SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference

Rowan Zellers♠   **Yonatan Bisk**♠   **Roy Schwartz**♠♡   **Yejin Choi**♠♡

♠Paul G. Allen School of Computer Science & Engineering, University of Washington
♡Allen Institute for Artificial Intelligence
{rowanz,ybisk,roysch,yejin}@cs.washington.edu
https://rowanzellers.com/swag

Abstract

Given a partial description like "she opened the hood of the car," humans can reason about the situation and anticipate what might come next ("then, she examined the engine"). In this paper, we introduce the task of grounded commonsense inference, unifying natural language inference and commonsense reasoning.

We present SWAG, a new dataset with 113k multiple choice questions about a rich spectrum of grounded situations. To address the recurring challenges of the annotation artifacts and human biases found in many existing datasets, we propose *Adversarial Filtering* (AF), a novel procedure that constructs a de-biased dataset by iteratively training an ensemble of stylistic classifiers, and using them to filter the data. To account for the aggressive adversarial filtering, we use state-of-the-art language models to massively oversample a diverse set of potential counterfactuals. Empirical results demonstrate that while humans can solve the resulting inference problems with high accuracy (88%), various competitive models struggle on our task. We provide comprehensive analysis that indicates significant opportunities for future research.

1 Introduction

When we read a story, we bring to it a large body of implicit knowledge about the physical world. For instance, given the context "on stage, a woman takes a seat at the piano," shown in Table 1, we can easily infer what the situation might *look* like: a woman is giving a piano performance, with a crowd watching her. We can furthermore infer her likely *next* action: she will most likely set her fingers on the piano keys and start playing.

This type of natural language inference requires commonsense reasoning, substantially broadening the scope of prior work that focused primarily on

---

On stage, a woman takes a seat at the piano. She
a) sits on a bench as her sister plays with the doll.
b) smiles with someone as the music plays.
c) is in the crowd, watching the dancers.
d) nervously sets her fingers on the keys.

A girl is going across a set of monkey bars. She
a) jumps up across the monkey bars.
b) struggles onto the monkey bars to grab her head.
c) gets to the end and stands on a wooden plank.
d) jumps up and does a back flip.

The woman is now blow drying the dog. The dog
a) is placed in the kennel next to a woman's feet.
b) washes her face with the shampoo.
c) walks into frame and walks towards the dog.
d) tried to cut her face, so she is trying to do something very close to her face.

Table 1: Examples from SWAG; the correct answer is **bolded**. Adversarial Filtering ensures that stylistic models find all options equally appealing.

linguistic entailment (Chierchia and McConnell-Ginet, 2000). Whereas the dominant entailment paradigm asks if two natural language sentences (the 'premise' and the 'hypothesis') describe the same set of possible worlds (Dagan et al., 2006; Bowman et al., 2015), here we focus on whether a (multiple-choice) ending describes a possible (*future*) world that can be anticipated from the situation described in the premise, even when it is not strictly entailed. Making such inference necessitates a rich understanding about everyday physical situations, including object affordances (Gibson, 1979) and frame semantics (Baker et al., 1998).

A first step toward grounded commonsense inference with today's deep learning machinery is to create a large-scale dataset. However, recent work has shown that human-written datasets are susceptible to *annotation artifacts*: unintended stylistic patterns that give out clues for the gold labels (Gururangan et al., 2018; Poliak et al., 2018). As a result, models trained on such datasets with hu-

man biases run the risk of over-estimating the actual performance on the underlying task, and are vulnerable to adversarial or out-of-domain examples (Wang et al., 2018; Glockner et al., 2018).

In this paper, we introduce Adversarial Filtering (AF), a new method to automatically detect and reduce stylistic artifacts. We use this method to construct SWAG: an adversarial dataset with 113k multiple-choice questions. We start with pairs of temporally adjacent video captions, each with a context and a follow-up event that we *know* is physically possible. We then use a state-of-the-art language model fine-tuned on this data to massively oversample a diverse set of possible negative sentence endings (or *counterfactuals*). Next, we filter these candidate endings aggressively and adversarially using a committee of trained models to obtain a population of de-biased endings with similar stylistic features to the real ones. Finally, these filtered counterfactuals are validated by crowd workers to further ensure data quality.

Extensive empirical results demonstrate unique contributions of our dataset, complementing existing datasets for natural langauge inference (NLI) (Bowman et al., 2015; Williams et al., 2018) and commonsense reasoning (Roemmele et al., 2011; Mostafazadeh et al., 2016; Zhang et al., 2017). **First**, our dataset poses a new challenge of grounded commonsense inference that is easy for humans (88%) while hard for current state-of-the-art NLI models (<60%). **Second**, our proposed adversarial filtering methodology allows for cost-effective construction of a large-scale dataset while substantially reducing known annotation artifacts. The generality of adversarial filtering allows it to be applied to build future datasets, ensuring that they serve as reliable benchmarks.

