premise and on the hypothesis. This results in pairs of video likelihoods which we feed into a graphical model in Stan to determine how the sentences are related. It is a categorical mixture model of three components with the mixture proportions as the Stan `parameters`. Entailment is modeled as the score for the hypothesis being a positive linear transformation of the score for the premise with noise proportional to the score. Contradiction is modeled as the score for the hypothesis being a negative linear transformation of the score for the premise with noise proportional to the score. Neutral is modeled as the scores being independent. The predicted label is the one corresponds to the component with the highest mixture proportion. Section 6.2 discusses several limitations of this approach.

Chapter 5

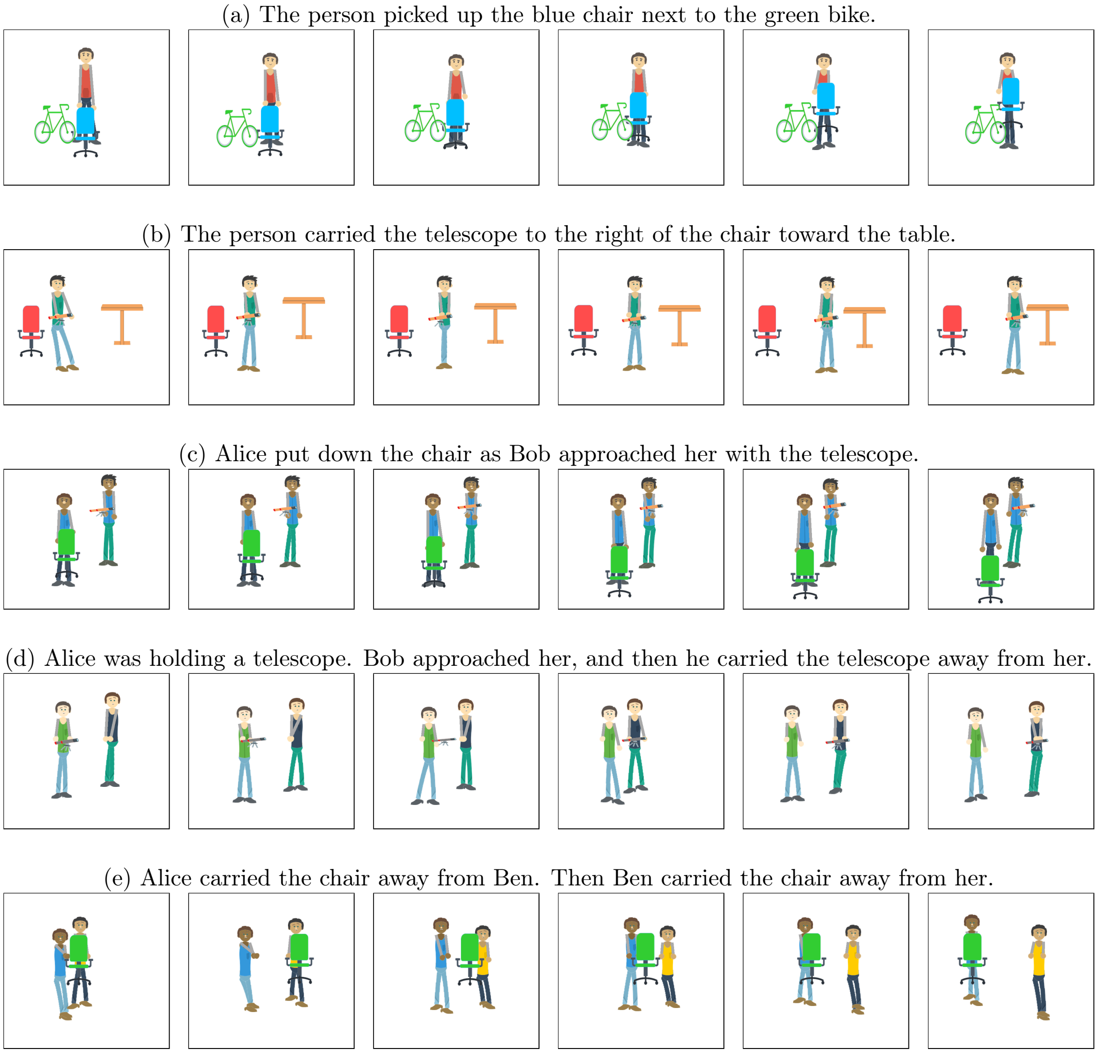
Evaluation

We evaluated our approach both qualitatively and quantitatively. For the vision tasks, we had to evaluate the generated videos and captions manually as there is no formal way to grade videos and captions. For the natural language inference task, we evaluated our approach using a custom evaluation corpus as briefly mentioned earlier. The sections below detail the evaluation procedure and results.

