Appendix to accompany "Dissecting contextual word embeddings: Architecture and Representation"

A biLM hyperparameters

For consistency and a fair comparison, all biLMs use a 512 dimensional representation in each direction at each layer, providing a 1024 dimensional contextual representation at each layer. All models use the same character based word embedding layer $\mathbf{x}_k$, with the exception of the 4-layer LSTM as described below. All systems use residual connections between the contextual layers (He et al., 2016).

LSTM Hyperparameters for the 4-layer LSTM closely follow those from the 2-layer ELMo model in Peters et al. (2018). It has four layers, with each direction in each layer having a 4096 dimension hidden state and a 512 dimensional projection. To reduce training time, the character based word representation is simplified from the other models. It uses the same 2048 character n-gram CNN filters as the other models, but moves the projection layer from 2048 to 512 dimensions after the convolutions and before the two highway layers.

Transformer The Transformer biLM uses six layers, each with eight attention heads and a 2048 hidden dimension for the feed forward layers. 10% dropout was used after the word embedding layer $\mathbf{x}_k$, multi-headed attention, the hidden layers in the feed forward layers and before the residual connections. Optimization used batches of 12,000 tokens split across 4 GPUs with, using the learning rate schedule from Vaswani et al. (2017) with 2,000 warm up steps. The final model weights were averaged over 10 consecutive checkpoints.

Gated CNN The Gated CNN has 16 layers of $[4, 512]$ residual blocks with 5% dropout between each block. Optimization used Adagrad with linearly increasing learning rate from 0 to 0.4 over the first 10,000 batches. The batch size was 7,500 split across 4 GPUs. Gradients were clipped if their norm exceeded 5.0. The final model weights were averaged over 10 consecutive checkpoints.

B Task model hyperparameters

MultiNLI Our implementation of the ESIM model uses 300 dimensions for all LSTMs and all feed forward layers. For regularization we used 50% dropout at the input to each LSTM and after each feed forward layer. Optimization used Adam with learning rate 0.0004 and batch size of 32 sentence pairs.

Semantic Role Labeling The SRL model uses the reimplementation of He et al. (2017) from Gardner et al. (2018). Word representations are concatenated with a 100 dimensional binary predicate representation, specifying the location of the predicate for the given frame. This is passed through an 8 layer bidirectional LSTM, where the layers alternate between forward and backward directions. Highway connections and variational dropout are used between every LSTM layer. Models are with a batch size of 80 sentences using Adadelta (Zeiler, 2012) with an initial learning rate of 1.0 and rho 0.95.

Constituency Parsing The constituency Parser is a reimplementation of Joshi et al. (2018), available in AllenNLP (Gardner et al., 2018). Word representations from the various biLMs models are passed through a two layer bidirectional LSTM with hidden size 250 and 0.2 dropout. Then, the span representations are passed through a feedforward layer (dropout 0.1, hidden size 250) with a relu non-linearity before classification. We use a batch size of 64 and gradients are normalized to have a global norm ≤ 5.0. Optmization uses Adadelta with initial learning rate of 1.0 and rho 0.95.

NER The NER model concatenates 128 character CNN filters of width 3 characters to the pretrained word representations. It uses two LSTM layers with hidden size 200 with 50% dropout at input and output. The final layer is a CRF, with constrained decoding to enforce valid tag sequences. We employ early stopping on the development set and report average F_1 across five random seeds.

C Contextual similarities

Figures 5, 6, and 7 show contextual similarities similar to Figure 1 for all layers from the 4-layer LSTM, the Transformer and gated CNN biLMs.

D Layer diagnostics

Tables 4, 5 and 6 list full results corresponding to the top three rows in Fig. 3.