2   SWAG: Our new dataset

We introduce a new dataset for studying physically grounded commonsense inference, called SWAG.[1] Our task is to predict which event is most likely to occur next in a video. More formally, a model is given a context $c = (s, n)$: a complete sentence s and a noun phrase n that begins a second sentence, as well as a list of possible verb phrase sentence endings $V = \{v_1, \ldots, v_4\}$. See Figure 1 for an example triple (s, n, v_i). The model must then select the most appropriate verb phrase $v_i \in V$.
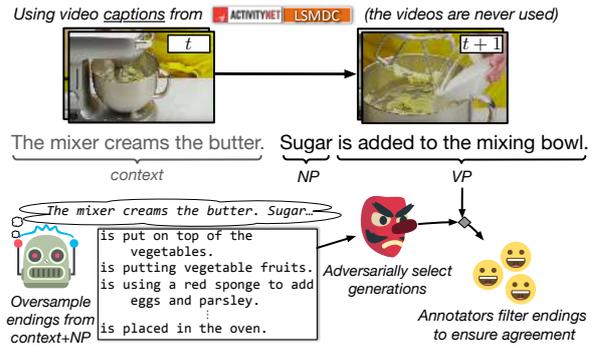

Figure 1: Overview of the data collection process. For a pair of sequential video captions, the second caption is split into noun and verb phrases. A language model generates many negative endings, of which a difficult subset are human-annotated.

Overview   Our corpus consists of 113k multiple choice questions (73k training, 20k validation, 20k test) and is derived from pairs of consecutive video captions from ActivityNet Captions (Krishna et al., 2017; Heilbron et al., 2015) and the Large Scale Movie Description Challenge (LSMDC; Rohrbach et al., 2017). The two datasets are slightly different in nature and allow us to achieve broader coverage: ActivityNet contains 20k YouTube clips containing one of 203 activity types (such as doing gymnastics or playing guitar); LSMDC consists of 128k movie captions (audio descriptions and scripts). For each pair of captions, we use a constituency parser (Stern et al., 2017) to split the second sentence into noun and verb phrases (Figure 1).[2] Each question has a human-verified gold ending and 3 distractors.

3   A solution to annotation artifacts

In this section, we outline the construction of SWAG. We seek dataset diversity while minimizing *annotation artifacts*, conditional stylistic patterns such as length and word-preference biases. For many NLI datasets, these biases have been shown to allow shallow models (e.g. bag-of-words) obtain artificially high performance.

To avoid introducing easily "gamed" patterns, we present Adversarial Filtering (AF), a generally-applicable treatment involving the iterative refinement of a set of assignments to increase the entropy under a chosen model family. We then discuss how we generate counterfactual endings, and

---

Algorithm 1 Adversarial filtering (AF) of negative samples. During our experiments, we set $N^{easy} = 2$ for refining a population of $N^- = 1023$ negative examples to $k = 9$, and used a 80%/20% train/test split.

---

while convergence not reached **do**
- Split the dataset $\mathcal{D}$ randomly up into training and testing portions $\mathcal{D}^{tr}$ and $\mathcal{D}^{te}$.
- Optimize a model f_θ on $\mathcal{D}^{tr}$.

for index i in $\mathcal{D}^{te}$ **do**
- Identify easy indices:
$\mathcal{A}_i^{easy} = \{j \in \mathcal{A}_i : f_\theta(x_i^+) > f_\theta(x_{i,j}^-)\}$
- Replace N^{easy} easy indices $j \in \mathcal{A}_i^{easy}$ with adversarial indices $k \notin \mathcal{A}_i$ satisfying $f_\theta(x_{i,k}^-) > f_\theta(x_{i,j}^-)$.

end for
end while

---

finally, the models used for filtering.

3.1 Formal definition

In this section, we formalize what it means for a dataset to be *adversarial*. Intuitively, we say that an adversarial dataset for a model f is one on which f will not generalize, even if evaluated on test data from the same distribution. More formally, let our input space be $\mathcal{X}$ and the label space be $\mathcal{Y}$. Our trainable classifier f, taking parameters θ is defined as $f_\theta : \mathcal{X} \rightarrow \mathbb{R}^{|\mathcal{Y}|}$. Let our dataset of size N be defined as $\mathcal{D} = \{(x_i, y_i)\}_{1 \leq i \leq N}$, and let the loss function over the dataset be $L(f_\theta, \mathcal{D})$. We say that a dataset is *adversarial* with respect to f if we expect high empirical error I over all leave-one-out train/test splits (Vapnik, 2000):

$$I(\mathcal{D}, f) = \frac{1}{N} \sum_{i=1}^{N} L(f_{\theta_i^\star}, \{(x_i, y_i)\}), \qquad (1)$$

where $\theta_i^\star = \underset{\theta}{\operatorname{argmin}}\, L(f_\theta, \mathcal{D} \setminus \{(x_i, y_i)\})$, (2)

with regularization terms omitted for simplicity.

3.2 Adversarial filtering (AF) algorithm

In this section, we outline an approach for generating an adversarial dataset $\mathcal{D}$, effectively maximizing empirical error I with respect to a family of trainable classifiers f. Without loss of generality, we consider the situation where we have N *contexts*, each associated with a single positive example $(x_i^+, 1) \in \mathcal{X} \times \mathcal{Y}$, and a large population of context-specific negative examples $(x_{i,j}^-, 0) \in \mathcal{X} \times \mathcal{Y}$, where $1 \leq j \leq N^-$ for each i. For instance, the negative examples could be incorrect relations in knowledge-base completion (Socher et al., 2013), or all words in a dictionary for a single-word cloze task (Zweig and Burges, 2011).