5.1 Video Generation

To evaluate the video generation capability, we generated videos for a set of manually written sentences and manually evaluated the videos one by one. Our evaluation sentences cover all words in the lexicon, and follow a grammar similar to the one in Yu et al. [39] but with additional rules for handling multiple clauses and sentences. The coordinate conjunctions we used include “and”, “as” and “while”, and the adverbs of sequence are “then” and “next”. This enabled us to evaluate videos with concurrent and complex actions extended in time. The results show that the model can reliably synthesize videos with static and dynamic objects involving concurrent and sequential actions, although complex videos require more sampling iterations to converge and more time to complete.

Figure 5-1 illustrates some generated videos. We can see that in cases where the event is underspecified as in (e), the model is able to extrapolate the action that should



(a) The person picked up the blue chair next to the green bike.

(b) The person carried the telescope to the right of the chair toward the table.

(c) Alice put down the chair as Bob approached her with the telescope.

(d) Alice was holding a telescope. Bob approached her, and then he carried the telescope away from her.

(e) Alice carried the chair away from Ben. Then Ben carried the chair away from her.

Figure 5-1: Videos generated by the generative language–vision model.

occur for the event to be realistic and coherent. That is, it knows that after Alice carried the chair away from Ben, she is at a distance from him; so in order for Ben to carry that chair back away from Alice, he needs to get close to her first although that action is not mentioned in the provided caption. In addition, we also found that the model can visualize multiple different scenarios for an event. For example, for the sentence "The person approached the chair", the videos we generate include both ones where the person approached the chair from the left and ones where the person approached the chair from the right. Similarly, for concurrent actions as shown in Figure 5-2, the model knows that "Alice left the chair as Ben approached it." could imply that Alice and Ben walked toward each other or walked in the same direction.

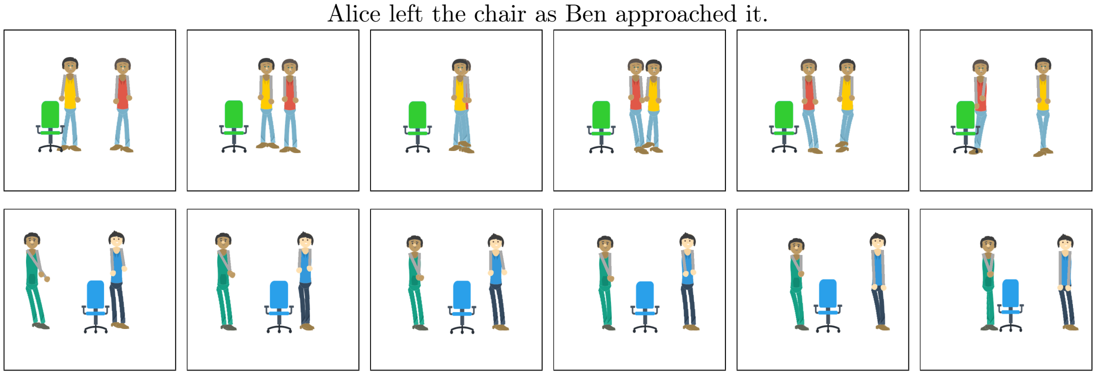


Figure 5-2: Two different generated videos for the exact same sentence.

This result shows that the model has realistic visual reasoning which enables it to understand language beyond just as a collection of linguistic forms. This is also reflected in the model's performance in the NLI task evaluated in Section 5.4 below.

