Learning Faithful Representations of Causal Graphs

Ananth Balashankar **Lakshminarayanan Subramanian**
New York University
New York, NY, USA
{ananth,lakshmi}@nyu.edu

Abstract

Learning contextual text embeddings that represent causal graphs has been useful in improving the performance of downstream tasks like causal treatment effect estimation. However, existing causal embeddings which are trained to predict direct causal links, fail to capture other indirect causal links of the graph, thus leading to spurious correlations in downstream tasks. In this paper, we define the faithfulness property of contextual embeddings to capture geometric distance-based properties of directed acyclic causal graphs. By incorporating these faithfulness properties, we learn text embeddings that are 31.3% more faithful to human validated causal graphs with about 800K and 200K causal links and achieve 21.1% better Precision-Recall AUC in a link prediction fine-tuning task. Further, in a crowdsourced causal question-answering task on Yahoo! Answers with questions of the form "What causes X?", our faithful embeddings achieved a precision of the first ranked answer (P@1) of 41.07%, outperforming the existing baseline by 10.2%.

1 Introduction

Learning distributed word representations that capture causal relationships are useful for real-world natural language processing tasks (Roberts et al., 2020; Veitch et al., 2020; Gao et al., 2018, 2019). Approximating the notion of causality with a similarity-based distance metric using separate vector representations for cause and effect tokens has led to significant improvement in the performance of downstream tasks like Question Answering, but can be too restrictive to generalize over unobserved edges in larger causal graphs (Sharp et al., 2016). In downstream causal reasoning based tasks like dialog systems (Ning et al., 2018), explanation generation (Grimsley et al., 2020), question answering (Sharp et al., 2016), it is important to align the models with the corresponding causal graph. However, words that have low cosine similarity capture various semantic similarities, like relatedness, synonyms, replaceability, or complementarity, but not directionality (Hamilton et al., 2017). Hence, any symmetric distance in an embedding space cannot convey the directed causal semantics for a downstream task (Mémoli et al., 2016). In this paper, we overcome these two shortcomings and propose to optimize for directed *faithfulness* (Spirtes et al., 1993) that word embeddings have to satisfy towards a causal graph.

Prior work on capturing sufficient information for causal inference tasks from embeddings aims to directly use them for average treatment effect estimation (Veitch et al., 2020). We are, however, interested in a complementary question: "Can we learn word embeddings based on a distance measure that maps the directed distance between nodes in a causal graph to that in the embedding space?". Unlike prior work, which aims to learn a causal aware embedding restricted to direct link prediction (Hamilton et al., 2017), we propose faithfulness constraints so that causal word embeddings aims to preserve the partial ordering over pairwise distances in the directed causal graph. In this paper, to achieve the goal of learning faithful word embeddings with a vocabulary of more than 100K tokens, we minimize faithfulness violations over pairwise samples of nodes in the causal graph. Through this constrained optimization, we learn an embedding that can be applied directly for causal inference tasks but also generalizes to emergent causal links. It has been shown that NLP models need to understand such causal links that persist in the real world for safe deployment (Gao et al., 2018; Mishra et al., 2019). Embeddings that violate the faithfulness property, can lead to spurious correlations based on co-location in the embedding space. For example, in a Yahoo! causal question-answering task's example: “What causes nosebleed?”: the answers were “dry air”, “heavy dust”, “damaged nasal cells” and “liver problems”. If we were to only rely on an undirected association based embeddings, the causes “dry air” and “liver problems” might be nearby (with distance of 2), but would be appropriately placed far in a *directed* causality based embedding space. To capture such asymmetric properties, we aim to preserve alignment with the causal graph by mapping causal links to an asymmetric quasi-pseudo distance measure during training to capture directionality of the causal graph as per Figure 1. Since human validated causal graphs can be used directly to answer questions of the type “What causes X?”, we demonstrate the utility of learning faithful representations by using our distance-based features to solve the Yahoo! causal question-answering (QA) task. A causal QA task, unlike a standard QA task, can directly benefit from incorporating a causal graph into word embeddings to answer anti-causal queries. Our key contributions are:

- We define a faithfulness property for word embeddings over a causal graph, that captures geometric properties of the causal graph, beyond the direct link prediction by ensuring global proximity preservation.
- We propose a methodology to learn faithful embeddings through violation minimization which improves neighborhood detection by 31.3%, uniformity by 42.6%, and distance correlation by 54.2% using a quasi-pseudo distance metric.
- The faithful BERT and RoBERTa-based embeddings we learn, when used as inputs to a causal QA task, increases the precision of the first ranked answer (P@1) over existing baselines by 10.2%.

2 Related Work

2.1 Causal Model Representations

Causal Inference, as outlined in (Pearl, 2009) formalizes cause and effects discovered through intervention based experiments and communicates them via directed acyclic graphs. With the availability of large observational datasets for machine learning, various methods and assumptions have been proposed for learning causal graphs (Schölkopf, 2019), data fusion and transportability properties (Bareinboim and Pearl, 2016; Bonner and Vasile, 2017). Specifically, our work closely aligns with the assumption of faithfulness (Spirtes et al., 1993), which requires that the observed probability distributions of nodes in a causal graph are conditionally independent as per the links in the graph. In our work, we use the probability distributions as modeled in a natural language model (Kuhn and De Mori, 1990) and align it with the causal links in a graphical causal model. We extend the faithfulness assumption to be reflected in embeddings learnt by a masked language model (Devlin et al., 2019; Liu et al., 2019b) for downstream tasks. This definition of faithfulness is different from the one proposed by (Jacovi and Goldberg, 2020) used to evaluate models for interpretability of models used for downstream tasks. Instead, our work builds on embeddings learnt in (Sharp et al., 2016), given a causal model and learn embeddings that are bootstrapped using a small set of cause-effect seeds. Causal models have also been used to learn auxiliary tasks (Feder et al., 2020) using adversarial training to ensure that a language model learns causal-inspired representations. Such approaches use causal models to learn counterfactual embeddings invariant to the presence of confounding concepts in a sentence, while we encode the geometrical properties of causal graphs into the embeddings and the distance measure to maintain their faithfulness. In principle, we adopt a similar approach to (Veitch et al., 2020) of fine-tuning towards a causal link prediction task. This is in contrast with approaches that use energy-based transition vectors used to represent the cause-to-effect and effect-to-cause links (Zhao et al., 2017). Our approach uses regularization constraints similar to the ones

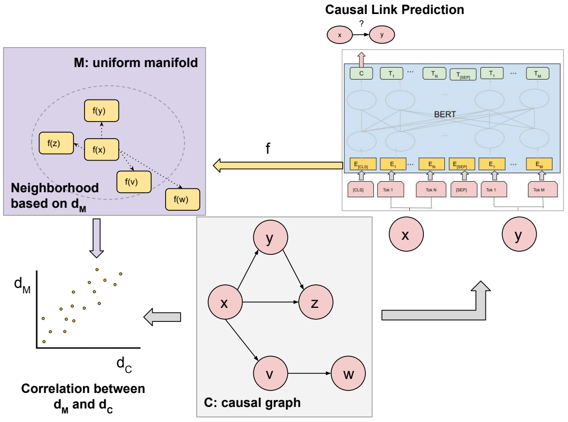

Figure 1: Schematic of our Faithful BERT-based model

proposed for information bottlenecks in word embeddings (Li and Eisner, 2019; Goyal and Durrett, 2019), text-based games (Narasimhan et al., 2015), activation links in neuroscience (Chalupka et al., 2016), causal consistency with ordinary differential equations (Rubenstein et al., 2017) and temporal Granger Causality (Tank et al., 2018). For an extensive survey of using text for causal inference tasks, we refer to (Keith et al., 2020).

