

Multimodal Natural Language Inference Final Report

Vincent Sitzmann, Martina Marek, Leonid Keselman
Stanford University

{sitzmann,martinam,leonidk}@stanford.edu

Abstract

We explore how natural language inference (NLI) tasks can be augmented with the use of visual information. Namely, we replicate and expand existing baselines for NLI, including recent deep learning methods. By adding image features to these models, we explore how the textual and visual modalities interact. Specifically, we show that image features can provide a small boost in classifier performance for simpler models, but are a subset of information provided in the premise statement and thus do not benefit complex models. Additionally, we demonstrate a weakness in the SNLI dataset, showing that textual entailment is predictable without reference to the premise statement.

1 Introduction

Given an image, as well as a human-generated caption (premise statement), we want to predict whether a second statement (hypothesis statement) is entailed, neutral, or contradictory with regard to the premise statement. This task is known as natural language inference (NLI) or recognizing textual entailment (RTE). While existing approaches have focused on tackling the inference task solely on the given statements, we plan to improve these results by combining the language features of the two statements with visual information from an image.

This task extends the scope of the classical NLI and requires a combination of both inference models and visual models. As such, it combines aspects from two major fields of artificial intelligence: natural language understanding and computer vision. To this end, we are working on the Stanford Natural Language Inference dataset which contains premise and hypothe-

sis statements, where the premise statements originate from image captions.



Caption	A person in a black wetsuit is surfing a small wave.
Entailment	A person is surfing a wave.
Contradiction	A woman is trying to sleep on her bed.
Neutral	A person surfing a wave in Hawaii.

Figure 1: An example of the data given in the SNLI dataset, with the associated image from the Flickr30k dataset

2 Related Work

The Stanford Natural Language Inference (SNLI) Corpus (Bowman et al., 2015) evaluated the performance of several natural language inference models on their new, larger dataset. They report three major baselines. The first and best-performing baseline is a linear classifier with both lexicalized and unlexicalized sentence features. The second baseline feeds sum-of-GloVe sentence embeddings (Pennington et al., 2014) into a fully connected neural network. The third uses an LSTM to embed both the premise and hypotheses sentences into a low-dimensional vector space and then feeds these vectors into a three-layer neural network. SNLI is described in further detail in section 4.

On previous NLI datasets, such as RTE (Giampiccolo et al., 2007) and SICK (Nakov and Zesch, 2014), high performing approaches usually drew from a rich set of hand-crafted features and used additional resources (Giampiccolo et al., 2007), (Hickl et al., 2006), (Lai and Hockenmaier, 2014). This was necessary to compensate for small training set sizes. These machine learning based models proved to be the most successful methods on these datasets, usually outperforming logic inference based approaches (Tatu and Moldovan, 2005).

The much larger SNLI dataset (Bowman et al., 2015) now enables complex deep learning approaches to outperform feature-based systems. Designed to model many aspects of the highly complex structure of language, these models are now applied successfully to inference tasks. (Bowman et al., 2016) introduces a tree-structured RNN that is able to capture the hierarchical structure of natural language without the need for expensive preprocessing such as parsing, and supports batched computation. (Liu et al., 2016) beat the previous encoder-based implementations by using a bidirectional LSTMN with an attention mechanism to build the sentence encodings. Classification was then done over the concatenation, product and difference, achieving near state-of-the-art performance.

The best-performing models on the SNLI corpus are attention models that are able to reference the premise statement when inferring on the hypothesis statement (Rocktäschel et al., 2015) (Wang and Jiang, 2015). The current state-of-the-art result on the SNLI corpus was achieved by the Long-Short-Term Memory Network (LSTMN) model proposed in (Cheng et al., 2016). It generalizes the standard LSTM architecture by introducing memory and hidden state tapes that grow dynamically with each input word and allow the hypothesis-parsing network to guide attention over the states of the premise-parsing network in a process they call deep attention fusion. They show a significant improvement against the original baselines, as shown in table 1.