5.2 Video Completion

We evaluated the video completion capability just as in the previous section except that we also gave the model several prefix and/or suffix frames which we got from separate video generation. We visualized the original and completed videos and compared them side by side. We found that the model is able to recover the videos just as expected. Figure 5-3 shows an example of a completed video. The model was given

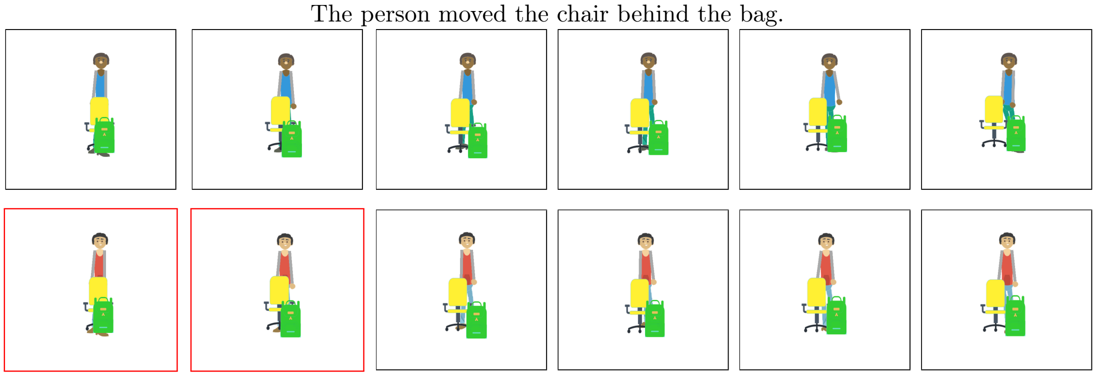


Figure 5-3: An example original and completed video pair. The top video is the original video, and the bottom video is completed video. The frames in red are the prefix frames given to the model. The look of the person is randomized every time so that difference here can be ignored.

the first 4 out of 12 frames (i.e., the 2 red frames in the figure) and was required to sample the rest of the video conditioned on the prefix frames and the sentence. As shown in the figure, the resampled video looks very similar to the original one—the chair behind the bag got moved to the left as intended. We saw similar results for more complicated sentences.

5.3 Caption Generation

This task is quite difficult to evaluate especially for complicated sentences where there are many ways to fill in the missing frames. Therefore, we evaluated it by looking at the top-scoring video–caption pairs. For simple videos involving only several (concurrent) actions such as "The person approached the chair" or "Alice picked up the chair as Ben put down the telescope", the model is able to complete the video and generate the correct caption given just the starting and final frame. We found that the top-scoring captions are generally just paraphrases of the original one. For example, for the video "The person carried the bag away from the chair", the top captions are "The person carried the bag to the right of the chair away from the chair", "The person left the chair with the bag" and "The person moved the bag to the right of the chair".

Videos with sequential actions take significantly more time to generate video–caption pairs for even after we enabled concurrent video sampling. We found that for sequential actions with very distinct start and end state such as "Alice put down the chair. Ben approached her, then he picked up the chair.", the model can recover the original video–caption pair quite reliably. However, for the less obvious actions, the model needs more context (i.e., prefix and suffix frames) and tends to produce captions that are a lot simpler than the intended one. This result is not necessarily negative because this setup is expected to be difficult even for humans. Instead, it is quite remarkable that the model can use visual imagination to perform this task without any additional training.

5.4 Paraphrases and Natural Language Inference

As described in Section 3.2, natural language inference is a popular task and there exist a number of benchmark corpora that could be used for evaluation. However, those corpora do not come with a lot of sentences that can be grounded in vision or that are in the grammatical structure that we target. Therefore, we compiled our own NLI evaluation corpus consisting of sentence pairs that involve visual reasoning and are challenging for existing NLI approaches. For comparison, we selected 4 different state-of-the-art models trained on the SNLI corpus as our baselines, namely BERT [12], Parikh + ELMO [25], ESIM + ELMO [10], and Infersent [11]. The subsections below describe the corpus itself, the other baselines used and the evaluation results.