2.2 Graph Representation Learning

Learning asymmetric transitive graph representations which generalize the causal graph have been studied extensively in Information Retrieval (Chen et al., 2007; Epasto and Perozzi, 2019; Li et al., 2019; Grover and Leskovec, 2016). They either utilize a random walk learning technique (Perozzi et al., 2014) or matrix factorization techniques (Lee and Seung, 2000; Tenenbaum et al., 2000; Wang et al., 2017; Mikolov et al., 2013) to incorporate priors such as the stationary transition probability matrix, community structure, etc. More recently, (Liu et al., 2019a; Ostendorff et al., 2019; Lu et al., 2020) have incorporated knowledge graphs in BERT and shown increased accuracy in knowledge-centric NLP tasks. (Zhou et al., 2017; Gordo and Perronnin, 2011; Ou et al., 2016; Sun et al., 2018; Tang et al., 2015) propose asymmetric higher order proximity preserving graph embedding methods by learning separate source and target embeddings. While we can learn faithful 3-dimension embeddings for any fixed finite undirected graph deterministically (Cohen et al., 1995), fine-tuning pre-trained word embeddings such that they generalize over all sub-graphs in a directed graph is known to be a hard graph kernel design problem that scales cubically with the number of nodes (Vishwanathan et al., 2010). Our approach builds on efforts to incorporate graph-like structure in BERT, but overcomes the issue of learning dual embeddings for cause-effect edges by learning unified embeddings for both cause and effect roles of words. Through such embeddings, we can further aid causal discovery that is not yet captured in a graphical notation (Chen et al., 2014).

2.3 Graph Neural Networks

Recently, Graph neural networks that capture the graph neighborhood structure have been employed in link prediction (Zhu et al., 2020; Abu-El-Haija et al., 2017). In (You et al., 2018), the problem is reduced to that of sequence prediction by reducing the graph to breadth-first search based deterministic sequence. In (Li et al., 2018), node embeddings are updated after several rounds of message passing, while in (Tu et al., 2016) a variant of the random walk is incorporated with a max-margin discriminative constraint. In (Velikovi et al., 2018), models are learned by attending over the neighborhood of nodes for context, while (Kipf and Welling, 2016) apply spectral graph convolutions for a self-supervised learning task. We adopt the incremental approach proposed in (Velikovi et al., 2018) which does not rely on knowing the entire graph structure apriori and fine-tune on cause-effect pairs for the link prediction task on a pre-trained BERT-based language model.

3 Learning Faithful Embeddings

3.1 Background

Causal inference (Pearl, 2009) aims to understand the cause and effect relationships between events. For events observed in the real world, learning purely based on correlations can lead to spurious causal links (e.g "wet grass" causes "rainfall") and can severely impact downstream tasks. Hence, intervention-based studies are conducted which carefully study the impact of a cause using controlled randomized experiments (e.g. vaccine trials) and other criterion to learn if links between causes and effects exist using observed data under specific assumptions. The findings of such studies are formalized using frameworks like Rubin Causal Models (Rubin, 1974), Structural Causal Models (Pearl, 2009), etc. While there are differences in abstractions between them, there is formal equivalence (Galles and Pearl, 1998) in modeling counterfactuals ("What is the effect when the *cause* is intervened?") and we refer the reader to (Pearl and Mackenzie, 2018) for a primer in causal modeling.

In this paper, we assume a graphical structural causal model C (Pearl, 2009) is given, whose nodes are linked with directed edges that denote the cause-effect relationship. For example, the cause-effect of "smoking" causes "cancer", references to the real world action of "smoking" in individuals that leads to the development of "cancer" kind of disease in those individuals. While causal models have a close relationship to the knowledge graph, the links of the causal graph have a well-defined causal interpretation that can be validated through counterfactual experiments. In this work, we assume the availability of such a causal graph and we do not aim to build one. Instead, we rely on human annotators who with the help of web crawlers (Heindorf et al., 2020a) and other information retrieval tools (Sharp et al., 2016) produce a directed graphical causal model as shown in Figure 1.