On the second aspect of the multimodal inference task, incorporating visual information into existing inference methods, no work has been done to the best of our knowledge. However, since the introduction of Convolutional Neural Networks (CNNs) for image classification in 2012

(Krizhevsky et al., 2012), their capability to learn rich image representations has enabled the use of images in multimodal settings. For instance, a lot of recent work has addressed generative models for text descriptions of images, such as (Karpathy and Fei-Fei, 2015) and (Vinyals et al., 2015). The former uses a bidirectional Recurrent Neural Network, multiple region proposals, and infers correspondences between sentences and proposals. The latter is simpler, in that it takes a full image input, uses a standard classification CNN, and learns a unidirectional neural network. For more sophisticated tasks, such as Visual Question Answering (Antol et al., 2015), models with selective attention have been designed, which allow the language model to selectively focus on parts of the image when it is asked to evaluate certain words (Xu et al., 2015). These are made possible by learning a fully-differential soft attention model over entire images.

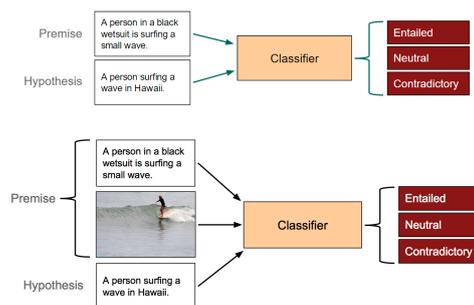


Figure 2: A comparison of classical natural language inference (above) and the multimodal natural language inference (below) explored in this paper.

3 Multimodal Natural Language Inference

In this paper, we extend the usual Natural Language Inference task with visual information. Originally, given a premise and hypothesis sentence, the task is to decide whether the given sentences entail, contradict or are neutral to each other. Multimodal Natural Language Inference, on the other hand, takes an image-caption pair as the premise, and then decides whether the hypothesis sentence is an entailment, contradiction or neutral to the premise pair. A visual overview of both tasks is shown in figure 2.

4 Data

The Stanford Natural Language Inference (SNLI) Corpus is the largest NLI dataset to date, with human generated annotations and enough data to enable deep learning models (Bowman et al., 2015). It comprises 570,152 image, caption and hypothesis triplets, where the caption represents the premise and the hypothesis sentence is either contradictory, neutral or entailed by the premise. Most of the captions were crowdsourced in response to images from the Flickr 30k corpus (Young et al., 2014), while 4,000 captions were taken from the VisualGenome dataset (Krishna et al., 2016). The hypothesis captions were collected in an Amazon Mechanical Turk setting, where the Turkers were only shown the caption and were asked to write a hypothesis statement that is entailed, contradictory or neutral with respect to the premise statement. As each image in the dataset is annotated with all three types of labels, the classes are balanced and there exists an entailed, neutral and contradictory statement for every caption. An example datum is shown in figure 1.

4.1 Image Feature Extraction

To extract rich image features, we use a Convolutional Neural Network (CNN). As demonstrated in recent literature (Razavian et al., 2014), using the top layer vector of a neural network trained for classification can create a compact image representation that can easily produce state-of-the-art results across a wide range of vision problems.

To extract these features, we used Google’s pre-trained Inception-v3 net (Szegedy et al., 2015). Its architecture is depicted in figure 4.1. This network is different than those used in other visual-language datasets (Antol et al., 2015), but is known to perform better for classification; scoring a top-5 accuracy of a stunning 96.54% on the ImageNet challenge (Deng et al., 2009). We fed all Flickr30k (Young et al., 2014) images through the inception network and extracted the activations before the final 1000 dimensional fully-connected layer. This yields a 2048-dimensional feature vector for each of the Flickr30k images. We also extract 8x8x1280 dimensional features for a soft attention model and the LSTMN, as described in sections 5.4 and 5.3. All other sections use the 2,048 dimensional vectors.

