A Appendix for SWAG

A.1 More detail about video datasets

As mentioned in the main paper, we obtained contexts and found endings from video data. The videos in the ActivityNet dataset are already broken up into into clips. However, the LSMDC dataset contains captions for the entire movie, so it is possible that temporally adjacent captions describe events that are far apart in time. Thus, we don't include any pair of captions that have a time-difference of more than 25 seconds.

In addition to the datasets we used, we also considered the DiDeMo dataset, which consists of (often several) referring expressions in a video (Hendricks et al., 2017). However, many of the referring expressions are themselves sentence fragments, (e.g. "first time we see people" so we ultimately did not use this dataset.) Additionally, we considered the Visual Madlibs dataset (Yu et al., 2015), as it contains 10k hypothetical captions written by Mechanical Turk workers about what might happen *next* given an image. However, these captions are fundamentally different from the rest of the data (as they're about what *might*) happen next; as a result, they use different types of language. They also have different tenses versus the other datasets that we considered (e.g. past tense), as a result of the "Mad-libs" style of data collection.

A.2 Details of the language model

Our language model follows standard best practices: the input and output embedding layers are tied (Inan et al., 2017; Press and Wolf, 2017), all embedding and hidden layers are set to 512, and we used recurrent dropout (Gal and Ghahramani, 2016) on the hidden states and embedding layer. We additionally train a *backwards* language model alongside the forward language model, and they share embedding parameters. This adds extra supervision to the embedding layer and gives us another way to score candidate generations. We first pretrain the language model for two epochs on pairs of two sentences in the Toronto Books dataset (Zhu et al., 2015), and then train on sentence pairs from ActivityNet Captions and LSMDC, validating on held-out perplexity. For optimization, we use Adam (Kingma and Ba, 2015) with a learning rate of 10^{-3} and clip gradients to norm 1.0.

All of the above details were validated using perplexity on a held-out set of the video datasets during early experimentation. Our final development set forward perplexity was 31.2 and backward perplexity was 30.4. We tried more complicated language modeling architectures, such as from (Józefowicz et al., 2016), but ended up not seeing an improvement due to overfitting.

A.3 Language model features for the MLP, during adversarial filtering

We obtained LM perplexity features to be used during adversarial filtering in the following ways, using both directions of the bidirectional language model. We extract perplexities for the context by itself (going forward), the ending given the context (going forward), the context given the ending (going backward), and the ending by itself (going backward). We also extract the probability of the final generated token going forward, since sentences sometimes reach the length limit of 25 tokens and end unnaturally.

A.4 Refinining the generated answers to four distractors

In the main paper, we noted that we started with 1023 negatives per example, which the adversarial filtering process filtered down to 9. Five of these were passed to mechanical turk workers, and we were left with anywhere between 0 and 4 of these per example as "distractors." (Note that we always were filtering out the second best option that the was selected by the turkers). This means that for many of our examples (62%) we actually have a fourth distractor. In these cases, we sorted the distractors by their "unlikely/likely" score, so that the fourth distractor was the one deemed most likely. We still provided the fourth distractor in the training set to be possibly used in future work, however we didn't train on it for simplicity.

A.5 More information about Mechanical turk

We used several tricks to keep the interannotator agreement high (with a pairwise percent agreement of 79% at classifying an ending as either in the Top 2). First, we had a screening HIT where turkers were given detailed instructions for the task, and only the best-scoring turk workers qualified for the remaining HITs. Second, we periodically dequalified turkers that had a low agreement with the gold endings: any turk worker with an accuracy of less than 55% of classifying the "gold"

Questions with only generated endings	25,618
Questions with one original ending	87,939
Questions in total	113,557
Sentence pairs from ActivityNet	51,439
Sentence pairs from LSMDC	62,118
Unique contexts	92,221
Unique endings	452,683

Table 1: Statistics of SWAG.

Freq	Topic words
5.0%	ball, pull, hit, wall, inside, time, game, rope, team
4.9%	window, red, long, drink, bowl, ingredient, mix
6.1%	arm, speak, appear, climb, tree, roll, like, roof, edge
4.0%	water, bar, board, blue, boat, fly, river, join, dive
5.3%	eye, smile, close, little, lean, cover, remove, lip
4.6%	walk, outside, street, wave, pass, beach, sidewalk
5.7%	field, drop, slide, drive, right, kick, park, road, chest
4.7%	watch, dog, flip, stick, land, demonstrate, trick, mat
4.5%	dance, lift, try, line, snow, gun, catch, hill, bend
4.6%	fall, crowd, pour, shake, finish, raise, grass, wooden
5.9%	perform, spin, house, stage, routine, fence, bow

Table 2: A visualization of the diversity of the dataset, using a topic model (Blei et al., 2003).

ending as the best or second best, over 10 or more HITs, had the qualification taken away. We also gave small bonuses to turkers with high accuracy.

During our crowdsourcing, we tried to pay the Turkers a fair wage (median $8.57 per hour) and they left positive comments for us on TurkOpticon and TurkerView. The total dataset cost was $23,000, or an average of 20 cents per example.

A.6 Implementation details of the models considered

We implemented the neural models in PyTorch using the AllenNLP library (Gardner et al., 2018). Our experiments use the Adam optimizer (Kingma and Ba, 2015), with a learning rate of 10^{-3} and gradient clipping, except for Decomposable Attention and ESIM, where we use the AllenNLP default configurations.

