Disentangling Representations of Text by Masking Transformers

Xiongyi Zhang
Northeastern University
`zhang.xio@northeastern.edu`

Jan-Willem van de Meent
University of Amsterdam
Northeastern University
`j.vandemeent@northeastern.edu`

Byron C. Wallace
Northeastern University
`b.wallace@northeastern.edu`

Abstract

Representations from large pretrained models such as BERT encode a range of features into monolithic vectors, affording strong predictive accuracy across a range of downstream tasks. In this paper we explore whether it is possible to learn *disentangled representations* by identifying existing subnetworks within pretrained models that encode distinct, complementary aspects. Concretely, we learn binary masks over transformer weights or hidden units to uncover subsets of features that correlate with a specific factor of variation; this eliminates the need to train a disentangled model from scratch for a particular task. We evaluate this method with respect to its ability to disentangle representations of *sentiment* from *genre* in movie reviews, *toxicity* from *dialect* in Tweets, and *syntax* from *semantics*. By combining masking with magnitude pruning we find that we can identify sparse subnetworks within BERT that strongly encode particular aspects (e.g., semantics) while only weakly encoding others (e.g., syntax). Moreover, despite *only* learning masks, disentanglement-via-masking performs as well as — and often better than — previously proposed methods based on variational autoencoders and adversarial training.

1 Introduction and Motivation

Large pretrained models such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), and XL-Net (Yang et al., 2019) have come to dominate modern NLP. Such models rely on self-supervision over large datasets to learn general-purpose representations of text that achieve strong predictive performance across a spectrum of downstream tasks (Liu et al., 2019). A downside of such learned representations is that it is not obvious what information they encode, which hinders model robustness and interpretability. The opacity of embeddings produced by models such as BERT has motivated NLP research on designing probing tasks as a means of uncovering the properties of input texts that are
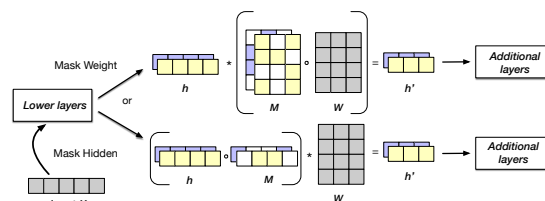

Figure 1: Masking weights and hidden activations in BERT. We show a linear layer with weights W, inputs h, and outputs h'. We learn a mask for each disentangled factor, which is either applied to the weights W or to intermediate representations h.

encoded in learned representations (Rogers et al., 2020; Linzen et al., 2019; Tenney et al., 2019).

In this paper we investigate whether we can uncover *disentangled representations* from pretrained models. That is, rather than mapping inputs onto a single vector that captures arbitrary combinations of features, our aim is to extract a representation that factorizes into distinct, complementary properties of inputs. Explicitly factorizing representations aids interpretability, in the sense that it becomes more straightforward to determine which factors of variation inform predictions in downstream tasks.

A general motivation for learning disentangled representations is to try and minimize — or at least expose — model reliance on spurious correlations, i.e., relationships between (potentially sensitive) attributes and labels that exist in the training data but which are not causally linked (Kaushik et al., 2020). This is particularly important for large pretrained models like BERT, as we do not know what the representations produced by such models encode. Here, learning disentangled representations may facilitate increased robustness under distributional shifts by capturing a notion of invariance: If syntactic changes do not affect the representation of semantic features (and vice versa) then we can hope to learn models that are less sensitive to any incidental correlations between these factors.

As one example that we explore in this paper, consider the task of identifying Tweets that contain hate speech (Founta et al., 2018). Recent work shows that models trained over Tweets annotated on a toxicity scale exhibit a racial bias: They have a tendency to over-predict that Tweets written by users who self-identify as Black are "toxic", owing to the use of African American Vernacular English (AAVE; Sap et al. 2019). In principal, disentangled representations would allow us to isolate relevant signal from irrelevant or spurious factors (such as, in this case, the particular English dialect used), which might in turn reveal and allow us to mitigate unwanted system biases, and increase robustness.

To date, most research on disentangled representations has focused on applications in computer vision (Locatello et al., 2019b; Kulkarni et al., 2015; Chen et al., 2016; Higgins et al., 2017), where there exist comparatively clear independent factors of variation such as size, position, and orientation, which have physical grounding and can be formalized in terms of actions of symmetry subgroups (Higgins et al., 2018). A challenge in learning disentangled representations of text is that it is less clear which factors of variation should admit invariance. Still, we may hope to disentangle particular properties for certain applications — e.g., protected demographic information (Elazar and Goldberg, 2018) — and there are general properties of language that we might hope to disentangle, e.g., syntax and semantics (Chen et al., 2019).

2 Methods

We are interested in learning a disentangled representation that maps inputs x (text) onto vectors $z^{(a)}$ and $z^{(b)}$ that encode two distinct factors of variation. To do so, we will learn two sets of masks $M^{(a)}$ and $M^{(b)}$ that can be applied to either the weights or the intermediate representations in a pretrained model (in our case, BERT). We estimate only the mask parameters and do not finetune the weights of the pretrained model.

To learn $M^{(a)}$ and $M^{(b)}$, we assume access to triplets (x_0, x_1, x_2) in which x_0 and x_1 are similar with respect to aspect a but dissimilar with respect to aspect b, whereas x_0 and x_2 are similar with respect to aspect b but dissimilar with respect to aspect a. In some of our experiments (e.g., when disentangling sentiment from genre in movie reviews) we further assume that we have access to class labels $y^{(a)} \in \{0, 1\}$ and $y^{(b)} \in \{0, 1\}$ for as-

pects of interest. In such cases, we build triplets using these labels, defining (x_0, x_1, x_2) such that $y_0^{(a)} = y_1^{(a)} \neq y_2^{(a)}$ and $y_0^{(b)} = y_2^{(b)} \neq y_1^{(b)}$.