Due to broken URLs, a small number (less than 1%) of the Flickr30k images are no longer

type	patch size/stride or remarks	input size
conv	3×3/2	299×299×3
conv	3×3/1	149×149×32
conv padded	3×3/1	147×147×32
pool	3×3/2	147×147×64
conv	3×3/1	73×73×64
conv	3×3/2	71×71×80
conv	3×3/1	35×35×192
3×Inception	As in figure 5	35×35×288
5×Inception	As in figure 6	17×17×768
2×Inception	As in figure 7	8×8×1280
pool	8×8	8×8×2048
linear	logits	1×1×2048
softmax	classifier	1×1×1000

Figure 3: Architecture of Google’s Inception Net. The layers we use for feature extraction are marked in red. 2,048 dimensional features are used for experiments with off-the-shelf sentence embeddings such as hand-crafted features and sum-of-GloVe embeddings. The 8x8x1280 features are used for deep-learned encoder-decoder architectures such as the LSTMN and soft attention models.

available, and we simply generate random 2048-dimensional feature vectors or a zero 8x8x1280 feature map for those images.

5 Models

5.1 Feature-Based Model

Similar to the previous approaches on the RTE (Giampiccolo et al., 2007) and SICK datasets (Nakov and Zesch, 2014), the SNLI paper (Bowman et al., 2015) proposed a simple feature-based baseline on their presented dataset. It draws its features from 6 different sets: BLEU scores (Pa-

Classifier/Model	Accuracy
Feature Baseline (Bowman et al., 2015)	78.1%
Neural Network (Bowman et al., 2015)	75.3%
LSTMN (Cheng et al., 2016)	86.3%
Feature Baseline	71.8%
Feature Baseline + Images	70.8%
Neural Network	73.3%
Neural Network + Images	74.0%
LSTMN	67.0%
LSTMN + Images	73.0%

Table 1: Comparison of our methods against the published state-of-the-art results

pineni et al., 2002), length difference between the sentence pairs, word overlap, uni- and bigram indicators for the hypothesis, and indicator features for cross-unigrams as well as cross-bigrams between the premise and hypothesis. They then train these features with a linear classifier.

We implemented this model with most of the features mentioned above, with the following differences: The overlap is computed only over all words, not separately over adjectives, adverbs, nouns and verbs on top of that. Furthermore, before extracting the features, we do some preprocessing: All stopwords and dots at the end of the sentence are removed, and words are stemmed to get a denser feature representation. When classifying over the hypothesis only, features were only drawn from the unigrams and bigrams of the hypothesis since no premise was given for this task. Consequently, there were no unlexicalized features.

We used NLTK (Loper and Bird, 2002) to stem our words. Since the original paper doesn't clarify its choice of linear classifiers, we tested both an SVM (Fan et al., 2008) and a Logistic Regression classifier with python's sklearn library (Pedregosa et al., 2011). We found the Logistic Regression classifier performed better and therefore use it as our reference baseline. For efficiency reasons, we trained with only the most common 10,000 lexicalized features, plus 7 additional unlexicalized features.

For the multimodal inference results, we took the 2048-dimensional vector of activations of the last fully connected layer of the inception architecture and concatenated it with the natural language features as described above.

5.2 Sum of GloVe Neural Network Model

This model aims at reproducing the sum-of-GloVe vectors baseline from the original SNLI paper (Bowman et al., 2015). Both textual inputs, the hypothesis and premise, are embedded into a 300 dimensional sentence encoding to create a compact representation for classification by summing the sentence's GloVe word vectors (Pennington et al., 2014). For words lacking a GloVe vector, we generate a random 300d number and store that. We tried two other models, ignoring unknown words and using a universal unknown word token, but we found the random generation technique was most useful across all classifier models.

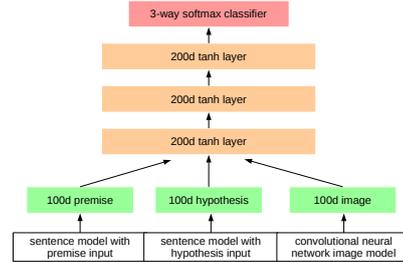


Figure 4: The neural network classification architecture: for each sentence embedding model, the model is run with the two sentences as input, and those outputs are concatenated with an image feature vector representation. The sentence models' outputs are used as the two 100d inputs.

While the original baseline method simply used the GloVe vectors as an initial seed and then back-propagated into the embedding, thus enabling the model to update those vectors, we simply used the GloVe vectors as-is. We spoke with the authors of the original paper and were told the improvement would be marginal and greatly increases training and model complexity. Additional results reported on SNLI (Kelcey, 2016) show the result is noisy, and only yields about 1% improvement. In fact, they report random embeddings outperform GloVe embeddings when given enough training time. Due to limited time and resources, we worked with the original GloVe vectors as our word representations.