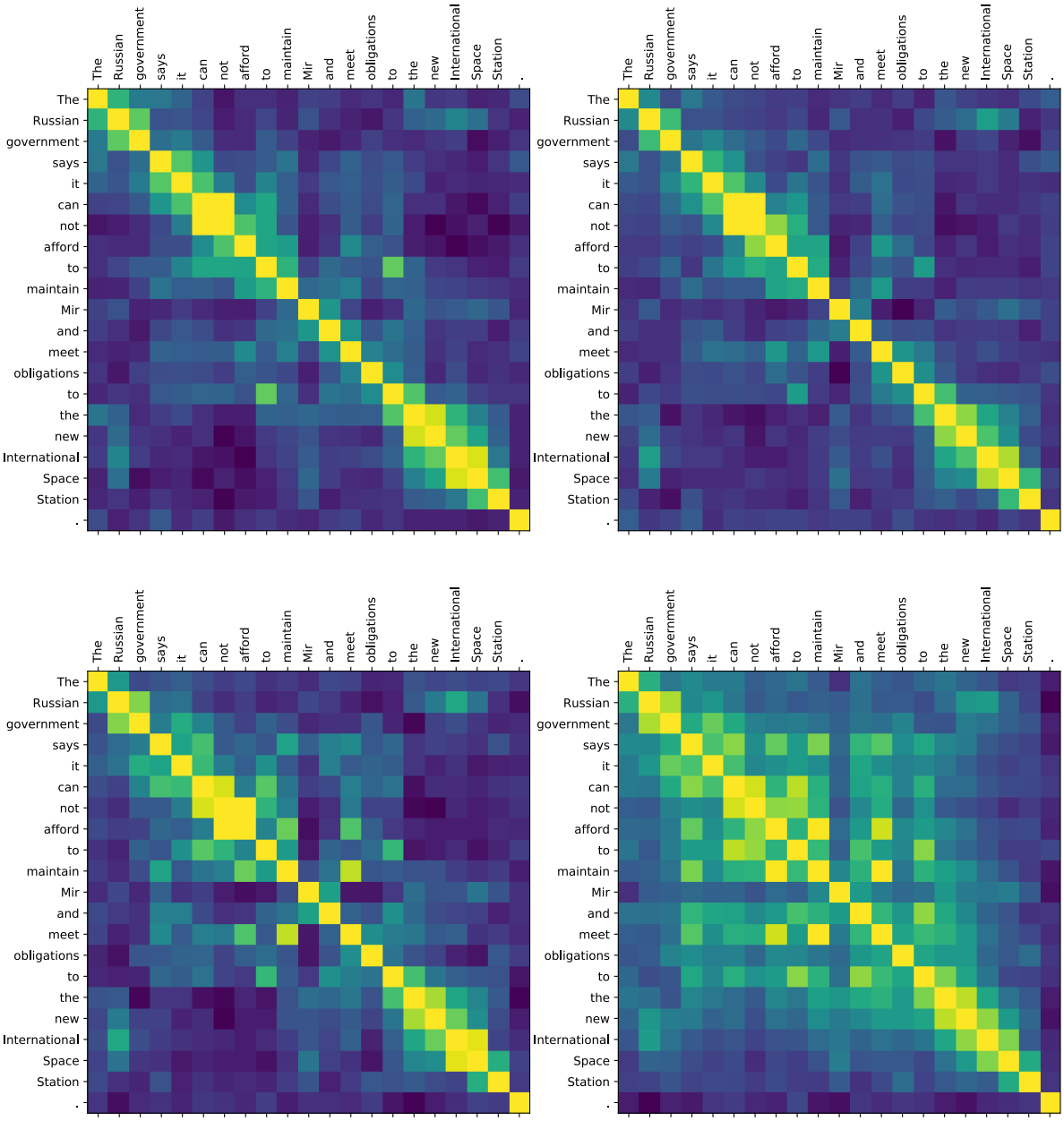

Figure 5: Visualization of contextual similarities from the 4-layer LSTM biLM. The first layer is at top left and last layer at bottom right, with the layer indices increasing from left to right and top to bottom in the image.