5.4.1 Grounded NLI Corpus

While our approach can parse and compare general sentences grounded in vision, we focus on more systematic cases that highlight the performance limitations of standard NLI approaches. We generated an NLI corpus that can complement existing corpora and can be used for future NLI research. Following [24], sentence pairs were generated from templates with varied syntactic structures that produce near-misses, require visual reasoning and avoid biases present in standard datasets. Our approach never sees an NLI dataset so it cannot learn the arbitrary correlations that often allow for high performance without any understanding of the task. The resulting dataset consists of 2,100 sentence pairs and is challenging compared to existing NLI datasets, with performance of state-of-the-art models being 10–20% lower than on SNLI and MultiNLI.

To generate premise–hypothesis pairs, we make use of templates that can be classified into nine different categories as shown in Table 5.1. We made sure to have at least two labels within each category and include equal number of sentence pairs (100 pairs) for each label in each category to avoid biases. The detailed templates can be found in Table A.1. These templates require verbs that are symmetrical (e.g., "The

Category	Description
Passive Voice	The premise is in active voice and the hypothesis is in passive voice, and the pair only differs in either the subject/object order or negation.
Verb Argument Order	The premise and hypothesis are both in active voice and only differ in the subject/object order.
Prepositional Phrase Argument Order	The premise and hypothesis only differ in the argument order of one spatial-relation prepositional phrase or motion prepositional phrase.
Related Verbs	The premise and hypothesis only differ in one verb chosen to be either synonyms or words that are related but are subtly different.
Related Verbs (Different Structures)	The difference is similar to the above but the verbs do not have to be both transitive or intransitive.
Related Prepositions	The premise and hypothesis only differ in one preposition chosen to be either antonyms or words that imply one another visually.
Indirect Implications	The premise and hypothesis are completely different in structure and word usage.
Conjunction (Subject and Object)	The premise and/or hypothesis have subjects or objects with conjunction and have either synonymous or opposite verbs.
Conjunction (Modifiers)	The premise and hypothesis have objects with modifiers and conjunction, and differ in at least one modifier.

Table 5.1: The descriptions for the templates used to generate the dataset.

person shook hands with the dog" and "The dog shook hands with the person") and verbs that are related but different (e.g., "feed" and "pet"). Thus, we introduced a number of additional verbs that also serve to diversify the sentences in the corpus: *rotate*, *throw*, *push*, *kick*, *punch*, *wash*, *feed*, *pet*, *meet*, *shake hands with*, *reunite with*, *confront with* and *jump*. Our approach handles these unknown words by using GloVe embeddings to map them to the words in the lexicon. While this is a promising solution, we expected this addition to lead to a notable decrease in the performance of our approach.

5.4.2 Baselines

In addition to the state-of-the-art models above, we used two other baselines referred to as Baseline 1 and Baseline 2. These baselines are ablations of our approach. They differ from the original approach in that we ablated the key aspects of action recognition from it. Specifically, we ablated changes in time by ignoring verbs in Baseline 1 and ablated both spatial relations and changes in time by ignoring both prepositions and verbs in Baseline 2. In other words, Baseline 1 can only imagine static scenes since it cannot reason about the dynamic relationships between participants, whereas Baseline 2 can only imagine static scenes where both dynamic and static relationships between participants are completely unconstrained. We expected these baselines to do significantly worse than the complete approach as one must reason about the properties of objects, relationships between objects, and how these change over time in order to perform inference with vision.

5.4.3 Results

Table 5.2 shows the accuracy of all the models on the SNLI corpus and the grounded corpus described in 5.4.1. Overall, our approach performed better than all the baselines despite having not seen any NLI examples. BERT has the next highest performance with 73.2% accuracy, whereas Infersent has the worst performance with 52.6% accuracy. Ablations of our model also perform quite poorly; removing reasoning over time reduces performance to 63.5% and removing that and reasoning about spatial relations reduces performance to 54.0%. As expected, Baseline 2 has chance performance on all categories except for the modifiers one. The accuracies by label show that Baseline 2 either always predicts one label or predicts all the labels uniformly. Similarly, Baseline 1 has near chance performance on all categories except for the modifiers and prepositions ones.