Before classification, the sentence embeddings are projected into a 100-dimensional feature space. The projection weights are shared across both premise and hypothesis; we experimented with using separate weights and found slightly worse performance. Following the SNLI baseline, these projections are then concatenated and fed into a 3-layer fully connected neural network with 200 hidden units per layer and TanH activations. As reported earlier, we additionally tried SVMs, Random Forests, and other classifiers, but found the neural network approach to work best.

As shown in figure 4, we extend the model to be multimodal by taking the 2048-dimensional image feature vectors, as explained in section 4.1. This feature vector is also projected into a 100-dimensional space with a fully connected layer and concatenated with the premise and hypothesis

projections.

5.3 Long-Short-Term-Memory Network

As a last model, we implemented the state-of-the-art model on the SNLI dataset, the LSTMN encoder-decoder architecture with deep attention fusion described in (Cheng et al., 2016). This model extends the classic LSTM architecture by generalizing the memory and hidden states to memory and hidden state *tapes* - for every new word that the LSTM parses, it adds a new entry to these two tapes. After each sentence, the tapes are reset. This allows for an attention mechanism over previously read words within a single sentence-parsing LSTM. Two such LSTMNs are implemented to parse the premise and the hypothesis. The hypothesis-parsing LSTMN is further equipped with an attention model over the hidden and memory tapes of the premise-parsing LSTMN. This architecture, which (Cheng et al., 2016) title deep attention fusion, allows the hypothesis-parsing LSTMN to consider the premise embedding with each and every word it parses and thus to notice parallels or contradictions. To generate a final prediction, the memory and hidden state tapes of encoder and decoder are averaged and concatenated, yielding a single embedding. This embedding is then fed into a three-layer fully connected network with ReLU activations and dropout that produces the final predictions. We attempt to reproduce the test performance of 86.3% on the SNLI dataset.

For our multimodal experiments, we consider the image and the premise statement as a multimodal representation of world knowledge. Consequently, the hypothesis statement can be imagined as a query into this world representation. We follow this concept when adding image features to the LSTMN architecture: First, we use a 1x1 convolution to reduce the number of channels in the image feature map from 1280 to the number of hidden units of the LSTMN, which is also the depth dimension of its hidden and memory tapes. We then concatenate the projected image with the hidden and memory tapes of the premise-parsing LSTMN, extending them by 64 states that represent 64 locations in the image. This allows the decoder with deep attention fusion to guide attention not only over the hidden states of the premise statement, but also over each location in the image.

5.4 Soft Image Attention

As an intermediary step of the LSTMN with image features and the sum-of-GloVe-based Neural Network, we implemented an architecture that takes as input the two sum-of-GloVe embeddings of the premise and hypothesis statements and uses them to guide attention over the 8x8x1280 feature map extracted from one of the last convolutional layers of the inception architecture. This is motivated by recent successes in visual question answering (Xu and Saenko, 2015) and image captioning (Xu et al., 2015).

Specifically, the sum-of-GloVe embeddings form the input to a LSTM which implements a soft attention model as presented in (Bahdanau et al., 2014). This LSTM then uses the input to calculate a soft attention map over the image feature map. This soft attention map is essentially a weight for each of the 64 spatial features in the image feature map. The image features are then fused in a weighted sum. In three hops, the LSTM thus accumulates information about relevant parts of the image. The last hidden state of the LSTM is then fed into a three-layer neural network with ReLU activations to produce a final prediction.

6 Results

This section describes the results obtained with the previously introduced models on the original SNLI task, when classifying over the hypothesis only, and when adding image features to these two tasks, respectively.

6.1 SNLI Results

Our implementation of the feature-based linear classifier from (Bowman et al., 2015) achieved an accuracy of 76%, very close to the 78.2% presented in the paper. After limiting the number of features to 10,000, though, the accuracy dropped to 71.8%. This adjustment was necessary to allow efficient computation when adding image features.

