Model	MR	CR	SUBJ	MPQA	SST2	TREC	MRPC	SICK-E	SICK-R	STS14	Macro	Micro
Baseline	77.14	77.24	91.14	87.27	81.11	79.20	73.45/81.80	79.54	0.804	0.54/0.55	80.76	83.08
LSTM SGD	73.44	76.66	85.76	87.71	77.87	75.00	72.75/80.94	83.30	0.861	0.59/0.58	79.06	81.23
LSTM Adam	75.87	79.34	87.94	87.92	80.89	74.20	73.68/81.60	83.86	0.857	0.62/0.60	80.46	82.81
BiLSTM SGD	73.15	77.61	89.39	86.69	77.54	86.00	74.26/81.39	85.75	0.872	0.61/0.59	81.30	82.27
BiLSTM Adam	75.96	79.12	89.76	87.86	81.22	85.20	74.84/82.24	82.77	0.870	0.59/0.58	82.09	83.27
BLM SGD	78.08	81.19	**92.15**	88.68	82.37	**90.40**	74.03/81.95	85.51	0.886	0.68/0.66	84.05	85.09
BLM SGD WD	77.70	81.62	91.98	88.51	81.55	87.60	74.61/81.99	86.54	0.887	0.68/0.65	83.76	85.03
BLM SGD DP	78.04	81.99	92.09	88.63	81.44	89.00	**75.83/83.22**	**86.77**	**0.889**	0.69/0.66	**84.22**	**85.28**
BLM Adam	77.72	81.27	91.45	88.94	**83.47**	86.60	74.26/82.49	85.49	0.885	**0.70/0.68**	83.65	84.92
Conneau et al.	**79.9**	**84.6**	**92.1**	**89.8**	83.3	88.7	75.1/82.3	86.3	0.885	0.68/0.65	83.7	85.2

TABLE I: Results on SentEval transfer tasks (test dataset). Best performance across models shown in bold.

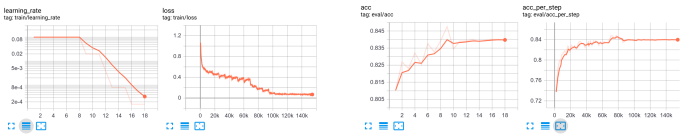
I. General results

Github repository for the code can be found here.

Model	Train	Val	$\text{Test}_{\text{micro}}$	$\text{Test}_{\text{macro}}$
Baseline	66.49%	67.79%	67.65%	67.54%
LSTM SGD	89.70%	82.34%	81.39%	81.35%
LSTM Adam	99.17%	82.89%	82.78%	82.75%
BiLSTM SGD	97.99%	81.51%	81.25%	81.19%
BiLSTM Adam	99.38%	82.99%	82.64%	82.55%
BL Max SGD	99.06%	84.23%	84.15%	84.09%
BL Max SGD WD	94.47%	84.76%	84.85%	84.80%
BL Max SGD DP	96.91%	84.57%	**85.41%**	**85.35%**
BL Max Adam	99.98%	**85.00%**	84.99%	84.93%
Conneau et al.	–	**85.0%**	84.5%	–

TABLE II: Result on SNLI dataset. Abbreviations: DP - dropout 0.1 in MLP, WD - weight decay 10^{-4} instead of 10^{-5}.

All models are significantly better than Baseline (sign-test), but no difference among LSTM/BiLSTM. BiLSTM Max SGD DP and Adam are significantly better than all LSTM/BiLSTM, but no difference among BiLSTM Max.



(a) Training loss curve (b) Evaluation accuracy

Fig. 1: Tensorboard plots showing performance during training. The x-axis is in (a) steps/batches and (b) epochs.