The state-of-the-art models share a set of systematic limitations. For instance, they all do poorly on passive voice. The accuracies by label show that the models predominantly predict *entailment* for this category likely due to the large lexical overlap and the matching noun order such as in “The person kicked the ball” and “The person was kicked by the ball”. BERT and ESIM appear to do well on related verbs and related prepositions, while the others tend to misclassify *neutral* pairs such as “The person went near the monkey” and “The person carried the dog toward the monkey” as *contradiction*. BERT appears to be the best at capturing the changes in

	BERT	Parikh	ESIM	Infersent	Ours	Baseline 1	Baseline 2
SNLI	**90.2**	86.4	88.5	84.6	N/A	N/A	N/A
Passive Voice	52.5	50.5	48.5	50.0	**73.5**	56.0	53.0
Verb Argument Order	**71.5**	58.0	60.5	50.0	61.5	56.5	50.0
Prepositional Phrase Argument Order	60.0	54.5	57.0	50.0	**82.5**	68.5	50.0
Related Verbs	96.0	65.0	88.5	53.5	**97.0**	54.0	50.0
Related Prepositions	81.0	77.5	92.0	53.5	**97.5**	94.5	50.0
Related Verbs (Different Structures)	**66.3**	62.3	59.0	51.7	57.3	46.0	49.7
Indirect Implications	**77.3**	67.3	75.0	51.2	70.0	42.3	33.3
Conjunction (Subject and Object)	**76.3**	58.0	77.0	62.0	74.0	56.3	50.0
Conjunction (Modifiers)	78.0	53.5	65.0	52.5	99.5	97.8	**100.0**
Overall μ	73.2	60.7	69.2	52.6	**79.2**	63.5	54.0
Overall σ	12.0	7.8	14.0	3.7	15.0	18.8	17.1

Table 5.2: The accuracy of models on the SNLI test set and our grounded NLI corpus broken down by category. Note that the SNLI accuracies are not included in the overall mean and standard deviation calculation.

argument order in the remaining categories.

Our approach has comparable performance to BERT in all categories although the dataset contains many words unknown to our model. Yet, according to the by-label accuracies, our approach also has a tendency to misclassify *neutral* pairs as *contradiction*. One plausible explanation is that although sentence like “The person walked away from the dog” does not imply “The person carried the chair away from the dog”, the sentences are still related visually so the likelihoods of those sentences on the same videos are related. Despite this, our model and even its ablations still outperform the state-of-the-art models in several categories, namely modifiers and passive voice. This is because while the syntactic difference might be subtle and difficult to capture, the difference in the visualized scenes is very significant. For instance, visually “The person picked up the blue chair next to the black table” is very different from “The person picked up the black chair next to the blue table”. Likewise, “The person approached the dog” and “The person was approached by the dog” are also very different. So in some cases solving language tasks using vision might be significantly less difficult than solving using linguistics. Overall, the results here demonstrate that it is possible to convert a language task to a vision task and solve it without additional training. This models a capacity to generalize that is central to human intelligence.

Chapter 6

Discussion

6.1 Summary

We introduced a generative language–vision model that can generate videos conditioned on sentences. We showed that this capability enabled us to do video completion, video captioning and natural language inference without task-specific training. The only training required is for making the model acquire a lexicon from captioned videos similar to the way children acquire language through exposure to perceptual cues. To do video generation, we made use of the Sentence Tracker's video–sentence scoring approach to build a generative model that can synthesize videos for a sentence by sampling the visual features of the participants of the described event. We evaluated the tasks above both quantitatively and qualitatively. We found that our approach can reliably generate videos with static and dynamic objects involving in concurrent and sequential actions. The promising results on NLI show that it is possible to reduce a language task to a vision task and take advantage of information obtained through vision to solve the task more robustly.

6.2 Challenges

Despite the positive results, our approach has a number of limitations. For instance, while the No-U-Turn sampler (NUTS) works well for our purpose, the whole video generation step is very expensive timewise, especially for more complicated sentences. As such, when evaluating the approach on the thousands of sentence pairs in the NLI corpus, we could only afford to sample the coordinates of the center of the participants and completely omitted the coordinates of hands and feet. This is likely to have led to a decrease in performance.