The sum-of-GloVe neural network classifier achieved an accuracy of 73.3% on the original SNLI task. As with the feature-based classifier, this model was setup to be quick to train and run experiments on, as our primary task was to test the value of multimodal input. As reported in section 5.2, we only trained for 10 epochs and did not include dropout. These results outperformed the feature-based classifier with a limited number of features, but did not reach the performance with

Classifier Inputs	Handcrafted Feature Accuracy	GloVe NN Accuracy	LSTMN Accuracy
Hypothesis	62.1%	64.6%	-
Hypothesis + Image	59.7%	67.1%	-
Hypothesis + Premise	71.8%	73.3%	70.35%
Hypothesis + Image + Premise	70.8%	74.0%	70.35%

Table 2: Results of our multimodal experiments run over a suite of different classifier modals

the full set of features. This parallels the results from (Bowman et al., 2015), whose feature-based baseline outperformed the sum-of-GloVe vectors neural network, as well as the simple LSTM neural network baselines.

Finally, we tested our implementation of the LSTMN model (Cheng et al., 2016). Due to limited time, we were unable to run the number of experiments that would have been necessary to fully explore the performance of this architecture - we were confronted with issues of severe overfitting and thus were forced to add significant dropout, thus slowing down training. As a result, our results stay far behind the performance reported in in (Cheng et al., 2016). Currently, we see a test accuracy of 70.4% and a train accuracy of 70%, but we are confident that given more time and compute we would be able to reproduce the results reported in (Cheng et al., 2016), as we saw the network successfully overfit the training data.

6.2 Hypothesis-only Results

As part of our exploration of multiple modalities and the value of different inputs, we ran our classifiers given only the hypothesis sentences, without either a textual or image premise. Since no correct inference should be possible when the premise is missing, the accuracy should mirror a random guessing, and only get about 33%. Surprisingly, though, we found that all classifiers perform considerably better, yielding accuracies over 60%. We explore the impact of these results in section 7.2. We additionally trained a premise-only baseline, but since the classes are balanced and input is identical to multiple labels, we achieved the expected 33 % training and testing accuracy.

6.3 Multimodal Results

In a next step, we added image features to both the hypothesis only, and original SNLI tasks to see whether the results can be improved. Again, results are shown in table 2.

For the feature-based linear classifier, we found that adding image features decreases the accuracy slightly. It seems the linear classifier has trouble adjusting to high dimensional features from the CNN classification task. In both the hypothesis-only and textual premise classifiers, accuracy decreased.

The more complex Neural Network-based GloVe model, on the other hand, is able to extract valuable information from the image features. The image plus hypothesis classifier outperforms the hypothesis-only model by several percentage points. However, it only provides a minor increase in the accuracy when added to the model that includes the premise.

We additionally tested our soft-attention model described in section 5.4, but saw it only obtain 72.3% on the full multimodal task. This was less than the much simpler sum-of-GloVe neural network presented in section 5.2. Based on our analysis in section 7.2, it seems that visual attention alone is not beneficial. Thus, the results from this model are included here for reference, but not further elaborated upon in the results and analysis.

For our most complex model, the LSTMN with deep attention fusion, we find that image features do not seem to improve performance. This seems to suggest that complex models such as the LSTMN are able to extract all relevant information from the hypothesis statement and the image features are thus redundant. However, while we trained the LSTMN for up to 10 hours on a high-performance GPU cluster, we did not have enough time to train these models for several days, which would have been required to train them up to their final performance. As a result, their performance is still not close to the performance of the original model by (Cheng et al., 2016).

Model	Entail	Neutral	Contra	Total
FT H	65.1%	60.0%	61.2%	61.2%
FT HI	78.9%	44.7%	54.6%	59.7%
FT HP	78.5%	64.2%	71.8%	71.8%
FT HIP	70.9%	69.7%	72.0%	70.8%
NN H	68.1%	65.3%	60.6%	64.6%
NN HI	66.6%	61.4%	72.4%	67.1%
NN HP	77.9%	70.9%	73.8%	73.3%
NN HIP	75.7%	72.1%	74.3%	74.0%

Table 3: (**H**) = Hypothesis, (**P**) = Premise, (**I**) = Image. (**NN**) = Neural Network, sum-of-GloVe model, (**FT**) = Hand-crafted features, linear classifier model. Accuracies of the different styles of classifier per class. The results from this table are discussed in detail in section 7.2.