3.2 Faithfulness

Given a graphical causal model C, we now present a faithfulness property an embedding that aims to closely align with the causal model has to satisfy. The faithfulness property was first proposed for any two causal spaces in (Bombelli et al., 2013) in the domain of quantum physics with the space-time dimension. Inspired by this, we propose an instantiation for word embeddings and a corresponding graphical causal model.

Definition 1 (Faithfulness). *An embedding $f : C \rightarrow M$ from a causal set (C, d_C) to a vector space (M, d_M) is faithful if:*

- $\exists \lambda, \forall x, y \in C, d_C(x, y) = 1 \Leftrightarrow d_M(f(x), f(y)) \leq \lambda$
- $f(C)$ *is distributed uniformly*
- $\forall x, y, w, z \in C, d_C(x, y) \leq d_C(w, z) \Leftrightarrow d_M(f(x), f(y)) \leq d_M(f(w), f(z))$

Note that we use the causal set (C, d_C) as a tuple of the graphical causal model C and a distance measure d_C which is used to measure the *directed* distance between nodes in the graph. The vector space in which we map our embeddings is also characterized by a tuple (M, d_M), where M is the multidimensional real number space R^m, and a distance measure d_M which identifies nearby words in that vector space. The three conditions posed by the faithfulness property, more concretely specify that there needs to be a real threshold, within the embedding space, which can cover all the neighboring nodes of a word, the embedding space needs to be uniformly distributed, and finally, any inequality relationships between two distance measures in the causal graph needs to hold in the embedding space too. An embedding that satisfies this property can then be used to sufficiently represent the causal graph in downstream tasks.

3.2.1 Distance Measures

The definition of faithfulness is dependent on the distance measure used in both the causal graph and the embedding domains. In this work, we assume that the causal graph is a directed acyclic graph, and hence we measure d_C as the shortest directed distance (number of edges in an unweighted graph) between two nodes. If no such path exists between two nodes, we consider the distance to be a large number, which in the case of an unweighted graph, can be set to $> n$, where n is the number of nodes in the acyclic graph. Note that weighted graphs can also be incorporated with minor changes based on the maximum path in the graph.

However, the distance measure in the embedding space faces challenges in evaluation of simple supervised tasks (Jastrzebski et al., 2017). To overcome these, we chose a distance measure that is closely tied to our faithfulness definition. We chose a unified set of embeddings for both the cause u and effect v, and, if there exists a causal edge from $u \rightarrow v$, then we would expect that $d_M(f(u), f(v)) << d_M(f(v), f(u))$. For this reason, symmetric distance choices like Euclidean distance, cosine similarity are not suitable. Our chosen distance measure, hence should follow the properties of quasi-pseudo metrics, defined as follows in (Moshokoa, 2005):

Definition 2 (Quasi-Pseudo Metric). *A measure $d_M : X \times X \rightarrow [0, \infty)$ is a quasi-pseudo metric if $\forall x, y, z \in X$,*

- $d_M(x, y) \geq 0$
- $d_M(x, x) = 0$, *but* $d_M(x, y) = 0$ *is possible for* $x \neq y$
- $d_M(x, z) \leq d_M(x, y) + d_M(y, z)$

Hence, quasi-psuedo metrics, which do *not* satisfy the symmetry property are best suited to measure the distance between any two embeddings. We can generate such metrics, given a measure d. If the cause phrase u has p word tokens, and the effect phrase v has q word tokens, we choose the Max-Matching method given in (Xie and Mu, 2019) in our definition of d_M by iterating through all pairs of words $(v_b, u_a) : v_b \neq u_a$. Note that the measure d computes the difference between v to u over the total m number of dimensions in $f(v_b), f(u_a)$.

$$d(u, v) = \min_{\substack{a=1..p \\ b=1..q \\ v_b \neq u_a}} \sum_{j=1}^{m} (f_j(v_b) - f_j(u_a)) \tag{1}$$

$$d_M(f(u), f(v)) = \begin{cases} d(u, v), & \text{if } d(u, v) > 0 \\ 10^{-d(u,v)} - 1, & \text{otherwise} \end{cases} \tag{2}$$

We chose this definition, as it is differentiable (except at 0, where we choose the gradient to be 0). Also, for each point u in the embedding space, there is a corresponding hyperplane that passes through it that defines the half-space which separates the reachable nodes $v : d(u, v) > 0$ - nodes which have either an indirect or direct causal link and the unreachable nodes $v : d(u, v) < 0$. Also, by the property of $d(u, v) = -d(v, u)$, we see that if v is reachable from u, then u is not reachable from v, thus affirming that this is suitable to represent a causal graph that is directed and acyclic.

