## A Experiment Details

**Dataset statistics** Table 2 shows the sizes of the datasets experimented with.

| | Training | Dev. | Test |
|---|---|---|---|
| **kitchen** | 3,298 | 822 | 4,118 |
| **dvd** | 14,066 | 3,514 | 17,578 |
| **books** | 20,000 | 5,000 | 25,000 |
| **original_mix** | 20,000 | 5,000 | 25,000 |

Table 2: Text classification dataset sizes. Each dataset follows the same training/dev./test split ratio as the original mix.

**Preprocessing** As preprocessing for the data for each individual category, we tokenize using the NLTK word tokenizer. We removed reviews with text shorter than 5 tokens.

We binarize the review score using the standard procedure, assigning 1- and 2-star reviews as negative, and 4- and 5-star reviews as positive (discarding 3-star reviews). Then, if there were more than 25,000 negative reviews, we downsample to 25,000 (otherwise we keep them all), and then downsample the positive reviews to be the same number as negative, to have a balanced dataset. We match the train, development, and test set proportions of 4:1:5 from the original mixture.

We generate the BERT embeddings using the sum of the last four hidden layers of the large uncased BERT model, so our embedding size is 1024. Summing the last four layers was the best performing approach in the ablation of Devlin et al. (2019) that had fewer than 4096 embedding size (which was too large to fit in memory). We embed each sentence individually (there can be multiple sentences within one example).

**Implementation details** For GloVe, we train rational models with 24 5-state WFSAs, each corresponding to a 4-gram soft-pattern (Fig. 2). For BERT, we train models with 12 WFSAs.[13]

**Experiments** For each model (regularized or baseline), we run random search to select our hyperparameters (evaluating 20 uniformly sampled hyperparameter configurations). For the hyperparameter configuration that leads to the best development result, we train the model again 5 times with different random seeds, and report the mean and standard deviation of the models' test performance.

**Parameters** The models are trained with Adam (Kingma and Ba, 2015). During training with group lasso we turn off the learning rate schedule (so the learning rate stays fixed), similarly to Gordon et al. (2018). This leads to improved stability in the learned structure for a given hyperparameter assignment.

Following Peng et al. (2018) we sample 20 hyperparameters uniformly, for which we train and evaluate our models. Hyperparameter ranges are presented in Table 4. For the BERT experiments, we reduced both the upper and lower bound on the learning rate by two orders of magnitude.

**Regularization strength search** We searched for model structures that were regularized down to close to 20, 40, 60, or 80 transitions (10, 20, 30, and 40 for BERT experiments). For a particular goal size, we uniformly sample 20 hyperparameter assignments from the ranges in Table 4, then sorted the samples by increasing learning rate. For each hyperparameter assignment, we trained a model with the current regularization strength. If the resulting learned structure was too large (small), we doubled (halved) the regularization strength, repeating until we were within 10 transitions of our goal (5 for BERT experiments).[14] Finally, we finetuned the appropriately-sized learned structure by continuing training without the regularizer, and computed the result on the development set. For the best model on the development set, we retrained (first with the regularizer to learn a structure, then finetuned) five times, and plot the mean and variance of the test accuracy and learned structure size.

## B Visualization

Table 3 shows the same visualization shown in §4 for another sparse rational RNN containing only four WFSAs and 11 main-path transitions, trained with BERT embeddings on **kitchen**. It also shows a few clear patterns (e.g., Patt. 2). Interpretation here is more challenging though, as contextual embeddings make every token embedding depend

[13]The BERT embedding dimension is significantly larger than GloVe (1024 compared to 300), so we used a smaller number of WFSAs. As our results show, the BERT models still substantially outperform the GloVe ones.

[14]If the regularization strength became larger than $10^2$ or smaller than $10^{-9}$, we threw out the hyperparameter assignment and resampled (this happened when, e.g., the learning rate was too small for any of the weights to actually make it to zero).

| | | transition$_1$ | transition$_2$ | transition$_3$ |
|---|---|---|---|---|
| Patt. 1 | Top | are | perfect | ...$_{SL}$ [CLS] |
| | | definitely | recommend | ...$_{SL}$ [CLS] |
| | | excellent | product | ...$_{SL}$ [CLS] |
| | | highly | recommend | ...$_{SL}$ [CLS] |
| | Bottom | not | ...$_{SL}$ [SEP] | ...$_{SL}$ [CLS] |
| | | very | disappointing | !$_{SL}$ [SEP]$_{SL}$ [CLS] |
| | | was | defective | ...$_{SL}$ had |
| | | would | not | ...$_{SL}$ [CLS] |
| Patt. 2 | Top | [CLS] | mine | broke |
| | | [CLS] | it | ...$_{SL}$ heat |
| | | [CLS] | thus | it |
| | | [CLS] | it$_{SL}$ does | it$_{SL}$ heat |
| | Bottom | [CLS] | perfect | ...$_{SL}$ cold |
| | | [CLS] | sturdy | ...$_{SL}$ cooks |
| | | [CLS] | evenly | ,$_{SL}$ withstand$_{SL}$ heat |
| | | [CLS] | it | is |
| Patt. 3 | Top | ‘ | pops | ’$_{SL}$ ’$_{SL}$ escape |
| | | ‘ | gave | out |
| | | that | had | escaped |
| | | ‘ | non | - |
| | Bottom | simply | does | not |
| | | [CLS] | useless | equipment$_{SL}$ ! |
| | | unit | would | not |
| | | [CLS] | poor | to$_{SL}$ no |
| Patt. 4 | Top | [CLS] | after | |
| | | [CLS] | our | |
| | | mysteriously | jammed | |
| | | mysteriously | jammed | |
| | Bottom | [CLS] | i | |
| | | [CLS] | i | |
| | | [CLS] | i | |
| | | [CLS] | we | |

Table 3: Visualization of a sparse rational RNN containing 4 WFSAs only, trained on **kitchen** using BERT.

on the entire context.[15] A particular example of this is the excessive use of the start token ([CLS]), whose contextual embedding has been shown to capture the sentiment information at the sentence level (Devlin et al., 2019).

**Regularization strength recommendation** If a practitioner wishes to learn a single small model, we recommend they start with $\lambda$ such that the loss $\mathcal{L}(\mathbf{w})$ and the regularization term are equal at initialization (before training). We found that having equal contribution led to eliminating approximately half of the states, though this varies with data set size, learning rate, and gradient clipping, among other variables.

[15]Indeed, contextual embeddings raise problems for interpretation methods that work by targeting individual words, e.g., attention (Bahdanau et al., 2015), as these embeddings also depend on other words. Interpretation methods for contextual embeddings are an exciting direction for future work.

| **Type** | **Range** |
|---|---|
| Learning rate | $[7 * 10^{-3}, 0.5]$ |
| Vertical dropout | $[0, 0.5]$ |
| Recurrent dropout | $[0, 0.5]$ |
| Embedding dropout | $[0, 0.5]$ |
| $\ell_2$ regularization | $[0, 0.5]$ |
| Weight decay | $[10^{-5}, 10^{-7}]$ |

Table 4: Hyperparameter ranges considered in our experiments.