Our goal will be to filter the population of negative examples for each instance i to a size of $k \ll N^-$. This will be captured by returning a set of *assignments* $\mathcal{A}$, where for each instance the assignment will be a k-subset $\mathcal{A}_i = [1 \ldots N^-]^k$. The filtered dataset will then be:

$$\mathcal{D}^{AF} = \{(x_i, 1), \{(x_{i,j}^-, 0)\}_{j \in \mathcal{A}_i}\}_{1 \leq i \leq N} \quad (3)$$

Unfortunately, optimizing $I(\mathcal{D}^{AF}, f)$ is difficult as $\mathcal{A}$ is global and non-differentiable. To address this, we present Algorithm 1. On each iteration, we split the data into dummy 'train' and 'test' splits. We train a model f on the training portion and obtain parameters θ, then use the remaining test portion to reassign the indices of $\mathcal{A}$. For each context, we replace some number of 'easy' negatives in $\mathcal{A}$ that f_θ classifies correctly with 'adversarial' negatives outside of $\mathcal{A}$ that f_θ misclassifies.

This process can be thought of as increasing the overall entropy of the dataset: given a strong model f_θ that is compatible with a random subset of the data, we aim to ensure it cannot generalize to the held-out set. We repeat this for several iterations to reduce the generalization ability of the model family f over arbitrary train/test splits.

3.3 Generating candidate endings

To generate counterfactuals for **SWAG**, we use an LSTM (Hochreiter and Schmidhuber, 1997) language model (LM), conditioned on contexts from video captions. We first pretrain on BookCorpus (Zhu et al., 2015), then finetune on the video caption datasets. The architecture uses standard best practices and was validated on held-out perplexity of the video caption datasets; details are in the appendix. We use the LM to sample $N^- =1023$ unique endings for a partial caption.[3]

Importantly, we *greedily* sample the endings, since beam search decoding biases the generated endings to be of lower perplexity (and thus easily distinguishable from found endings). We find this process gives good counterfactuals: the generated endings tend to use *topical* words, but often make little sense physically, making them perfect for our task. Further, the generated endings are marked as "gibberish" by humans only 9.1% of the time (Sec 3.5); in that case the ending is filtered out.

---

[3]To ensure that the LM generates unique endings, we split the data into five validation folds and train five separate LMs, one for each set of training folds. This means that each LM never sees the found endings during training.
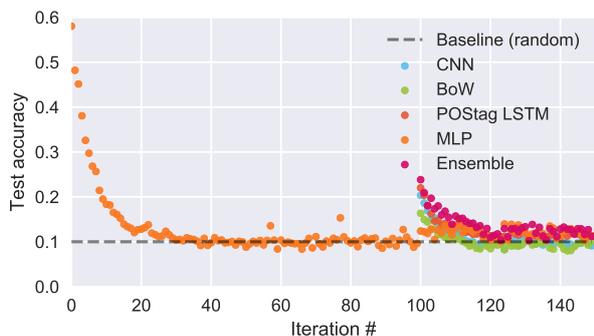
Figure 2: Test accuracy by AF iteration, under the negatives given by $\mathcal{A}$. The accuracy drops from around 60% to close to random chance. For efficiency, the first 100 iterations only use the MLP.

3.4 Stylistic models for adversarial filtering

In creating **Swag**, we designed the model family f to pick up on low-level *stylistic features* that we posit should not be predictive of whether an event happens next in a video. These stylistic features are an obvious case of annotation artifacts (Cai et al., 2017; Schwartz et al., 2017).[4] Our final classifier is an ensemble of four stylistic models:

1. A multilayer perceptron (MLP) given LM perplexity features and context/ending lengths.
2. A bag-of-words model that averages the word embeddings of the second sentence as features.
3. A one-layer CNN, with filter sizes ranging from 2-5, over the second sentence.
4. A bidirectional LSTM over the 100 most common words in the second sentence; uncommon words are replaced by their POS tags.

We ensemble the models by concatenating their final representations and passing it through an MLP. On every adversarial iteration, the ensemble is trained jointly to minimize cross-entropy.

The accuracies of these models (at each iteration, evaluated on a 20% split of the test dataset before indices of $\mathcal{A}$ get remapped) are shown in Figure 2. Performance decreases from 60% to close to random chance; moreover, confusing the perplexity-based MLP is not sufficient to lower performance of the ensemble. Only once the other stylistic models are added does the ensemble accuracy drop substantially, suggesting that our approach is effective at reducing stylistic artifacts.

Imagine that you are watching a video clip. The clip has a caption, but it is missing the final phrase. Please choose the best 2 caption endings, and classify each as:
- **likely**, if it completes the caption in a reasonable way;
- **unlikely**, if it sounds ridiculous or impossible;
- **gibberish** if it has such serious errors that it doesn't feel like a valid English sentence.

Example: Someone is shown sitting on a fence and talking to the camera while pointing out horses. Someone
- stands in front of a podium. (**likely**, **second best**)
- rides a horse using a hose. (**unlikely**)
- is shown riding a horse. (**likely**, **best**)
- , the horse in a plaza field. (**gibberish**)

Figure 3: Mechanical Turk instructions (abridged).

3.5 Human verification

The final data-collection step is to have humans verify the data. Workers on Amazon Mechanical Turk were given the caption context, as well as six candidate endings: one found ending and five adversarially-sampled endings. The task was twofold: Turkers ranked the endings independently as likely, unlikely, or gibberish, and selected the best and second best endings (Fig 3).