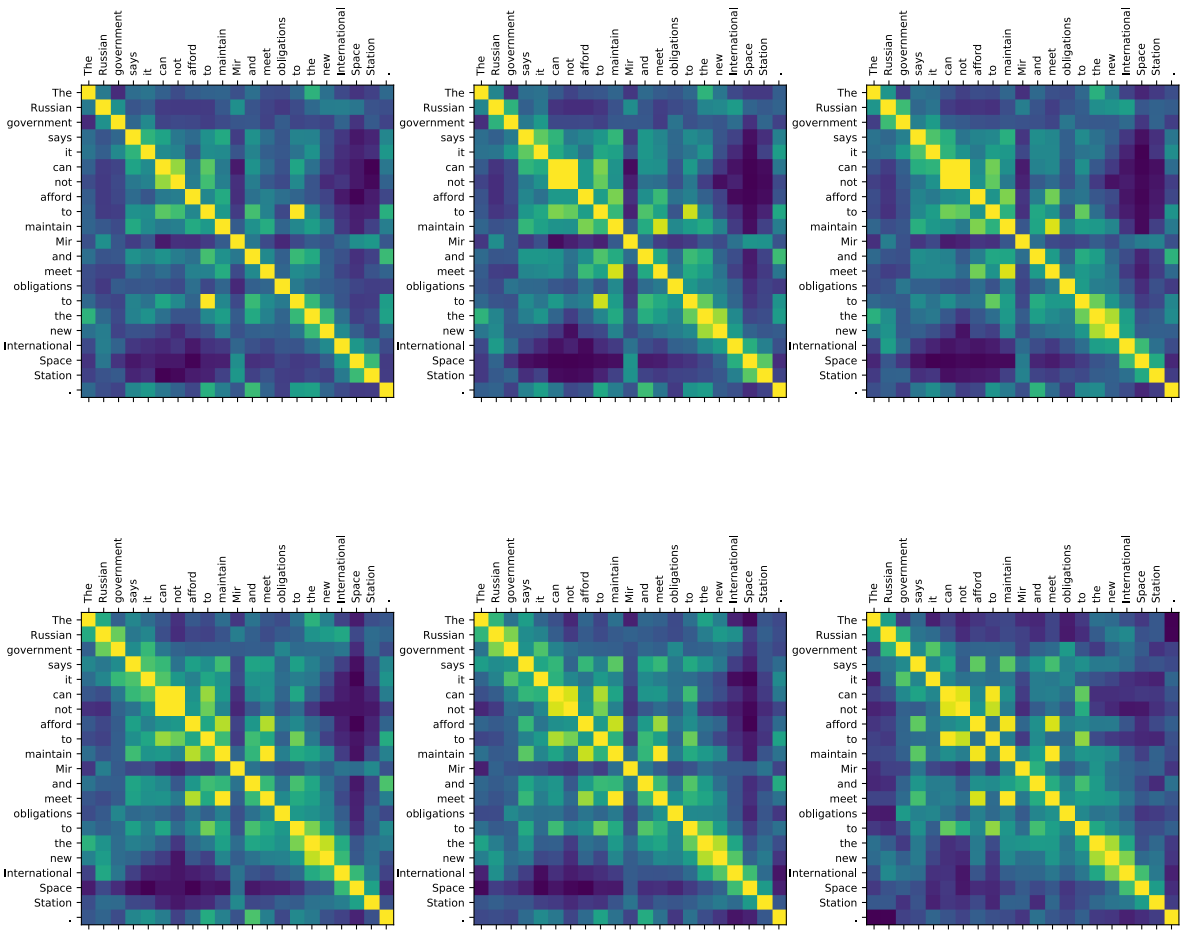

Figure 6: Visualization of contextual similarities from the Transformer biLM. The first layer is at top left and last layer at bottom right, with the layer indices increasing from left to right and top to bottom in the image.