A.7 More info about dataset diversity

The final dataset has a vocabulary size of 21000. We also visualize the coverage of the dataset with a Topic model (see Table 2).

A.8 Comparing the distribution of verbs with MultiNLI

We also produced an extension to Figure 4 of the main paper, that involves verbs from MultiNLI, in Figure 1. We ended up not including it in the paper because we wanted to focus our comparison between SNLI and SWAG (as they are both grounded datasets). Interestingly, we find that SWAG has a less skewed cumulative distribution of verbs up to around 120, when afterwards MultiNLI has a slightly less skewed distribution. This is possibly due to the broader set of domains considered by MultiNLI, whereas we consider videos (which is also a broad domain! but still underrepresents words highly used in newswire text, for instance.)

A.9 More examples

We have more qualitative examples in Table 3.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027.

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

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Hakan Inan, Khashayar Khosravi, and Richard Socher. 2017. Tying word vectors and word classifiers: A loss framework for language modeling. In *ICLR*. ICLR.

Rafal Józefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *CoRR*, abs/1602.02410.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*. ICLR.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 157–163.

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

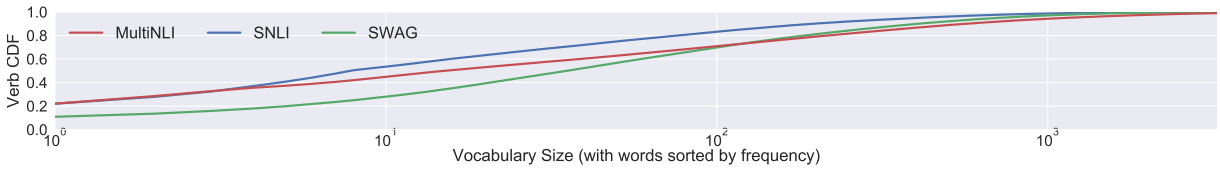

Figure 1: Bottom: CDF for verbs in SNLI, SWAG, and MultiNLI.

The lady demonstrates wrapping gifts using her feet. The lady a) shows us the different shapes of the ornaments. (99.67%) b) continues playing when the lady talks to the camera. (0.26%) c) takes the desserts from the box and continues talking to the camera . (0.07%) **d) cuts the paper with scissors. (0.01%)**	In a cafeteria, someone holds a combination tray and bowl in one hand. With the other, he a) heads into his own study. (80.67%) b) glances around and studies the photo of the blonde someone. (8.45%) **c) struggles to serve himself food with chopsticks. (6.82%)** d) opens the wall , revealing an expanse of bed within. (4.06%)
As he approaches , his kayak flips upside-down. As the view follows him, we a) see silhouetted black clouds making him zoom out of the trees, catching smoke. (42.54%) b) drift over a busy city street , like down buildings on the tarmac. (41.41%) c) find someone climbing into a tawny grave atop a road drawn among german soldiers. (13.73%) **d) notice another man seated on the rocks to the right in red with a white helmet. (2.32%)**	A man is bending over a sink. He a) takes a rag from over the sink, putting it in his mouth. (89.54%) **b) is spraying a small dog with a hose. (6.07%)** c) is carrying a shaving machine with a pressure washer. (4.29%) d) is putting a pair of shaving glass on the side of his face. (0.10%)
People are walking next to the camels leading them. A building a) is shown riding the camels. (90.72%) **b) is shown in the background. (8.39%)** c) with a rifle is leading them. (0.87%) d) is then shown for several clip. (0.01%)	A hockey game is in progress. two hockey players a) walked together in the middle of a field. (48.11%) b) walk past with a goal. (44.00%) c) sit around a rope watching the other team. (5.30%) **d) ram into each other and begin fighting. (2.58%)**
Meanwhile, someone parries another giant 's attacks. The giant a) strikes a fight and thuds into someone as he rushes in, who briefly flees. (89.96%) **b) knocks someone 's sword out of his hand. (5.25%)** c) spins him across the bars. (4.55%) d) throws stick to the bat, dragging around. (0.24%)	A lady pours ice in a glass. The lady a) pours ice into the glass. (65.14%) b) measures the contents of the glass. (33.56%) c) pours lemon mixture into a glass and pours liquids into asian juice. (0.87%) **d) adds 3 liquors and lemon juice. (0.43%)**
The stars emerge from behind the clouds. Someone a) backs away from the windows of the clip, as lightning billows over the sky. (96.59%) b) walks back across the room with nothing of his own. (1.82%) **c) stands on his boat and looks at a deep orange and red sunset. (1.47%)** d) shoots the man 's shoulder sideways, but neither do anything for a few seconds. (0.12%)	Someone stands waiting with the bridesmaids. Everyone a) seems to be ecstatic. (78.33%) **b) looks around as someone walks down the aisle, arm-in-arm with someone 's uncle. (8.97%)** c) holds someone 's eyebrow. (8.84%) d) looks at her anxiously as someone walks and sits in his seat. (3.85%)

Table 3: More (incorrect) questions answered by the best model, ESIM+Elmo, sorted by model probability. The right answers are **bolded**.

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*.