We obtained the correct answers to each context in two ways. If a Turker ranks the found ending as either best or second best (73.7% of the time), we add the found ending as a gold example, with negatives from the generations not labelled best or gibberish. Further, if a Turker ranks a generated ending as best, and the found ending as second best, then we have reason to believe that the generation is good. This lets us add an additional training example, consisting of the generated best ending as the gold, and remaining generations as negatives.[5] Examples with ≤ 3 non-gibberish endings were filtered out.[6]

We found after 1000 examples that the annotators tended to have high agreement, also generally choosing found endings over generations (see Table 2). Thus, we collected the remaining 112k examples with one annotator each, periodically verifying that annotators preferred the found endings.

4 Experiments

In this section, we evaluate the performance of various NLI models on **Swag**. Recall that models

---

[4]A broad definition of annotation artifacts might include aspects besides lexical/stylistic features: for instance, certain events are less likely semantically regardless of the context (e.g. riding a horse using a hose). For this work, we erred more conservatively and only filtered based on style.

[5]These two examples share contexts. To prevent biasing the test and validation sets, we didn't perform this procedure on answers from the evaluation sets' context.

[6]To be data-efficient, we reannotated filtered-out examples by replacing gibberish endings, as well as generations that outranked the found ending, with candidates from $\mathcal{A}$.

	Label distribution by ending type		Inter-annotator agreement	
Labels	Found end	Gen. end	α	ppa
Best	53.5%	9.3%		
Second Best	20.2%	15.9%	0.43	72%
Neither	26.3%	74.8%		
Likely	80.3%	33.3%		
Unlikely	19.0%	57.5%	0.39	64%
Gibberish	0.7%	9.1%		

Table 2: Annotators tend to label the found ending as likely and within the top 2 (column 2), in other cases the example is filtered out. Both label groups have high inter-annotator agreement, in terms of Krippendorff's α and pairwise percent agreement.

for our dataset take the following form: given a sentence and a noun phrase as context $c = (s, n)$, as well as a list of possible verb phrase endings $V = \{v_1, \ldots, v_4\}$, a model f_θ must select a verb $\hat{i}$ that hopefully matches i_{gold}:

$$\hat{i} = \operatorname*{argmax}_i f_\theta(s, n, v_i) \qquad (4)$$

To study the amount of bias in our dataset, we also consider models that take as input just the ending verb phrase v_i, or the entire second sentence (n, v_i). For our learned models, we train f by minimizing multi-class cross-entropy. We consider three different types of word representations: 300d GloVe vectors from Common Crawl (Pennington et al., 2014), 300d Numberbatch vectors retrofitted using ConceptNet relations (Speer et al., 2017), and 1024d ELMo contextual representations that show improvement on a variety of NLP tasks, including standard NLI (Peters et al., 2018). We follow the final dataset split (see Section 2) using two training approaches: training on the found data, and the found and highly-ranked generated data. See the appendix for more details.

4.1 Unary models

The following models predict labels from *a single span* of text as input; this could be the ending only, the second sentence only, or the full passage.

a. fastText (Joulin et al., 2017): This library models a single span of text as a bag of n-grams, and tries to predict the probability of an ending being correct or incorrect independently.[7]

b. Pretrained sentence encoders We consider two types of pretrained RNN sentence encoders, SkipThoughts (Kiros et al., 2015) and InferSent

---

[7]The fastText model is trained using binary cross-entropy; at test time we extract the prediction by selecting the ending with the highest positive likelihood under the model.

(Conneau et al., 2017). SkipThoughts was trained by predicting adjacent sentences in book data, whereas InferSent was trained on supervised NLI data. For each second sentence (or just the ending), we feed the encoding into an MLP.

c. LSTM sentence encoder Given an arbitrary span of text, we run a two-layer BiLSTM over it. The final hidden states are then max-pooled to obtain a fixed-size representation, which is then used to predict the potential for that ending.

4.2 Binary models

The following models predict labels from *two spans* of text. We consider two possibilties for these models: using just the second sentence, where the two text spans are n, v_i, or using the context and the second sentence, in which case the spans are $s, (n, v_i)$. The latter case includes many models developed for the NLI task.

d. Dual Bag-of-Words For this baseline, we treat each sentence as a bag-of-embeddings (c, v_i). We model the probability of picking an ending i using a bilinear model: $\operatorname{softmax}_i(cWv_i^T)$.[8]

e. Dual pretrained sentence encoders Here, we obtain representations from SkipThoughts or InferSent for each span, and compute their pairwise compatibility using either 1) a bilinear model or 2) an MLP from their concatenated representations.

f. SNLI inference Here, we consider two models that do well on SNLI (Bowman et al., 2015): Decomposable Attention (Parikh et al., 2016) and ESIM (Chen et al., 2017). We use pretrained versions of these models (with ELMo embeddings) on SNLI to obtain 3-way entailment, neutral, and contradiction probabilities for each example. We then train a log-linear model using these 3-way probabilities as features.

g. SNLI models (retrained) Here, we train ESIM and Decomposable Attention on our dataset: we simply change the output layer size to 1 (the potential of an ending v_i) with a softmax over i.