2.1 Masking Weights and Hidden Activations

Figure 1 illustrates the two forms of masking that we consider in our approach (we depict only a single linear layer of the model). Here $h = (h^{(a)}, h^{(b)})$ are input activations, W are the weights in the pretrained model,[1] and $h' = (h'^{(a)}, h'^{(b)})$ are output activations. We augment each layer of the original network with two (binary) masks $M = (M^{(a)}, M^{(b)})$, applied in one of two ways:

1. Masking Weights Here masks $M^{(a)}$ and $M^{(b)}$ have the same shape as weights W, and outputs are computed using the masked weights tensor

$$h' = h \cdot (W \circ M). \qquad (1)$$

2. Masking Hidden Activations In this case masks $M^{(a)}$ and $M^{(b)}$ have the same shape as the intermediate (hidden) activations $h^{(a)}$ and $h^{(b)}$. Output activations are computed by applying the original weights W to masked inputs

$$h' = (h \circ M) \cdot W. \qquad (2)$$

In both methods, we follow (Zhao et al., 2020) and only mask the last several layers of BERT, leaving bottom layers unchanged.[2]

2.2 Triplet Loss

To learn masks, we assume that we have access to supervision in the form of triplets, as introduced above. Passing (x_0, x_1, x_2) through our model yields two representations for each instance: $(z_0^{(a)}, z_0^{(b)}), (z_1^{(a)}, z_1^{(b)}), (z_2^{(a)}, z_2^{(b)})$, for which we define the losses

$$\mathcal{L}_{\text{trp}}^{(a)} = \max\left(\|z_0^{(a)} - z_1^{(a)}\| - \|z_0^{(a)} - z_2^{(a)}\| + \alpha, 0\right),$$

$$\mathcal{L}_{\text{trp}}^{(b)} = \max\left(\|z_0^{(b)} - z_2^{(b)}\| - \|z_0^{(b)} - z_1^{(b)}\| + \alpha, 0\right),$$

$$\mathcal{L}_{\text{trp}} = \frac{1}{2}\left(\mathcal{L}_{\text{trp}}^{(a)} + \mathcal{L}_{\text{trp}}^{(b)}\right). \qquad (3)$$

Here α is a hyperparameter specifying a margin for the loss, which we set to $\alpha = 2$ in all experiments.

---

[1] We omit the bias term, which we do not mask.
[2] We mask the last six layers, which seems to work well.

2.3 Supervised Loss

In some settings we may have access to more direct forms of supervision. For example, when learning representations for the genre and sentiment in a movie review, we have explicit class labels $y^{(a)}$ and $y^{(b)}$ for each aspect. To exploit such supervision when available, we add classification layers $C^{(a)}$ and $C^{(b)}$ and define classification losses

$$\mathcal{L}_{\text{cls}}^{(a)} = \text{CrossEntropy}\Big(C^{(a)}(z^{(a)}), y^{(a)}\Big), \quad (4)$$

$$\mathcal{L}_{\text{cls}}^{(b)} = \text{CrossEntropy}\Big(C^{(b)}(z^{(b)}), y^{(b)}\Big), \quad (5)$$

$$\mathcal{L}_{\text{cls}} = \frac{1}{2}\Big(\mathcal{L}_{\text{cls}}^{(a)} + \mathcal{L}_{\text{cls}}^{(b)}\Big). \quad (6)$$

2.4 Disentanglement Loss

To ensure that the two aspect representations are distinct, we encourage the masks to overlap as little as possible. To achieve this we add a term in the loss for each layer $l \in L$

$$\mathcal{L}_{\text{ovl}} = \frac{1}{|L|} \sum_{l \in L} \sum_{i,j} \mathbb{1}_{(M_{i,j}^{(a)} + M_{i,j}^{(b)} > 1)}. \quad (7)$$

2.5 Binarization and Gradient Estimation

The final loss of our model is

$$\mathcal{L} = \lambda_{\text{trp}} \cdot \mathcal{L}_{\text{trp}} + \lambda_{\text{ovl}} \cdot \mathcal{L}_{\text{ovl}} \ (+\lambda_{\text{cls}} \cdot \mathcal{L}_{\text{cls}}). \quad (8)$$

We parenthetically denote the classification loss, which we only include when labels are available. We minimize this loss to estimate M (and classifier parameters), keeping the pretrained BERT weights fixed. Because the loss is not differentiable with respect to a binary mask, we learn continuous masks M that are binarized during the forward pass by applying a threshold τ, a global hyperparameter,

$$M_{ij}^* = \begin{cases} 1 & \text{if } M_{ij} \geq \tau \\ 0 & \text{if } M_{ij} < \tau \end{cases}. \quad (9)$$

We then use a *straight-through* estimator (Hinton et al., 2012; Bengio et al., 2013) to approximate the derivative, which is to say that we evaluate the derivative of the loss with respect to the continuous mask M at the binarized values $M = M^*$,

$$M = M - \eta \left.\frac{\partial L}{\partial M}\right|_{M=M^*}. \quad (10)$$

IMDB

	Sentiment / Genre	Drama	Horror
Original Dataset	Positive	41.2	7.4
	Negative	20.0	31.4
Our Train (Correlated)	Positive	42.5	7.5
	Negative	7.5	42.5
Our Test (Uncorrelated)	Positive	25.0	25.0
	Negative	25.0	25.0

Twitter

	Toxicity / Race	Black	White
Original Dataset	Toxic	15.4	16.9
	Non-toxic	20.8	46.9
Train (Correlated)	Toxic	42.5	7.5
	Non-toxic	7.5	42.5
Test (Uncorrelated)	Toxic	25.0	25.0
	Non-toxic	25.0	25.0

Table 1: Percentage of each class in the original dataset, and in the two subsets we sampled, one correlated training set and one uncorrelated test set. We train models on the former and test on the latter. This is meant to assess the robustness of models to shifts in spurious correlations that might exist in training data.