7 Analysis

7.1 Performance of different classifiers

We were able to reproduce two major baselines that were introduced in (Bowman et al., 2015).

For linear classifiers, none of the models we built was able to combine the sparse, strictly positive textual features with the dense, continuous valued image features. In fact, for the hand-tuned features with an image vector and a linear classifier, the performance of the model decreases slightly, as seen in table 2. This is true of both the hypothesis-only and textual-premise models.

As can be seen in table 3, the accuracies of the neutral class are always worse than for the other classes. This is intuitive, as deciding neutrality can be challenging due to subtle differences between sentences. Additionally, adding image features to the hypothesis baseline significantly boosts the performance of the entailment class, outperforming the hypothesis-only model by over 13%. This suggests that adding the image features helps the classifier determine when a sentence was entailed in an image. However, due to the use of a simple linear classifier, it considerably hurt the performance on the neutral and contradiction classes.

We found that with a small number of training epochs, the neural network model that incorporates image features outperforms our baseline model by a small amount (0.8%, comparable to the 2.2% difference reported between using an LSTM and simply using a sum of GloVe vectors). Our attempts to train the image-added model further, with various amounts of regularization, showed either no additional performance or drastic overfit-

ting, as the neural network would get training accuracy near 90% while testing accuracy actually decreased to 70%. This is in contrast to the published SNLI baselines, which reported much less overfitting with a sum-of-words sentence embedding (79% training, 75% testing) (Bowman et al., 2015).

Furthermore, we implemented the deep attention fusion model from (Cheng et al., 2016), but due to limited training time, we were not able to reproduce its published performance. Likewise, our soft-attention model, presented in section 5.4, received limited training time and was unable to meet or exceed the performance of our simpler top-level image feature vector representation. These networks were fairly sophisticated and required a large amount of hyper-parameter tuning to converge well, and while we got good results, we ran out of time to fully explore their possibilities.

7.2 Shortcomings of the SNLI dataset

Based on the results in section 6.2, we show that even without the premise sentence, which intuitively would be necessary to do correct inference, both baseline models are able to perform far better than random and even outperform some simple baselines from (Bowman et al., 2015). This demonstrates a weakness in the variety of hypothesis statements in the SNLI dataset, as they are predictable based on their content alone. We present a numerical breakdown in table 3 and explore it further in this section.

While new analysis of the SNLI dataset suggests that many of the contradiction examples in SNLI simply include a negation of the premise (Bowman, 2016), we actually found that the hypothesis-only baseline performs well across all classes ($\geq 60\%$ accuracy). In fact, its strongest performance comes in entailment, for both feature and neural network baselines. This suggests possible grammatical bias in the dataset. That is, when people are asked to write entailed sentences, they might use active tone or other predictable features. Because of this result, it follows that creating unbiased textual entailment datasets is very hard, or that natural language has a predictability when users write entailed sentences. It seems possible that the annotators in SNLI lacked variety in tone, or maybe it is possible to detect agreement in large part due to tone alone.

Another interesting result from our multimodal analysis is that the neural network model is able to use the image vector to detect contradiction with about the same power as the textual premise vector. Image vectors allow for a 11.3% increase in detecting entailment accuracy.

This leads to another interesting observation from table 3, which is that adding an additional set of inputs sometimes decreased accuracy. This is certainly true for the simple linear classifier, where adding images decreases the overall accuracy. However, the neural network classifier also sees a drop in accuracy when adding additional data. Adding images to the hypothesis-only model decreases the quality of Neutral predictions by 4%, while adding images to the full textual modal leads to a 2% drop in the accuracy of Entailed predictions. We attribute this degradation in quality to our short training time (10 epochs), and insufficient hyper-parameter tuning for network capacity, learning rate, and regularization as we change the inputs of our models. Since we kept our configuration constant across all examples, we expect the configurations to be sub-optimal for certain cases.

7.3 Value of Image features

However, since the image features are high-dimensional, they contribute millions of additional parameters to the network and may require additional training. Since we held our training settings consistent across all models, there may be an additional performance boost available by using image features with more training time or with additional training data.