In addition, we should have formalized the Stan model for classifying the entailment relationship further. In particular, we did not manage to formally account for the difference in the number of words in the premise and hypothesis. As discussed earlier, scores decrease with the sentence length. So to tackle that, we placed priors on the variance of the key features in each word model during training to force the learned parameters to have tight variance. This is to ensure that a false video always gets significantly lower scores than a true video regardless of the length of the sentence. Given this, we distinguished entailment from contradiction by simply setting a threshold for the difference in scores to be proportional to the premise score. We did not tune this threshold; however, we believe that a more formal solution is needed for more accurate and precise evaluation. We experimented with inserting placeholder words with flat distributions for all features but found that it led to score distributions that are even harder to reason about.

Another challenge faced by this approach is the lack of a direct way to obtain the linking function. In this work, we parse the Stanford Dependencies representation and handle a subset of 50 grammatical relations using manually-constructed rules. This is only feasible because we only target grammatically simple sentences. However, this step will not scale to a more flexible set of natural language sentences.

6.3 Future Work

Our approach has several potential extensions. So far, we see that by simply crafting features that capture the physical implications of word meanings, we are able to generate realistic visualizations of actions described in a sentence. We intend to move a step further by incorporating physics into the model to predict consequences of actions and do commonsense question answering. For example, given “The person released the ball”, using physics we should be able to predict that the ball will fall toward the ground due to gravity and is likely to bounce multiple times. Likewise, given “Two people ran toward each other”, we should be able to predict that they will meet and maybe run into each other somewhere between the points they were at in the beginning. If precise calculation is used, we should also be able to track the positions of participants over time which might be useful for other applications.

We also intend to do paraphrase recognition between two languages (i.e., translation) similar to the way the NLI task is solved above. To do this, we will need to train two models on two different captioned video corpora in the languages of interest. With that, we can then check if two sentences in different languages are equivalent by generating videos for one of the sentences and comparing the likelihoods of those videos conditioned on each of the sentences just as before. Since those sentences can be grounded in vision, their scoring functions should agree with each other and we should be able to tell whether they have similar meanings. We believe that the likelihood pairs computed here should also be useful for loss calculation when training translation generation models.

Finally, we intend to use this approach to help robots learn to follow simple commands. In particular, we should be able to take in a command and a video recording of the robot so far, and run the Viterbi algorithm on the parsed command and extracted tracks to predict the most likely state sequence representing the robot’s action up until that point and return the likelihood that the action has been completed. We also plan to explore using the video completion and captioning capability to help robots make plans. That is, given the start and target states, we should in principle be able to synthesize possible videos of the robot doing the action and generate a corresponding plan for that.

6.4 Contributions

This thesis introduced a visual approach to language understanding that can reliably solve a set of language–vision tasks without task-specific training. We presented a probabilistic model that can robustly synthesize videos for sentences describing concurrent and sequential actions, and therefore outperforms standard video generation approaches. We demonstrated how to reduce language tasks such as paraphrase recognition and NLI to vision tasks, and solve the tasks just as effectively as approaches trained on thousands of examples. In doing that, we created an NLI evaluation corpus that is difficult for standard approaches and can be used with large benchmark corpora. Overall, this work laid the foundation for applications of vision and perception to natural language processing, and presented a promising approach to generalizing across language and vision tasks.

Appendix A

Tables

Table A.1: The templates used to generate sentences in our corpus. Here P stands for spatial-relation preposition, whereas P_M stands for motion preposition. For simplicity, we hide the optional PP in the NP in the templates above, and only expand the NP when we need to show the difference between sentences.

Category	Template	Example
Passive Voice	Premise: NP_1 VBD [P] NP_2	The person moved the table behind the bike.
	Entailment: NP_2 was VBN by NP_1	The table behind the bike was moved by the person.
	Contradiction: NP_1 was VBN by NP_2	The person was moved by the table behind the bike.
	Premise: NP_1 VBD [P] NP_2	The person fought with the cat to the right of the chair.
	Entailment: NP_1 was VBN by NP_2	The person was fought by the cat to the right of the chair.
	Contradiction: NP_2 was not VBN by NP_1	The cat to the right of the chair was not fought by the person.
Verb Argument Order	Premise: NP_1 VBD [P] NP_2	The person met the dog.
	Entailment: NP_2 VBD [P] NP_1	The dog met the person.
	Premise: NP_1 VBD [P] NP_2	The person kicked the table in front of the bag.
	Contradiction: NP_2 VBD [P] NP_1	The table in front of the bag kicked the person.
Prepositional Phrase Argument Order	Premise: NP_1 VBD [P] NP_2 P_1 NP_3	The monkey approached the table next to the bike.
	Entailment: NP_1 VBD [P] NP_3 P_1 NP_2	The monkey approached the bike next to the table.