II. SNLI Bias

SNLI dataset is highly biased due to the way it was created.
Easy: `P`: A man poses in front of an ad. `H1`: A man poses in front of an ad. `L1`: entailment, `H2`: A man poses in front of an ad for beer. `L2`: neutral
Hard: `H3`: A man walks by an ad. `L3`: contradiction
Easy: `P`: A young boy in green shorts balances on a pipe above a river. `H1`: Nobody is balancing. `L1`: contra., `H2`: A person balancing. `L2`: entail.
Hard: `H3`: A clever person balancing. `L3`: neutral

Model	Test easy	Test hard	Test combined
Baseline	79.98%	42.84%	67.65%
LSTM Adam	91.68%	64.86%	82.78%
BiLSTM Adam	91.56%	64.70%	82.64%
BLM Adam	92.63%	69.61%	84.99%
BLM SGD DP	**93.22%**	**69.70%**	**85.41%**

TABLE III: Test evaluation on subset easy/hard

III. Transfer tasks

See Table I. For Image Caption Retrieval, see Table IV.

Model	Caption retrieval				Image retrieval			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Baseline	34.12	66.24	79.16	3.0	26.84	61.12	77.34	3.8
LSTM SGD	29.62	62.70	77.28	3.2	25.08	58.86	75.40	4.0
LSTM Adam	33.92	67.70	81.12	3.0	27.42	62.45	78.46	3.2
BL SGD	32.82	67.56	81.14	2.8	27.83	62.64	78.14	3.2
BL Adam	37.14	70.54	82.80	2.2	29.04	64.63	80.07	3.0
BLM Adam	43.18	76.40	**88.38**	**2.0**	33.82	70.13	83.64	3.0
BLM SGD DP	**43.66**	**76.78**	88.06	**2.0**	**34.18**	**69.72**	**83.79**	**2.8**
Conneau et al.	42.6	75.3	87.3	**2**	33.9	**69.7**	**83.8**	**3**

TABLE IV: Image Caption Retrieval task

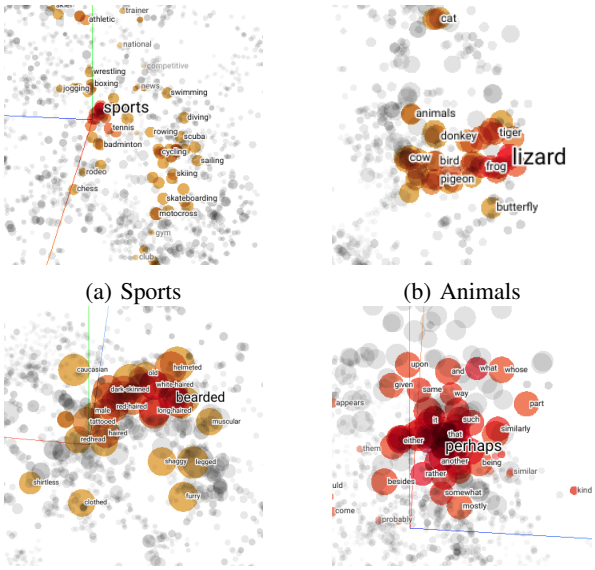
Fig. 2: 2D tSNE on BoW embedding of first two words for the TREC dataset. Blue: NUM, Green: ENTY, Light blue: DESC, Red: LOC, Pink: HUM.

IV. Word importance

For the Bi-LSTM with max pooling, we can measure the importance of a word by the number of times its hidden state is selected by the max pooling operation. The analysis is performed over the whole SNLI dataset.

Most important words: `skydiving, napping, snowing, barbecuing, wakeboarding, nobody, gardening, breakdancing, humans, golfing, sandcastle, awake`
Least important words: `",", ";", as, which, where, that, be, such, so, one, but, "``", and, perhaps, "</s>"`

Most important words per channel/feature:
Feature 19: `breakdancing, sawing, carving, harvesting, skateboarding, barbecuing, wakeboarding, kayaking`
Feature 29: `walking, vegetables, pasta, toast, awake, meal, tomatoes, spaghetti, dusk, walks`



(a) Sports (b) Animals

(c) Human properties (d) Low importance words

Fig. 3: 3D tSNE visualization of feature max pooling. Every word is represented by the channel-wise *importance*, and the selected words show the closest neighbor of the largest word.