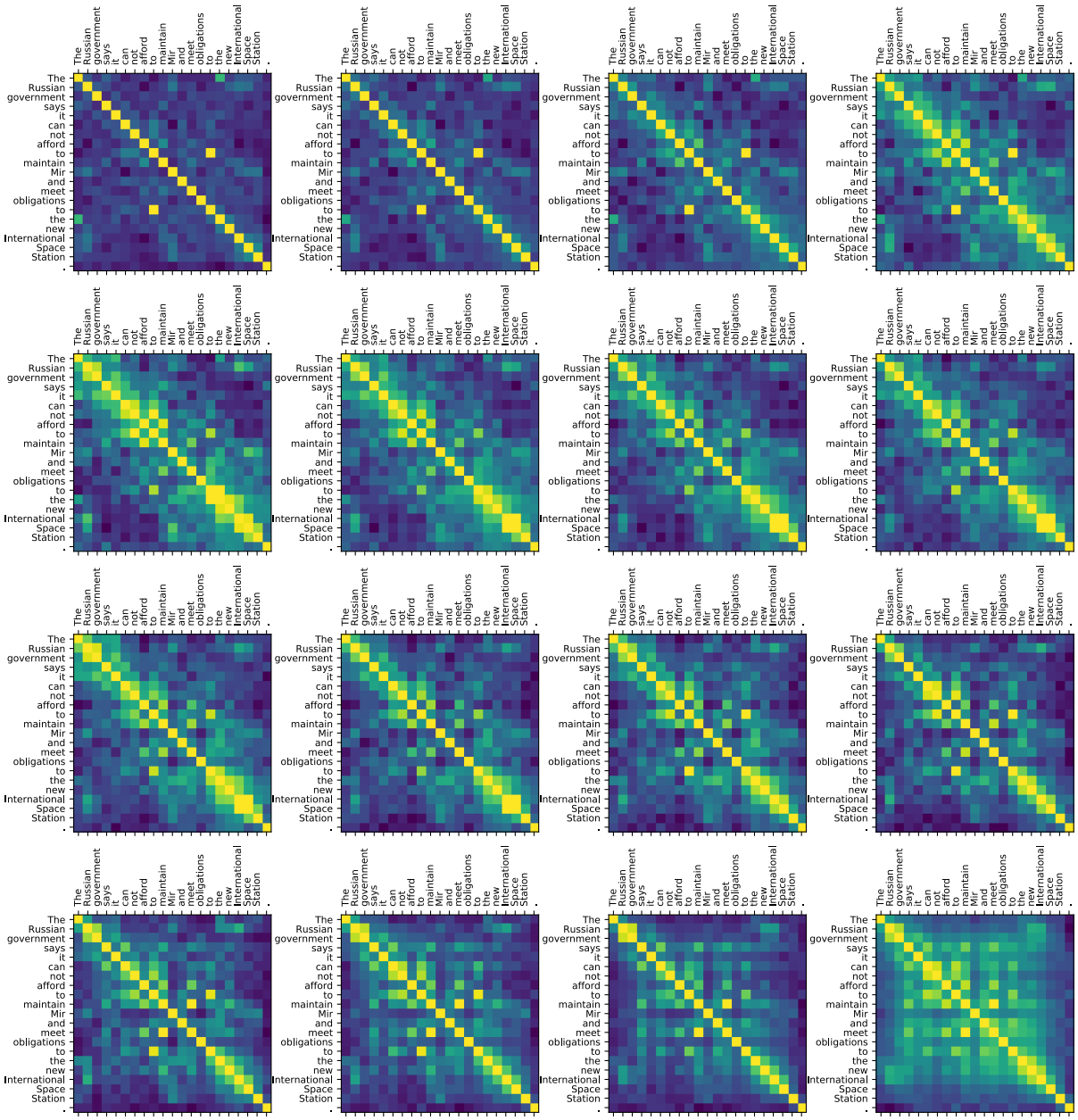

Figure 7: Visualization of contextual similarities from the gated CNN biLM. The first layer is at top left and last layer at bottom right, with the layer indices increasing from left to right and top to bottom in the image.

Layer	Accuracy
Elmo - 4 Layer	
Layer 1	46.5
Layer 2	46.0
Layer 3	48.8
Layer 4	**52.0**
Transformer	
Layer 1	47.8
Layer 2	52.9
Layer 3	55.7
Layer 4	**56.7**
Layer 5	54.7
Layer 6	51.5
Gated CNN	
Layer 1	41.5
Layer 2	44.2
Layer 3	44.2
Layer 4	45.7
Layer 5	45.9
Layer 6	48.5
Layer 7	47.4
Layer 8	49.6
Layer 9	51.7
Layer 10	47.8
Layer 11	52.1
Layer 12	**52.4**
Layer 13	50.3
Layer 14	51.3
Layer 15	52.0
Layer 16	51.6

Table 4: Unsupervised pronominal accuracies using the CoNLL 2012 development set.

Layer	Accuracy
GloVe Only	88.61
Elmo - 4 Layer	
Layer 1	**97.36**
Layer 2	97.16
Layer 3	96.90
Layer 4	96.58
Weighted Layers	97.22
Transformer	
Layer 1	97.30
Layer 2	97.35
Layer 3	97.25
Layer 4	97.15
Layer 5	96.90
Layer 6	96.82
Weighted Layers	**97.48**
Gated CNN	
Layer 1	97.09
Layer 2	97.16
Layer 3	97.19
Layer 4	97.16
Layer 5	97.11
Layer 6	97.09
Layer 7	97.08
Layer 8	97.01
Layer 9	97.00
Layer 10	96.97
Layer 11	96.96
Layer 12	96.97
Layer 13	96.85
Layer 14	96.80
Layer 15	96.64
Layer 16	96.43
Weighted Layers	**97.26**

Table 5: POS tagging accuracies for the linear models on the PTB dev set.

Layer	F_1	Precision	Recall
GloVe Only	18.1	11.2	45.9
LSTM - 4 Layer			
Layer 1	80.8	87.0	74.1
Layer 2	76.8	83.7	71.4
Layer 3	75.8	84.6	68.7
Layer 4	76.5	81.6	72.1
Weighted Layers	**80.9**	**87.9**	**75.4**
Transformer			
Layer 1	75.8	80.4	71.7
Layer 2	77.3	82.6	72.6
Layer 3	78.5	82.5	75.0
Layer 4	79.2	80.5	77.9
Layer 5	77.7	78.3	77.0
Layer 6	76.1	77.3	75.0
Weighted Layers	**82.8**	**87.5**	**78.6**
Gated CNN			
Layer 1	67.4	79.5	58.5
Layer 2	71.3	81.8	63.1
Layer 3	73.3	83.8	65.1
Layer 4	75.0	84.6	67.3
Layer 5	76.5	85.3	69.3
Layer 6	77.4	85.6	70.7
Layer 7	77.6	85.9	70.7
Layer 8	78.0	86.9	71.1
Layer 9	78.6	85.0	**73.3**
Layer 10	78.5	85.9	72.3
Layer 11	78.5	84.5	**73.3**
Layer 12	77.4	85.4	70.8
Layer 13	76.7	84.7	70.1
Layer 14	75.9	83.3	69.9
Layer 15	75.5	82.8	69.4
Layer 16	75.7	83.0	69.6
Weighted Layers	**78.6**	**86.4**	72.0

Table 6: Labeled Bracketing F_1, Precision and Recall for the linear parsing models on the PTB dev set.