3 Experiments

We conduct a series of experiments to evaluate the degree to which the proposed masking strategy achieves disentanglement, as compared to existing methods for disentanglement in NLP. As a first illustrative example, we consider a corpus of movie reviews, in which sentiment is correlated with film genre (3.1). We treat this as a proxy for a spurious correlation, and evaluate the robustness of the models to shifts in conditional probabilities of one attribute (sentiment) given another (genre). We then consider a more consequential example: Hate speech classification on Twitter (3.2). Prior work (Sap et al., 2019) has shown that models exploit a spurious correlation between "toxicity" and African American Vernacular English (AAVE); we aim to explicitly disentangle these factors in service of fairness. We evaluate whether the model is able to achieve *equalized odds*, a commonly used fairness metric. Finally, following prior work, we investigate disentangling semantics from syntax (insofar as this is possible) in Section 3.3.

3.1 Disentangling Sentiment From Genre

Experimental Setup In this experiment we assume a setting in which each data point x has both a 'main' label y and a secondary (possibly sensi-
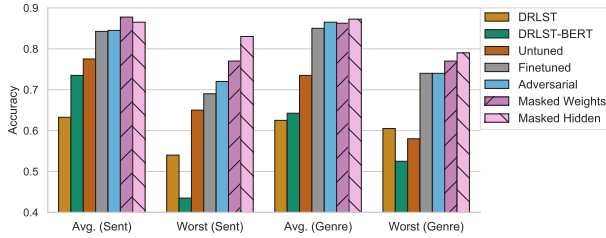
Figure 2: Average and worst main task performance across sentiment/genre combinations. Masked variants (proposed in this paper) are cross-hatched. Large gaps between average and worst performance for a model suggest that it is using the non-target attribute when making predictions for the main task.

	Sentiment ↑	Leakage (Genre) ↓
DRLST	62.1	59.0
DRLST-BERT	67.5	66.3
Untuned	82.3	81.5
Finetuned	87.5	85.5
Adversarial	86.8	80.3
Masked Weights	88.0	72.0
Masked Hidden	88.0	79.0

Table 2: Performance on the main task of sentiment analysis and gender information leakage. The DRLST baselines perform poorly on the main task; the proposed masking approaches have achieve comparable results for sentiment, and expose less genre information.

tive) attribute z. We are interested in evaluating the degree to which explicitly disentangling representations corresponding to these may afford robustness to shifts in the conditional distribution of y given z. As a convenient, illustrative dataset with which to investigate this, we use a set of movie reviews from IMDB (Maas et al., 2011) in which each review has both a binary sentiment label and a genre label.

We pick the two genres of movies that exhibit a strong correlation with review sentiment: Drama (reviews tend to be positive) and Horror (negative), excluding reviews corresponding to other genres and the (small) set of instances that belong to both genres. To investigate robustness to shifts in correlations between z and y we sampled two subsets from the training set such that in the first sentiment and genre are highly correlated, while in the second they are uncorrelated. We report the correlations between these variables in the two subsets in Table 1. We train models on the correlated subset, and then evaluate them on the uncorrelated set.

We compare the proposed masking approaches to several baselines. **Untuned** is a dense classification layer on top of BERT representations (without finetuning). In the **finetuned** variant we omit masks and instead minimize the loss with respect to BERT weights. In the **adversarial** model we adopt 'adversarial debiasing': In addition to minimizing loss on the main task, we train an adversarial classifier to predict the non-target attribute, and the encoder is trained to mitigate the adversaries' ability to do so. We implement this via gradient-reversal (Ganin and Lempitsky, 2015). We also compare to two variational autoencoder baselines: **DRLST** (John et al., 2019) is a VAE model with multi-task loss and adversarial loss; and **DRLST-BERT** is the same model, except we use BERT as the encoder

in place of a GRU (Cho et al., 2014).

Leakage of the Non-target Attribute We evaluate the degree to which representations "leak" non-target information. Following (Elazar and Goldberg, 2018), we first train the model to predict the main task label on the correlated dataset. Then we fix the encoder and train a single layer MLP on the uncorrelated dataset to probe the learned representations for the non-target attribute. Because this probe is trained and tested on only uncorrelated data, it cannot simply learn the main task and exploit the correlation. We report results for our proposed masking models and baselines in Table 2. We also report the results with genre classification as the main task and sentiment as the protected attributes in the Appendix (Section A.1). The DRLST baselines generally underperform, which translates to low leakage numbers but also poor performance on the main task. Compared to the baselines, our masking variants perform comparably with respect to predicting the main task label, but do so with less leakage.