4.3 Other models

We also considered the following models:

h. Length: Although length was used by the adversarial classifier, we want to verify that human validation didn't reintroduce a length bias. For this baseline, we always choose the shortest ending.

i. ConceptNet As our task requires world knowledge, we tried a rule-based system on top of the

---

[8]We also tried using an MLP, but got worse results.

	Model	Ending only found only Val	Test	Ending only found+gen Val	Test	2nd sentence only found only Val	Test	2nd sentence only found+gen Val	Test	Context+2nd sentence found only Val	Test	Context+2nd sentence found+gen Val	Test
misc	Random	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
	Length	26.7	27.0	26.7	27.0								
	ConceptNet					26.0	26.0	26.0	26.0				
Unary models — Sentence encoders	fastText	27.5	26.9	29.9	29.0	29.2	27.8	29.8	29.0	29.4	28.0	30.3	29.8
	SkipThoughts	32.4	32.1	32.2	31.8	33.0	32.4	32.8	32.3				
	InferSent	30.6	30.2	32.0	31.9	33.2	32.0	34.0	32.6				
Unary models — LSTM sequence model	LSTM+GloVe	31.9	31.8	32.9	32.4	32.7	32.4	34.3	33.5	43.1	43.6	45.6	45.7
	LSTM+Numberbatch	32.4	32.6	32.3	31.9	31.9	31.9	34.1	32.8	39.9	40.2	41.2	40.5
	LSTM+ELMo	**43.6**	**42.9**	**43.3**	**42.3**	**47.4**	**46.7**	**46.3**	**46.0**	51.4	50.6	51.3	50.4
Binary models — DualBoW	DualBoW+GloVe					31.3	31.3	31.9	31.2	34.5	34.7	32.9	33.1
	DualBoW+Numberbatch					31.9	31.4	31.6	31.3	35.1	35.1	34.2	34.1
Binary models — Dual sentence encoders	SkipThoughts-MLP					34.6	33.9	36.2	35.5	33.4	32.3	37.4	36.4
	SkipThoughts-Bilinear					36.0	35.7	34.7	34.5	36.5	35.6	35.3	34.9
	InferSent-MLP					32.9	32.1	32.8	32.7	35.9	36.2	39.5	39.4
	InferSent-Bilinear					32.0	31.3	31.6	31.3	40.5	40.3	39.0	38.4
Binary models — SNLI inference	SNLI-ESIM									36.4	36.1	36.2	36.0
	SNLI-DecompAttn									35.8	35.8	35.8	35.7
Binary models — SNLI models (retrained)	DecompAttn+GloVe					29.8	30.3	31.1	31.7	47.4	47.6	48.5	48.6
	DecompAttn+Numberbatch					32.4	31.7	32.5	31.9	47.4	48.0	48.0	48.3
	DecompAttn+ELMo					43.4	43.4	40.6	40.3	47.7	47.3	46.0	45.4
	ESIM+GloVe					34.8	35.1	36.3	36.7	51.9	52.7	52.5	52.5
	ESIM+Numberbatch					33.1	32.6	33.0	32.4	46.5	46.4	44.0	44.6
	ESIM+ELMo					46.0	45.7	45.9	44.8	**59.1**	**59.2**	58.7	58.5
Human	1 turker											82.8	
	3 turkers											85.1	
	5 turkers											**88.0**	
	Expert											85.0	

Table 3: Performance of all models in accuracy (%). All models substantially underperform humans, although performance increases as more context is provided (left to right). We optionally train on found endings only, or found and human-validated generated endings (found+gen).

ConceptNet knowledge base (Speer et al., 2017). For an ending sentence, we use the spaCy dependency parser to extract the head verb and its dependent object. The ending score is given by the number of ConceptNet causal relations[9] between synonyms of the verb and synonyms of the object.
j. **Human performance** To benchmark human performance, five Mechanical Turk workers were asked to answer 100 dataset questions, as did an 'expert' annotator (the first author of this paper). Predictions were combined using a majority vote.

4.4 Results

We present our results in Table 3. The best model that only uses the ending is the LSTM sequence model with ELMo embeddings, which obtains 43.6%. This model, as with most models studied, greatly improves with more context: by 3.1% when given the initial noun phrase, and by an ad-

ditional 4% when also given the first sentence.

Further improvement is gained from models that compute pairwise representations of the inputs. While the simplest such model, DualBoW, obtains only 35.1% accuracy, combining InferSent sentence representations gives 40.5% accuracy (InferSent-Bilinear). The best results come from pairwise NLI models: when fully trained on SWAG, ESIM+ELMo obtains 59.2% accuracy.

When comparing machine results to human results, we see there exists a lot of headroom. Though there likely is some noise in the task, our results suggest that humans (even untrained) converge to a consensus. Our in-house "expert" annotator is outperformed by an ensemble of 5 Turk workers (with 88% accuracy); thus, the effective upper bound on our dataset is likely even higher.