	Premise: NP$_1$ VBD [P] NP$_2$ P$_1$ NP$_3$	The person punched the bag to the left of the chair.
	Contradiction: NP$_1$ VBD [P] NP$_3$ P$_1$ NP$_2$	The person punched the chair to the left of the bag.
	Premise: NP$_1$ VBD [P] NP$_2$ P$_{\text{M}}$ NP$_3$	The person carried the bag toward the cat.
	Contradiction: NP$_1$ VBD [P] NP$_3$ P$_{\text{M}}$ NP$_2$	The person carried the cat toward the bag.
	Premise: NP$_1$ VBD [P] NP$_2$ and NP$_3$ P$_1$ NP$_4$	The dog left the car and the telescope next to the bag.
	Entailment: NP$_1$ VBD [P] NP$_2$ and NP$_4$ P$_1$ NP$_3$	The dog left the car and the bag next to the telescope.
	Premise: NP$_1$ VBD [P] NP$_2$ and NP$_3$ P$_1$ NP$_4$	The dog passed by the cat and the pizza in front of the car.
	Contradiction: NP$_1$ VBD [P] NP$_2$ and NP$_4$ P$_1$ NP$_3$	The dog passed by the cat and the car in front of the pizza.
	Premise: NP$_1$ VBD [P] NP$_2$ and NP$_3$ P$_{\text{M}}$ NP$_4$	The person transported the bike and the pizza toward the cat.
	Contradiction: NP$_1$ VBD [P] NP$_2$ and NP$_4$ P$_{\text{M}}$ NP$_3$	The person transported the bike and the cat toward the pizza.
Related Verbs	Premise: NP$_1$ VBD$_1$ [P] NP$_2$	The monkey stared at the pizza.
	Entailment: NP$_1$ VBD$_2$ [P] NP$_2$	The monkey looked at the pizza.
	Premise: NP$_1$ VBD$_1$ [P] NP$_2$	The person fed the dog near the bag.
	Contradiction: NP$_1$ VBD$_2$ [P] NP$_2$	The person petted the dog near the bag.
Related Prepositions	Premise: NP$_1$ VBD [P] NP$_2$ P$_1$ NP$_3$	The person moved the pizza in front of the bag.
	Entailment: NP$_1$ VBD [P] NP$_2$ P$_2$ NP$_3$	The person moved the pizza next to the bag.
	Premise: NP$_1$ VBD [P] NP$_2$ P$_1$ NP$_3$	The person picked up the bag to the left of the monkey.
	Contradiction: NP$_1$ VBD [P] NP$_2$ P$_2$ NP$_3$	The person picked up the bag to the right of the monkey.
Related Verbs (Different Structures)	Premise: NP$_1$ VBD$_1$ [P] NP$_2$ P$_{\text{M}}$ NP$_3$	The person carried the dog toward the monkey.
	Entailment: NP$_1$ VBD$_2$ [P] NP$_3$ with NP$_2$	The person approached the monkey with the dog.
	Contradiction: NP$_1$ VBD$_3$ [P] NP$_3$ without NP$_2$	The person walked toward the monkey without the dog.
	Premise: NP$_1$ VBD$_1$ [P] NP$_2$	The person went near the monkey.
	Neutral: NP$_1$ VBD$_2$ [P] NP$_2$ P$_{\text{M}}$ NP$_3$	The person carried the dog toward the monkey.