Worst Group Performance In addition to non-target attribute leakage, we measure how models perform on the main task for each subgroup: (Positive, Drama), (Positive, Horror), (Negative, Drama), and (Negative, Horror). Because the distribution of the four groups is unequal in the train set, we expect that models will perform better on attribute combinations that are over-represented in this set, and worse on those that are under-represented, suggesting that the model is implicitly exploiting the correlation between these attributes.

We report both the average and worst performance on the four subgroups; the latter is a proxy to measure robustness when subgroup compositions
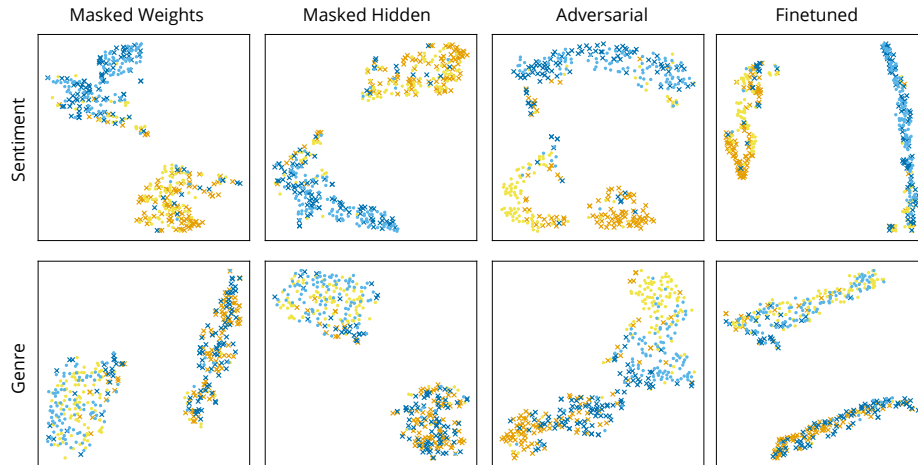
Figure 3: t-SNE projection of sentiment and genre representations of different models. Marker colors denote sentiment (**blue** for positive and **yellow** for negative); marker shapes denote genre (× for drama and ● for horror). In the upper row we expect the points of the same color to be clustered together, but not the points with the same marker shapes, and for the lower row we expect the points of the same marker shapes to be clustered together, but not those of the same colors.

shift between the train to the test set. Figure 2 plots the results. We observe that the masking variants realize similar average performance as the baselines, but consistently outperform these in terms of worst performance. This indicates that the proposed variants rely less on the correlation between the two attributes when predicting the main label.

Qualitative Evaluation In Figure 3 we plot t-SNE visualizations (Maaten and Hinton, 2008) of the representations induced by different models. If the representations are disentangled as desired, instances with different sentiment will be well separated, while those belonging to different genres within each sentiment will **not** be separated. Similarly, for genre representations, instances of the same genre should co-locate, but clusters should not reflect sentiment. No method perfectly realizes these criteria, but the proposed masking approaches achieve better results than do the two baselines. For instance, in the embeddings from the adversarial (Sentiment) and finetuned (Sentiment), instances that have negative sentiment but different genres (● and ×) are separated, indicating that these sentiment representations still carry genre information.

3.2 Disentangling Toxicity from Dialect

Experimental Setup In this experiment we evaluate models on a more consequential task: Detecting hate speech in Tweets (Founta et al., 2018). Prior work (Sap et al., 2019) has shown that existing hate speech datasets exhibit a correlation

between African American Vernacular English (AAVE) and toxicity ratings, and that models trained on such datasets propagate these biases. This results in Tweets by Black individuals being more likely to be predicted as "toxic". Factorizing representations of Tweets into dialectic and toxicity subvectors could ameliorate this problem.

We use (Founta et al., 2018) as a dataset for this task. This comprises 100k Tweets, each with a label indicating whether the Tweet is considered toxic, and self-reported information about the author. We focus on the self-reported race information. Specifically, we subset the data to include only users who self-reported as being either white or Black. The idea is that Tweets from Black individuals will sometimes use AAVE, which in turn could be spuriously associated with 'toxicity'.

Similar to the above experiment, we sampled two subsets of the data such that in the first the (annotated) toxicity and self-reported race are highly correlated, while in the second they are uncorrelated (see Table 1). We train models on the correlated subset, and evaluate them on the uncorrelated set. This setup is intended to measure the extent to which models are prone to exploiting (spurious) correlations, and whether and which disentanglement methods render models robust to these.

Leakage of Race Information We evaluate the degree to which representations of Tweets "leak" information about the (self-reported) race of their authors using the same method as above, and report

	Toxicity ↑	Leakage (Race) ↓
DRLST	66.4	86.6
DRLST-BERT	68.0	90.0
Untuned	68.2	76.8
Finetuned	70.4	93.0
Adversarial	70.2	67.0
Masked Weights	70.4	**59.8**
Masked Hidden	**71.4**	63.5

Table 3: Performance on the main task of toxicity prediction, and leakage of race information. Compared to the baselines, the proposed approaches achieve performance competitive with or better than baselines, while minimizing leakage of the protected attribute.

results in Table 3. We observe that the proposed masking variants perform comparably to baselines with respect to predicting the toxicity label, but leak considerably less information pertaining to the sensitive attribute (race).

Fairness Implications In addition to the degree to which representations encode race information, we are interested in how the model performs on instances comprising (self-identified) Black and white individuals, respectively. More specifically, we can measure the **True Positive Rate (TPR)** and the **True Negative Rate (TNR)** on these subgroups, which in turn inform *equalized odds*, a standard metric used in the fairness literature.