Furthermore, it would be possible to fit the CNN features directly to this task. Currently, the features are taken from a classifier that was fit to the ImageNet challenge. While this challenge is very broad and models that were pre-trained on it yielded good results for various other tasks in the past, it might be beneficial to train a CNN end-to-end to yield features that were adapted to this particular task.

Potentially, the 30k image dataset might be too small to properly train a complex model using both image and textual features, since image features alone add an approximate 1M additional parameters. As (Bowman et al., 2015) noted, the SNLI dataset might even be too small to even fully exhaust the means of deep textual models. The complexity of deep multimodal models would proba-

bly require an even larger dataset to fully exhaust their learning capacity.

8 Conclusion and Next Steps

In this paper, we replicated the models from (Bowman et al., 2015) and (Cheng et al., 2016). Further, we demonstrated how image features could be used to augment natural language inference tasks. The results indicate that for simple models such as linear classifiers, images worsen accuracy, while they allow a small performance boost for shallow learning models such as Neural Networks fed with sum-of-words GloVe. For sufficiently complex models, our results indicate that image features cannot contribute additional information and do not seem to improve classifier performance. Due to time and computational limits, however, we were unable to train our most sophisticated model, the LSTMN with deep attention fusion, up to its final accuracy, so that there is a chance that we did not explore the model’s full capabilities.

Additionally, we demonstrated possible structural flaws in the SNLI dataset, where hypotheses are predictable without the premise sentences. One would expect that without knowledge of a premise statement, a hypotheses-only classifier should not achieve a higher accuracy than random guessing would - yet, our hypotheses-only baseline outperforms random guessing by a factor of two, and is only about 10% worse than the initial SNLI baselines. There is interesting future work in exploring what structure in the hypothesis allow the classifier to perform that well. The natural question is whether recognizing textual entailment is predictable based on grammatical content alone and if not, how one can design a Mechanical Turk task such that these structures are not inhibited.

An interesting direction for future work is the basic concept of how to join multiple modalities representing world knowledge. In the case of multimodal NLI, this is captured by the question how one should join image and textual information to allow effective querying into the premise. However, the concept of multiple modalities that represent common world knowledge could be extended to many other fields - for instance, an interesting approach for object localization in images could be to learn common object relationships (such as "children standing on skis") and use these relationships to infer object localization.

References

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.
- 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)*. Association for Computational Linguistics.
- S. R. Bowman, J. Gauthier, A. Rastogi, R. Gupta, C. D. Manning, and C. Potts. 2016. A Fast Unified Model for Parsing and Sentence Understanding. *ArXiv e-prints*, March.
- Samuel R. Bowman. 2016. The stanford nli corpus revisited. <http://nlp.stanford.edu/blog/the-stanford-nli-corpus-revisited/>.
- Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. *CoRR*, abs/1601.06733.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.
- Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.
- Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9. Association for Computational Linguistics.
- Andrew Hickl, John Williams, Jeremy Bensley, Kirk Roberts, Bryan Rink, and Ying Shi. 2006. Recognizing textual entailment with lccs groundhog system. In *Proceedings of the Second PASCAL Challenges Workshop*.
- Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.
- Mat Kelcey. 2016. Snli hacking (in tensorflow). https://github.com/matpalm/snli_nn_tf.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.
- Alice Lai and Julia Hockenmaier. 2014. Illinois-1h: A denotational and distributional approach to semantics. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 329–334, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.
- Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. 2016. Learning natural language inference using bidirectional lstm model and inner-attention. *arXiv preprint arXiv:1605.09090*.
- Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1, ETMTNLP '02*, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Preslav Nakov and Torsten Zesch, editors. 2014. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics and Dublin City University, Dublin, Ireland, August.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.
- Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. *CoRR*, abs/1403.6382.

- Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kociský, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *CoRR*, abs/1509.06664.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.
- Marta Tatu and Dan Moldovan. 2005. A semantic approach to recognizing textual entailment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 371–378. Association for Computational Linguistics.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.
- Shuohang Wang and Jing Jiang. 2015. Learning natural language inference with LSTM. *CoRR*, abs/1512.08849.
- Huijuan Xu and Kate Saenko. 2015. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. *CoRR*, abs/1511.05234.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. 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.