5 Analysis

5.1 SWAG versus existing NLI datasets

The past few years have yielded great advances in NLI and representation learning, due to the availability of large datasets like SNLI and MultiNLI

---

[9]We used the relations 'Causes', 'CapableOf', 'ReceivesAction', 'UsedFor', and 'HasSubevent'. Though their coverage is low (30.4% of questions have an answer with ≥ 1 causal relation), the more frequent relations in ConceptNet, such as 'IsA', at best only indirectly relate to our task.
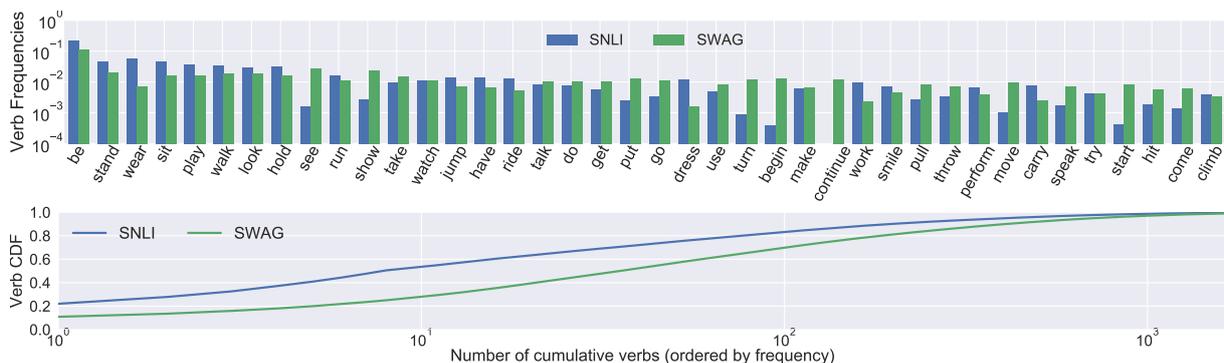
Figure 4: Top: Distribution of the 40 top verbs in the union of SNLI and **SWAG**. Our dataset shows a greater variety of dynamic verbs, such as "move", as well as temporal verbs such as "start" and "come." "Continue" is cut off for SNLI (it has frequency $6 \cdot 10^{-5}$). Bottom: CDF for verbs in SNLI and **SWAG**.

(Bowman et al., 2015; Williams et al., 2018). With the release of **SWAG**, we hope to continue this trend, particularly as our dataset largely has the same input/output format as other NLI datasets. We observe three key differences between our dataset and others in this space:

First, as noted in Section 1, **SWAG** requires a unique type of temporal reasoning. A state-of-the-art NLI model such as ESIM, when bottlenecked through the SNLI notion of entailment (SNLI-ESIM), only obtains 36.1% accuracy.[10] This implies that these datasets necessitate different (and complementary) forms of reasoning.

Second, our use of videos results in wide coverage of dynamic and temporal situations Compared with SNLI, with contexts from Flickr30K (Plummer et al., 2017) image captions, **SWAG** has more active verbs like 'pull' and 'hit,' and fewer static verbs like 'sit' and 'wear' (Figure 4).[11]

Third, our dataset suffers from few lexical biases. Whereas fastText, a bag of n-gram model, obtains 67.0% accuracy on SNLI versus a 34.3% baseline (Gururangan et al., 2018), fastText obtains only 29.0% accuracy on **SWAG**.[12]

5.2 Error analysis

We sought to quantify how human judgments differ from the best studied model, ESIM+ELMo. We randomly sampled 100 validation questions

---

[10] The weights of SNLI-ESIM pick up primarily on entailment probability (0.59), as with neutral (0.46), while contradiction is negatively correlated (-.42).

[11] Video data has other language differences; notably, character names in LSMDC were replaced by 'someone'

[12] The most predictive individual words on SWAG are infrequent in number: 'dotted' with P($+$|dotted) = 77% with 10.3 counts, and P($-$|similar) = 81% with 16.3 counts. (Counts from negative endings were discounted 3x, as there are 3 times as many negative endings as positive endings).

Reason	Explanation	Freq.
Situational	The good ending is better *in context*.	53.7%
Plausibility	The bad ending is implausible *regardless of context*.	14.4%
Novelty	The bad ending seems redundant; it is entailed by the context.	1.8%
Weirdness	The bad ending is semantically or grammatically malformed, e.g. 'the man is getting out of the horse.'	18.1%
Ambiguous	Both endings seem equally likely.	12.0%

Table 4: Justifications for ranking the gold answer over a wrong answer chosen by ESIM+ELMo.

that ESIM+ELMo answered incorrectly, for each extracting both the gold ending and the model's preferred ending. We asked 5 Amazon Mechanical Turk workers to pick the better ending (of which they preferred the gold endings 94% of the time) and to select one (or more) multiple choice reasons explaining why the chosen answer was better.

The options, and the frequencies, are outlined in Table 4. The most common reason for the turkers preferring the correct answer is <u>situational</u> (52.3% of the time), followed by <u>weirdness</u> (17.5%) and <u>plausibility</u> (14.4%). This suggests that ESIM+ELMo already does a good job at filtering out <u>weird</u> and <u>implausible</u> answers, with the main bottleneck being grounded physical understanding. The ambiguous percentage is also relatively low (12.0%), implying significant headroom.