We report the TPR and TNR of each model achieved over white and Black individuals, respectively, as well as the difference across the two groups in Figure 4. We observe that the proposed model variants achieve a smaller TPR and TNR gap across the two races (see rightmost subplots), indicating that performance is more equitable across the groups, compared to baselines.

3.3 Disentangling Semantics from Syntax

Experimental Setup As a final experiment, we follow prior work in attempting to disentangle semantic from syntactic information encoded in learned (BERT) representations of text. Because we have proposed exploiting triplet-loss, we first construct triplets (x_0, x_1, x_2) such that x_0 and x_1 are similar semantically but differ in syntax, while x_0 and x_2 are syntactically similar but encode different semantic information. We follow prior work (Chen et al., 2019; Ravfogel et al., 2020) in deriving these triplets. Specifically, we obtain x_0, x_1 from the ParaNMT-50M (Wieting and Gimpel, 2018) dataset. Here x_1 is obtained by applying back-

translation to x_0, i.e., by translating x_0 from English to Czech and then back into English. To derive x_2 we keep all function words (from a list introduced in Ravfogel et al. 2020) in x_0, and replace content words by masking each in turn, running the resultant input forward through BERT, and randomly selecting one of the top predictions (that differs from the original word) as a replacement.

We compare our disentanglement-via-masking strategies against models that represent state-of-the-art approaches to disentangling syntax and semantics. In particular, we compare against VGVAE (Chen et al., 2019), though we implement this on top of BERT-base to allow fair comparison. Following prior work that has used triplet loss for disentanglement, we also compare against a model in which we finetune BERT using the same triplet loss that we use to train our model, but in which we update all model parameters (as opposed to only estimating mask parameters).

To evaluate learned representations with respect to the semantic and syntactic information that they encode, we evaluate them on four tasks. Two of these depend predominantly on semantic information, while the other two depend more heavily on syntax.[3] For the semantics tasks we use: (i) A word content (WC) (Conneau et al., 2018) task in which we probe sentence representations to assess whether the corresponding sentence contains a particular word; and (ii) A semantic textual similarity (STS) benchmark (Nakov et al., 2013), which includes human provided similarity scores between pairs of sentences. We evaluate the former in terms of accuracy; for the latter (a ranking task) we use Spearman correlation. To evaluate whether representations encode syntax, we use: (i) A task in which the aim is to predict the length of the longest path in a sentence's parse tree from its embedding (Depth) (Conneau et al., 2018); and (ii) A task in which we probe sentence representations for the type of their top constituents immediately below the S node (TopConst).[4]

Figure 5 shows the signed differences between the performance achieved on semantics- and syntax-oriented tasks by BERT embeddings (we mean-pool over token embeddings) and the 'syntax' representations from the disentangled models considered (see the Appendix for the analogous plot for the 'semantics' representations in figure

---

[3]This is a (very) simplified view of 'semantics' / 'syntax'.

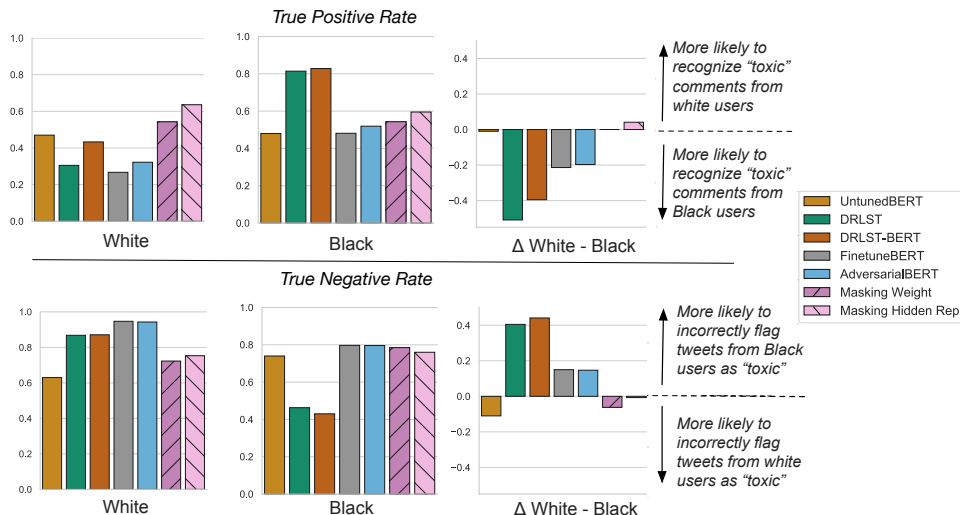[4]See (Conneau et al., 2018) for more details regarding WC, Depth, and TopConst tasks.

Figure 4: True Positive and True Negative Rates achieved on white and Black individuals, respectively, and the (signed) difference between these (rightmost subplots). The proposed masked variants (cross-hatched) are more equitable in performance, while other methods tend to over-predict Tweets written by Black individuals as "toxic".

A.2). Ideally, syntax embeddings would do well on the syntax-oriented tasks (Depth and TopCon) and poorly on the semantic tasks (WC and STS). With respect to syntax-oriented tasks, the proposed masking methods outperform BERT base representations, as well as the alternative disentangled models considered. These methods also considerably reduce performance on semantics-oriented tasks, as we would hope. We emphasize that this is achieved *only via masking, and without modifying the underlying model weights*.