Indirect Implications	Premise: NP_1 VBD [P] NP_2 P_1 NP_3	The person picked up the chair to the left of the cat.
	Entailment: NP_2 was P_2 NP_1	The chair was near the person.
	Contradiction: NP_2 was P_3 NP_1	The chair was far away from the person.
	Premise: NP_1 VBD [P] NP_2	The monkey looked at the pizza.
	Neutral: NP_2 was P_1 NP_1	The pizza was far away from the monkey.
Conjunction (Subject and Object)	Premise: NP_1 VBD [P] NP_2 and NP_3	The monkey walked past the car and the pizza near the bag.
	Entailment: NP_1 VBD [P] NP_2	The monkey walked past the car near the bag.
	Entailment: NP_1 VBD [P] NP_3	The monkey walked past the pizza near the bag.
	Premise: NP_1 VBD_1 [P] NP_2	The monkey picked up the pizza on the table.
	Contradiction: NP_1 VBD_2 [P] NP_2 and NP_3	The monkey put down the telescope and the pizza on the table.
	Premise: NP_1 VBD [P] NP_2	The person raised the bike in front of the bag.
	Neutral: NP_1 VBD [P] NP_2 and NP_3	The person raised the bike and the cat in front of the bag.
Conjunction (Modifiers)	Premise: NP_1 VBD [P] JJ_2 N_2 and JJ_3 N_3	The dog passed by the yellow table and the black monkey.
	Entailment: NP_1 VBD [P] JJ_3 N_3 and JJ_2 N_2	The dog passed by the black monkey and the yellow table.
	Entailment: NP_1 VBD [P] JJ_2 N_2	The dog passed by the the yellow table.
	Entailment: NP_1 VBD [P] JJ_3 N_3	The dog passed by the black monkey.
	Contradiction: NP_1 VBD [P] JJ_3 N_2 and JJ_2 N_3	The dog passed by the black table and the yellow monkey.
	Contradiction: NP_1 VBD [P] JJ_3 N_2	The dog passed by the black table.
	Contradiction: NP_1 VBD [P] JJ_2 N_3	The dog passed by the yellow table.

Bibliography

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970.

[3] Yevgeni Berzak, Andrei Barbu, Daniel Harari, Boris Katz, and Shimon Ullman. Do you see what I mean? Visual resolution of linguistic ambiguities. *EMNLP*, 2016.

[4] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[5] Haoye Cai, Chunyan Bai, Yu-Wing Tai, and Chi-Keung Tang. Deep video generation, prediction and completion of human action sequences. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 366–382, 2018.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[7] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[8] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.

[9] Danqi Chen and Christopher Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750, 2014.

[10] Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced LSTM for natural language inference. *ACL*, 2016.

[11] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics, 2017.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*, 2018.

[14] Zoubin Ghahramani and Michael I Jordan. Factorial Hidden Markov Models. In *Advances in Neural Information Processing Systems*, pages 472–478, 1996.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[16] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*, 2018.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[18] Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[19] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[20] Yitong Li, Martin Renqiang Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[21] Xiao Lin and Devi Parikh. Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2984–2993, 2015.

[22] Bill MacCartney and Christopher D Manning. *Natural language inference*. Citeseer, 2009.

[23] Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 1–8, 2014.

[24] R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*, 2019.

[25] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *EMNLP*, 2016.

[26] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[27] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL*, 2018.

[28] Steven Pinker. Formal models of language learning. *Cognition*, 7(3):217–283, 1979.

[29] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, 2018.

[30] Lawrence R Rabiner and Biing-Hwang Juang. An introduction to Hidden Markov Models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.

[31] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[32] Candace Ross, Andrei Barbu, Yevgeni Berzak, Battushig Myanganbayar, and Boris Katz. Grounding language acquisition by training semantic parsers using captioned videos. In *Conference on empirical methods in natural language processing (EMNLP)*, pages 740–750, 2018.

[33] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. LabelMe: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008.

[34] Narayanaswamy Siddharth, Andrei Barbu, and Jeffrey Mark Siskind. Seeing what you're told: Sentence-guided activity recognition in video. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2014.

[35] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621, 2016.

[36] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *ACL*, 2017.

[37] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[38] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[39] Haonan Yu, N Siddharth, Andrei Barbu, and Jeffrey Mark Siskind. A compositional framework for grounding language inference, generation, and acquisition in video. *Journal of Artificial Intelligence Research*, 52:601–713, 2015.

[40] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*, 2018.