5.3 Qualitative examples

Last, we show several qualitative examples in Table 5. Though models can do decently well by identifying complex alignment patterns between the two sentences (e.g. being "up a tree" implies that "tree" is the end phrase), the incorrect model predictions suggest this strategy is insuffi-

A waiter brings a fork. The waiter
   a) starts to step away. (74.76%)
   b) adds spaghetti to the table. (21.57%)
   c) brings a bunch of pie to the food (2.67%)
   d) drinks from the mug in the bowl. (0.98%)
	He is up a tree. Someone
	   a) stands underneath the tree. (97.44%)
	   b) is at a pool table holding a cup. (1.14%)
	   c) grabs a flower from a paper. (0.96%)
	   d) is eating some cereal. (0.45%)

An old man rides a small bumper car. Several people
   a) get in the parking lot. (76.58%)
   b) wait in the car. (15.28%)
   c) get stuck with other bumper cars. (6.75%)
   d) are running down the road. (1.39%)
	He pours the raw egg batter into the pan. He
	   a) drops the tiny pan onto a plate. (93.48%)
	   b) lifts the pan and moves it around to shuffle the eggs. (4.94%)
	   c) stirs the dough into a kite. (1.53%)
	   d) swirls the stir under the adhesive. (0.05%)

Table 5: Example questions answered by the best model, ESIM+Elmo, sorted by model probability. Correct model predictions are in **blue**, incorrect model predictions are red. The right answers are **bolded**.

cient. For instance, answering "An old man rides a small bumper car" requires knowledge about *bumper cars* and how they differ from regular cars: bumper cars are tiny, don't drive on roads, and don't work in parking lots, eliminating the alternatives. However, this knowledge is difficult to extract from existing corpora: for instance, the ConceptNet entry for Bumper Car has only a single relation: bumper cars are a type of vehicle. Other questions require intuitive physical reasoning: e.g, for "he pours the raw egg batter into the pan," about what happens next in making an omelet.

5.4 Where to go next?

Our results suggest that 𝗦𝘄𝗔𝗚 is a challenging testbed for NLI models. However, the adversarial models used to filter the dataset are purely stylistic and focus on the second sentence; thus, subtle artifacts still likely remain in our dataset. These patterns are ostensibly picked up by the NLI models (particularly when using ELMo features), but the large gap between machine and human performance suggests that more is required to solve the dataset. As models are developed for commonsense inference, and more broadly as the field of NLP advances, we note that AF can be used again to create a more adversarial version of 𝗦𝘄𝗔𝗚 using better language models and AF models.

6 Related Work

Entailment NLI There has been a long history of NLI benchmarks focusing on linguistic entailment (Cooper et al., 1996; Dagan et al., 2006; Marelli et al., 2014; Bowman et al., 2015; Lai et al., 2017; Williams et al., 2018). Recent NLI datasets in particular have supported learning broadly-applicable sentence representations (Conneau et al., 2017); moreover, models trained on these datasets were used as components

for performing better video captioning (Pasunuru and Bansal, 2017), summarization (Pasunuru and Bansal, 2018), and generation (Holtzman et al., 2018), confirming the importance of NLI research. The NLI task requires a variety of commonsense knowledge (LoBue and Yates, 2011), which our work complements. However, previous datasets for NLI have been challenged by unwanted annotation artifacts, (Gururangan et al., 2018; Poliak et al., 2018) or scale issues. Our work addresses these challenges by constructing a new NLI benchmark focused on grounded commonsense reasoning, and by introducing an adversarial filtering mechanism that substantially reduces known and easily detectable annotation artifacts.

Commonsense NLI Several datasets have been introduced to study NLI beyond linguistic entailment: for inferring likely causes and endings given a sentence (COPA; Roemmele et al., 2011), for choosing the most sensible ending to a short story (RocStories; Mostafazadeh et al., 2016; Sharma et al., 2018), and for predicting likelihood of a hypothesis by regressing to an ordinal label (JOCI; (Zhang et al., 2017)). These datasets are relatively small: 1k examples for COPA and 10k cloze examples for RocStories.[13] JOCI increases the scale by generating the hypotheses using a knowledge graph or a neural model. In contrast to JOCI where the task was formulated as a regression task on the degree of plausibility of the hypothesis, we frame commonsense inference as a multiple choice question to reduce the potential ambiguity in the labels and to allow for direct comparison between machines and humans. In addition, 𝗦𝘄𝗔𝗚's use of adversarial filtering increases diversity of situations and counterfactual generation quality.

---

[13]For RocStories, this was by design to encourage learning from the larger corpus of 98k sensible stories.

Last, another related task formulation is sentence completion or cloze, where the task is to predict a single word that is removed from a given context (Zweig and Burges, 2011; Paperno et al., 2016).[14] Our work in contrast requires longer textual descriptions to reason about.

Vision datasets Several resources have been introduced to study temporal inference in vision. The Visual Madlibs dataset has 20k image captions about hypothetical next/previous events (Yu et al., 2015); similar to our work, the test portion is multiple-choice, with counterfactual answers retrieved from similar images and verified by humans. The question of 'what will happen next?' has also been studied in photo albums (Huang et al., 2016), videos of team sports, (Felsen et al., 2017) and egocentric dog videos (Ehsani et al., 2018). Last, annotation artifacts are also a recurring problem for vision datasets such as Visual Genome (Zellers et al., 2018) and Visual QA (Jabri et al., 2016); recent work was done to create a more challenging VQA dataset by annotating complementary image pairs (Goyal et al., 2016).