3.4 Identifying *Sparse* Disentangled Sub-networks for Semantic and Syntax

We next assess if we are able to identify *sparse* disentangled subnetworks by combining the proposed masking approaches with *magnitude pruning* (Han et al., 2015a). Specifically, we use the loss function defined in Equation 8 to finetune BERT for k iterations, and prune weights associated with the m smallest magnitudes after training. We then initialize masks to the sparse sub-networks identified in this way, and continue refining these masks via the training procedure proposed above. We compare the resultant sparse network to networks similarly pruned (but not masked). Specifically, for the latter we consider: Standard magnitude tuning applied to BERT, without additional tuning (Pruned + Untuned), and a method in which after magnitude pruning we resume finetuning of the subnetwork until convergence, using the aforementioned loss function (Pruned + Finetuned).

We compare the performance achieved on the semantic and syntax tasks by the subnetworks identified using the above strategies at varying levels of sparsity, namely after pruning: {0, 20%, 40%, 60%, 80%, 85%, 90%, 95%} of weights.[5] We report full results in Appendix Figure A.3, but here observe that combining the proposed masking strategy with magnitude pruning consistently yields representations of semantics that perform comparatively strongly on the semantics-oriented tasks (STS, WC), even at very high levels of sparsity; these semantics representations also perform comparatively poorly on the syntax-oriented tasks (Depth, TopCon), as one would hope. Similarly, syntax representations perform poorly on semantics-oriented tasks, and outperform alternatives on the syntax-oriented tasks. In sum, this experiment suggests that we are indeed able to identify *sparse* disentangled subnetworks via masking.

4 Related Work

Disentangled and structured representations of images. The term *disentangled representations* has been used to refer to a range of methods with differing aims. Much of the initial focus in this space was on learning representations of images, in which certain dimensions correspond to interpretable factors of variation (Kulkarni et al., 2015;

---

[5]Technically, in the Pruned + Masked Weights method, refining the masks may change subnetwork sparsity, but empirically we find this to change the sparsity only slightly (~1% in all of our experiments).
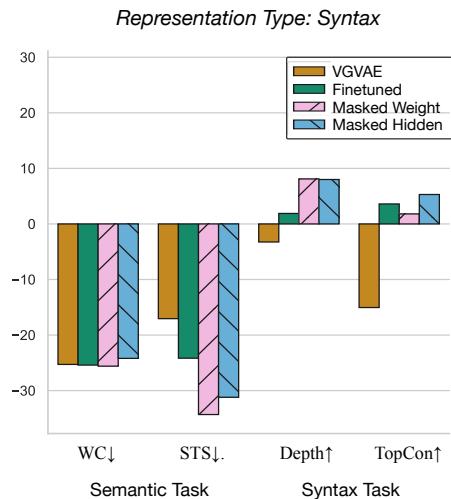
Figure 5: Differences between performances achieved via BERT embeddings and the disentangled model variants considered on semantics-oriented (WC, STS) and syntax-oriented (Depth, TopCon) tasks compared with BERT embeddings. We plot this difference with respect to the syntax embeddings induced by the models.

Higgins et al., 2017; Chen et al., 2016). In the context of variational autoencoders (Kingma and Welling, 2014; ?) this motivated work that evaluates to what extent such representations can recover a set of ground-truth factors of variation when learned without supervision (Eastwood and Williams, 2018; Kim and Mnih, 2018; Chen et al., 2018). Other work has investigated representations with the explicit motivation of fairness (Locatello et al., 2019a; Creager et al., 2019), which disentanglement may help to facilitate.

Disentangling representations in NLP. Compared to vision, there has been relatively little work on learning disentangled representations of text. Much of the prior work on disentanglement for NLP that does exist has focused on using such representations to facilitate *controlled generation*, e.g., manipulating sentiment (Larsson et al., 2017).

A related notion is that of *style transfer*, for example, separating style from content in language models Shen et al. (2017); Mir et al. (2019). There has also been prior work on learning representations of particular aspects to facilitate domain adaptation (Zhang et al., 2017), and aspect-specific information retrieval (Jain et al., 2018). Esmaeili et al. (2019) focus on disentangling user and item representations for product reviews. Moradshahi et al. (2019) combine BERT with Tensor-Product Representations to improve its transferability across

different tasks. Recent work has proposed learning distinct vectors coding for semantic and syntactic properties of text (Chen et al., 2019; Ravfogel et al., 2020); these serve as baseline models in our experiments.

Finally, while not explicitly framed in terms of disentanglement, efforts to 'de-bias' representations of text are related to our aims. Some of this work has used adversarial training to attempt to remove sensitive information (Elazar and Goldberg, 2018; Barrett et al., 2019).

Network pruning. A final thread of relevant work concerns selective pruning of neural networks. This has often been done in the interest of model compression (Han et al., 2015a,b). Recent intriguing work has considered pruning from a different perspective: Identifying small subnetworks — winning 'lottery tickets' (Frankle and Carbin, 2019) — that, trained in isolation with the right initialization, can match the performance of the original networks from which they were extracted. Very recent work has demonstrated that winning tickets exist within BERT (Chen et al., 2020).

5 Discussion

We have presented a novel perspective on learning disentangled representations for natural language processing in which we attempt to uncover existing subnetworks within pretrained transformers (e.g., BERT) that yield disentangled representations of text. We operationalized this intuition via a *masking* approach, in which we estimate *only* binary masks over weights or hidden states within BERT, leaving all other parameters unchanged. We demonstrated that — somewhat surprisingly — we are able to achieve a level of disentanglement that often exceeds existing approaches (e.g., a varational auto-encoder on top of BERT), which have the benefit of finetuning all model parameters.

Our experiments demonstrate the potential benefits of this approach. In Section 3.1 we showed that disentanglement via masking can yield representations that are comparatively robust to shifts in correlations between (potentially sensitive) attributes and target labels. Aside from increasing robustness, finding sparse subnetworks that induce disentangled representations constitutes a new direction to pursue in service of providing at least one type of model interpretability for NLP. Finally, we note that sparse masking (which does not mutate the underlying transformer parameters) may offer

efficiency advantages over alternative approaches.

6   Acknowledgements

References

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6330–6335, Hong Kong, China. Association for Computational Linguistics.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *ArXiv preprint*, abs/1308.3432.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. A multi-task approach for disentangling syntax and semantics in sentence representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2453–2464, Minneapolis, Minnesota. Association for Computational Linguistics.

Tian Qi Chen, Xuechen Li, Roger B. Grosse, and David Duvenaud. 2018. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2615–2625.

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. 2020. The lottery ticket hypothesis for pre-trained BERT networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2172–2180.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A. Weis, Kevin Swersky, Toniann Pitassi, and Richard S. Zemel. 2019. Flexibly fair representation learning by disentanglement. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1436–1445. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cian Eastwood and Christopher K. I. Williams. 2018. A framework for the quantitative evaluation of disentangled representations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Babak Esmaeili, Hongyi Huang, Byron C. Wallace, and Jan-Willem van de Meent. 2019. Structured neural topic models for reviews. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 3429–3439. PMLR.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yaroslav Ganin and Victor S. Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1180–1189. JMLR.org.

Song Han, Huizi Mao, and William J Dally. 2015a. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *ArXiv preprint*, abs/1510.00149.

Song Han, Jeff Pool, John Tran, and William J. Dally. 2015b. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1135–1143.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. 2018. Towards a Definition of Disentangled Representations. *ArXiv preprint*, abs/1812.02230.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Geoffrey Hinton, Nitsh Srivastava, and Kevin Swersky. 2012. Neural networks for machine learning. *Coursera, video lectures*, 264(1).

Sarthak Jain, Edward Banner, Jan-Willem van de Meent, Iain J. Marshall, and Byron C. Wallace. 2018. Learning disentangled representations of texts with application to biomedical abstracts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4683–4693, Brussels, Belgium. Association for Computational Linguistics.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning the difference that makes A difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Hyunjik Kim and Andriy Mnih. 2018. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2654–2663. PMLR.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Tejas D. Kulkarni, William F. Whitney, Pushmeet Kohli, and Joshua B. Tenenbaum. 2015. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2539–2547.

Maria Larsson, Amanda Nilsson, and Mikael Kågebäck. 2017. Disentangled representations for manipulation of sentiment in text. *ArXiv preprint*, abs/1712.10066.

Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes, editors. 2019. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Florence, Italy.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. 2019a. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14584–14597.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019b. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124. PMLR.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of*

the *Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

Mehrad Moradshahi, Hamid Palangi, Monica S Lam, Paul Smolensky, and Jianfeng Gao. 2019. Hubert untangles bert to improve transfer across nlp tasks. *ArXiv preprint*, abs/1910.12647.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 task 2: Sentiment analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Jacob Goldberger, and Yoav Goldberg. 2020. Unsupervised distillation of syntactic information from contextualized word representations. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 91–106, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text

by cross-alignment. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6830–6841.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Yuan Zhang, Regina Barzilay, and Tommi Jaakkola. 2017. Aspect-augmented adversarial networks for domain adaptation. *Transactions of the Association for Computational Linguistics*, 5:515–528.

Mengjie Zhao, Tao Lin, Fei Mi, Martin Jaggi, and Hinrich Schütze. 2020. Masking as an efficient alternative to finetuning for pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2226–2241, Online. Association for Computational Linguistics.

A   Appendix

A.1   Additional IMDB Results

In Table A.1 we report results treating genre classification as the main task, and sentiment as the 'protected' attribute.

|	Genre ↑	Leakage (Sentiment) ↓
DRLST	65.6	61.1
DRLST-BERT	71.4	70.3
Untuned	81.5	82.3
Finetuned	87.3	86.0
Adversarial	85.0	75.5
Masked Weights	87.0	73.0
Masked Hidden	85.0	79.0

Table A.1: Performance of models when treating genre prediction as the main task and sentiment as the 'protected' variable. The DRLST baselines perform poorly on the main task. The proposed masking methods achieve performance comparable to other baselines, while encoding less of the sentiment information.

## A.2   Distribution of Learned Masks Across BERT Layers

Here we inspect the subnetworks (i.e., the weights or hidden activations that are not masked) uncovered by our model, which may provide insights regarding where pretrained (masked) language models encode different sorts of linguistic information. Figure A.1 shows the distributions of the two types of masks (weights and hidden activations, respectively) over the layers within BERT for for the semantics/syntax tasks. We observe that the learned 'semantic' mask zeros out fewer elements at higher layers in the network, while the 'syntax' mask prefers to keep non-zero entries in lower layers. This suggests that semantic information may be captured mostly in higher layers of BERT, while syntactic information may be encoded in lower layers, consistent with observations in prior work (Tenney et al., 2019).