Reducing gender/racial bias Prior work has sought to reduce demographic biases in word embeddings (Zhang et al., 2018) as well as in image recognition models (Zhao et al., 2017). Our work has focused on producing a dataset with minimal annotation artifacts, which in turn helps to avoid some gender and racial biases that stem from elicitation (Rudinger et al., 2017). However, it is not perfect in this regard, particularly due to biases in movies (Schofield and Mehr, 2016; Sap et al., 2017). Our methodology could potentially be extended to construct datasets free of (possibly intersectional) gender or racial bias.

Physical knowledge Prior work has studied learning grounded knowledge about objects and verbs: from knowledge bases (Li et al., 2016), syntax parses (Forbes and Choi, 2017), word embeddings (Lucy and Gauthier, 2017), and images and dictionary definitions (Zellers and Choi, 2017). An alternate thread of work has been to learn scripts: high-level representations of event chains (Schank and Abelson, 1975; Chambers and Jurafsky, 2009). SWAG evaluates both of these strands.

---

[14]Prior work on sentence completion filtered negatives with heuristics based on LM perplexities. We initially tried something similar, but found the result to still be gameable.

7 Conclusion

We propose a new challenge of physically situated commonsense inference that broadens the scope of natural language inference (NLI) with commonsense reasoning. To support research toward commonsense NLI, we create a large-scale dataset SWAG with 113k multiple-choice questions. Our dataset is constructed using Adversarial Filtering (AF), a new paradigm for robust and cost-effective dataset construction that allows datasets to be constructed at scale while automatically reducing annotation artifacts that can be easily detected by a committee of strong baseline models. Our adversarial filtering paradigm is general, allowing potential applications to other datasets that require human composition of question answer pairs.

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642.

Zheng Cai, Lifu Tu, and Kevin Gimpel. 2017. Pay attention to the ending: Strong neural baselines for the

roc story cloze task. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 616–622.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised Learning of Narrative Schemas and Their Participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 602–610, Stroudsburg, PA, USA. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1657–1668.

Gennaro Chierchia and Sally McConnell-Ginet. 2000. *Meaning and Grammar (2Nd Ed.): An Introduction to Semantics*. MIT Press, Cambridge, MA, USA.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Robin Cooper, Dick Crouch, JV Eijckl, Chris Fox, JV Genabith, J Japars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. A framework for computational semantics (fracas). Technical report, Technical report, The FraCaS Consortium.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

Kiana Ehsani, Hessam Bagherinezhad, Joseph Redmon, Roozbeh Mottaghi, and Ali Farhadi. 2018. Who let the dogs out? modeling dog behavior from visual data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Panna Felsen, Pulkit Agrawal, and Jitendra Malik. 2017. What will happen next? forecasting player moves in sports videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3342–3351.

Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 266–276.

JJ Gibson. 1979. The ecological approach to visual perception. *Houghton Mifflin Comp*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. *arXiv preprint arXiv:1612.00837*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970.

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

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1638–1649. Association for Computational Linguistics.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239.

Allan Jabri, Armand Joulin, and Laurens van der Maaten. 2016. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 427–431.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-Captioning Events in Videos. In *International Conference on Computer Vision (ICCV)*.

Alice Lai, Yonatan Bisk, and Julia Hockenmaier. 2017. Natural language inference from multiple premises. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 100–109, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1445–1455, Berlin, Germany. Association for Computational Linguistics.

Peter LoBue and Alexander Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 329–334. Association for Computational Linguistics.

Li Lucy and Jon Gauthier. 2017. Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 76–85.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1314.

Nasrin Mostafazadeh, Nathanael Chambers, Xiadong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. In *NAACL*.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernandez. 2016. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255.

Ramakanth Pasunuru and Mohit Bansal. 2017. Multitask video captioning with video and entailment generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1273–1283, Vancouver, Canada. Association for Computational Linguistics.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 646–653. Association for Computational Linguistics.

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

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Int. J. Comput. Vision*, 123(1):74–93.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. In *Joint Conference on Lexical and Computational Semantics (StarSem)*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie Description. *International Journal of Computer Vision*, 123(1):94–120.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL*

Workshop on Ethics in Natural Language Processing, pages 74–79.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334.

Roger C. Schank and Robert P. Abelson. 1975. Scripts, plans, and knowledge. In *Proceedings of the 4th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'75, pages 151–157, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Alexandra Schofield and Leo Mehr. 2016. Gender-distinguishing features in film dialogue. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 32–39.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A. Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *Proc. of CoNLL*.

Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. 2018. Tackling the story ending biases in the story cloze test. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 752–757.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems*, pages 926–934.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Mitchell Stern, Jacob Andreas, and Dan Klein. 2017. A minimal span-based neural constituency parser. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 818–827.

Vladimir Vapnik. 2000. *The Nature of Statistical Learning Theory*, 2 edition. Information Science and Statistics. Springer-Verlag, New York.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. 2015. Visual Madlibs: Fill in the blank Image Generation and Question Answering. *ICCV*.

Rowan Zellers and Yejin Choi. 2017. Zero-shot activity recognition with verb attribute induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *Conference on Computer Vision and Pattern Recognition*.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Conference on Artificial Intelligence, Ethics and Society*.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal Common-sense Inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *arXiv preprint arXiv:1506.06724*.

Geoffrey Zweig and Christopher JC Burges. 2011. The microsoft research sentence completion challenge. Technical report, Citeseer.