A.3   Semantic Representation Performance (vs. BERT)

We show the signed differences between the performance achieved on semantics- and syntax-oriented tasks by BERT embeddings (we mean-pool over token embeddings) and the 'semantic' representations from the disentangled models in figure A.2.
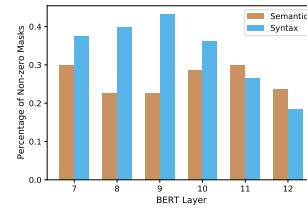
(a) Masking Weights
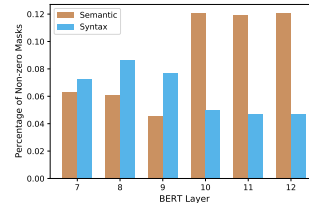
(b) Masking Hidden Activations

Figure A.1: Non-zero (unmasked) elements of the semantic and syntax subnetworks induced by masking weights (a) and hidden representations / activations (b).
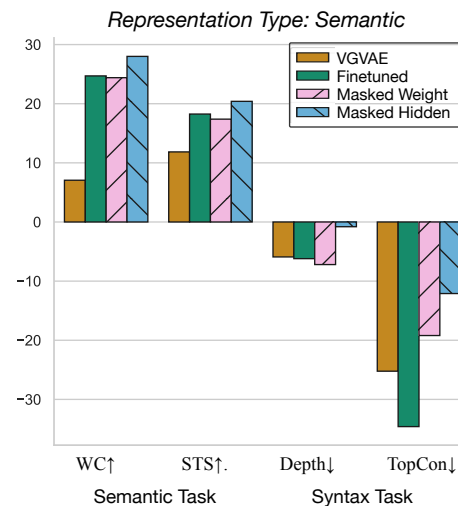

Figure A.2: Differences between the performances achieved via BERT embeddings and the disentangled model variants considered on semantics-oriented (WC, STS) and syntax-oriented (Depth, TopCon) tasks compared with BERT embeddings. We plot this difference with respect to the semantic embeddings induced by the models.
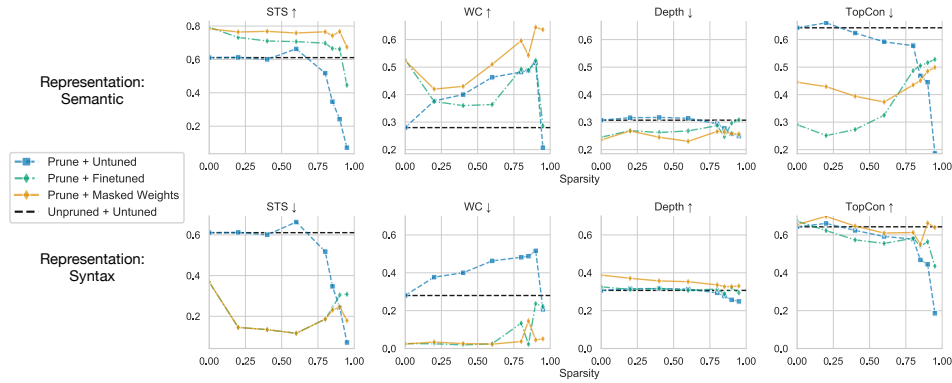
Figure A.3: Model performance as a function of the degree of pruning. The x-axis corresponds to the subnetwork sparsities (percent of weights dropped), while the y axes are performance measures — accuracy for all tasks except for STS, where we report Pearson's correlation. We compare the performance of models trained on the semantic (top) and syntax representations (bottom) learned by the disentangling strategies considered, after pruning to varying levels of sparsity.

Layers Masked	Sentiment ↑	Leakage (Genre) ↓
Last 3	86.3	73.5
Last 6	88.0	72.0
Last 9	88.0	72.8
All 12	87.5	72.3

Table A.2: Performance of model (Masking Weights) when masking the last 3, 6, 9 or 12 layers of BERT.

α	Sentiment ↑	Leakage (Genre) ↓
0.5	83.8	74.8
1.0	86.3	73.0
2.0	88.0	72.0
5.0	87.8	70.3

Table A.3: Performance of model (Masking Weights) when choosing different α for the triplet loss.

A.4 Model performance with iterative magnitude pruning

We report full results of combining our method with magnitude pruning to uncover sparse subnetworks in Figure A.3. We compare our method to several alternative pruning strategies: Standard magnitude tuning applied to BERT, without additional tuning (Pruned + Untuned), and a method in which after magnitude pruning we resume finetuning of the subnetwork for a fixed number of steps, using the aforementioned loss function (Pruned + Finetuned).

A.5 Additional Experiments:Perturbation Study of Hyper-parameters

We report model performance when masking different number of layers of BERT (Table A.2) and when choosing different values for α(Table A.3).

A.6 Model performance with varying degree of correlation in the training set

We report the comparison of our model (Masking Weights) with two baselines (Finetuned and Adversarially trained BERT) with varying degree of correlation in the training set. The task is sentiment classification and we control the correlation between sentiment and genre into 3 different settings: strong, moderate and weak (if any) correlation. We report the results in Table A.4. Our model significantly outperforms the baselines when the correlation is strong, and the advantage begins to diminish as the correlation becomes weaker, as we would expect.

	Finetuned	**Adversarial**	**Masking Weights**
Strong Correlation: 15% of "Drama" Reviews are Positive			
Avg.(Sentiment)	84.3	84.5	87.8
Worst (Sentiment)	69.3	72.0	77.0
Moderate Correlation: 25% of "Drama" Reviews are Positive			
Avg.(Sentiment)	86.8	86.5	86.3
Worst (Sentiment)	72.0	76.8	80.5
No Correlation: 50% of "Drama" Reviews are Positive			
Avg.(Sentiment)	88.5	88.0	86.8
Worst (Sentiment)	86.0	87.0	85.3

Table A.4: Performance of our model (Masking Weights) compared to baseline models with varying degree of correlation in the training set.