3.3 Causal Graph Link Prediction

There are currently many approaches to learning causal representations, one which uses a masked language modeling approach where the word tokens in the cause are paired with word tokens in the effect using a skip-gram technique in an unsupervised setting. In the supervised setting, models align the cause-effect embeddings to solve either a sequence-to-sequence translation task or logistic classification task. Since we aim to capture all the nodes of the causal graph into a single set of word embeddings, we choose this approach. Further, in the supervised setting, we make explicit the causal relationship between cause and effect, thereby capturing the directionality of the linkage. Thus, a supervised model could translate a cause to an effect or predict the link that exists from a cause to an effect. Among these supervised modeling choices, we choose the binary classification task of predicting if a directed edge exists between two nodes in the causal graph. This supervised learning is achieved by following the technique of fine-tuning as proposed in (Veitch et al., 2020). Formally, given a cause phrase u, an effect phrase v, let an $i(u, v)$ be an edge indicator variable $i(u, v) = \mathbb{1}_{u \to v}$ that takes binary values of $\{0, 1\}$ based on the existence of an edge from $u \to v$ in the causal graph.

Pre-trained Contextual Models: Pre-trained models based on transformers like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b) learn contextual embeddings of words or tokens by optimizing for the self-supervision task of predicting randomly masked tokens in a sentence. These pre-trained embeddings for word tokens have been used extensively for fine-tuning. Here, we use such fine-tuned models denoted as $\tilde{g}$ to predict the existence of an edge between the cause and effect u, v, by embedding them into $f(u), f(v)$ respectively and further optimizing them in the fine-tuning stage on the following cross-entropy classification loss

$$\mathcal{L}_s = \mathbb{E}_{u,v \sim C}\, \text{CrossEnt}(i(u, v), \tilde{g}(u, v)) \tag{3}$$

3.4 Violation Minimization

Given the faithfulness definition, our goal is to learn an embedding that minimizes the number of violations of the faithfulness property. For each of the 3 conditions present in the faithfulness property, we define how we measure their adherence and incorporate it in the loss function. In addition to the causal graph link prediction task, we now present how the faithfulness properties are incorporated through regularization constraints.

3.4.1 Neighborhood

Since we expect a single embedding distance threshold that perfectly encapsulates the neighborhood of a node, we can measure this by varying distance thresholds for neighborhood detection and compute the area under the curve of the precision-recall curve. Since we aim to retain all the neighbors of a node in the causal graph within an upper bound of the distance in the embedding space, we add the sum of the distance between the nodes and their neighbors as an L1 regularization loss.

$$\mathcal{L}_n = \mathbb{E}_{\substack{u \sim C \\ v \in Neigh(u)}} |d_M(f(u), f(v))| \tag{4}$$

3.4.2 Uniformity

Since checking for true uniformity can be computationally intractable, we approximate by computing the per-dimension aggregate of all the word embeddings and compute the Wasserstein distance (Olkin and Pukelsheim, 1982) between the observed distribution and the expected uniform distribution centered around zero (0^m). Since, in the uniformity constraint, we would expect that the embeddings are centered around zero, the mean of the embeddings should be close to zero. We measure the distance from this expected centroid and penalize the model for a high distance. If C_b denote the set of nodes chosen in a batch b, with size $|b|$, and $f_j(p)$ denote the j^{th} dimension of the embedding of node p, then we present the uniformity regularization loss:

$$\mathcal{L}_u = \sum_{j=1}^{m} \frac{1}{|b|} \sum_{p \in C_b} f_j(p) \tag{5}$$

3.4.3 Distance Correlation

To measure if inequalities between two distances in the causal graph hold in the embedding space, we measure the Pearson correlation coefficient between samples of distances between words in the causal graph and that of the embeddings. To ensure that any two distances sampled from the causal graph maintain the same inequality in the embedding space, we sample random nodes from the causal graph and compute the empirical Pearson Correlation Coefficient of their distances in the embedding space. A perfect correlation would lead to a coefficient of +1, so we penalize any deviation from that ideal correlation and present the distance correlation loss:

$$\begin{aligned}\mathcal{L}_c &= 1 - \rho_{d_C, d_M} \\ &= 1 - \frac{cov(d_C, d_M)}{\sigma_{d_C}\sigma_{d_M}}\end{aligned} \tag{6}$$

Note that all the above constraints are at a batch level and hence is added on to the batch cross-entropy loss during every back-propagation step. Since the losses are differentiable, we have used the auto-diff capability available in Tensorflow. The contribution of each of the above losses are combined using the Augmented Lagrangian method (Hestenes, 1969) and controlled using 3 parameters α, β, γ as follows:

$$\mathcal{L} = (1 - \alpha - \beta - \gamma)\mathcal{L}_s + \alpha\mathcal{L}_n + \beta\mathcal{L}_u + \gamma\mathcal{L}_c \tag{7}$$

The values of these hyperparameters were chosen to be $0.1, 0.15, 0.1$ respectively after cross-validation to optimize causal link prediction accuracy and faithfulness metrics. A summary of our approach is outlined in Algorithm 1.

The learning rate $a = 0.01$, $\mathcal{L}_u, \mathcal{L}_c$ are computed per batch by maintaining the required variables $f(u), f(v), d_C(u,v), d_M(f(u), f(v))$ in memory. These are implemented using Tensorflow's eager execution framework.

4 Evaluation

4.1 Causal Evidence Graphs

The causal evidence graphs we use contain phrases like “heavy rainfall” as causes and effects, which require us to learn the combined embeddings of the phrases. Restricting ourselves to just individual words would leave out the context required to understand the context to understand the cause-effect pairs. For example, the kind of effects “heavy rainfall” might have could be different from just “rainfall”. We thus utilize the contextual embedding framework used to learn language models in BERT (Devlin et al., 2019), as a way to learn contextual embeddings that align with a given graphical causal model. Note that there may be more than one causal model provided by experts based on their domains, and it is important to view our contribution as a way to align with domain expertise (for example, medical, legal, privacy, etc) with their respective causal models as a common mechanism to represent the said domain knowledge.

Algorithm 1 Faithful Embedding Training

```
Input: Pre-trained BERT based model $\tilde{g}$, causal graph $C$, distance measures: $d_C, d_M$,
for e=1..$epochs$ do
    $\mathcal{L} = 0$
    for j=1..b do
        $u, v \sim C : \sum \mathbb{1}_{i(u,v)=0} = \sum \mathbb{1}_{i(u,v)=1}$
        $\mathcal{L}_s$ += CrossEnt($i(u,v), \tilde{g}(u,v)$)
        $\mathcal{L}_n$ += $\sum_{w \in Neigh(u)} d_M(f(u), f(w))$
        Store $f(u), f(v)$ to update $\mathcal{L}_u$
        Store $d_C(u,v), d_M(f(u), f(v))$ to update $\mathcal{L}_c$
    end for
    Update $\mathcal{L}_u, \mathcal{L}_c$ and compute $\mathcal{L}$ (Eqn 7)
    Backprop $\tilde{g} \leftarrow \tilde{g} - a(\frac{\partial \mathcal{L}}{\partial \tilde{g}})$
end for
```

We use two causal graphs to construct their respective faithful embeddings, and demonstrate the utility of the embeddings in downstream tasks. The first causal graph we use is identical to the one used in (Sharp et al., 2016), which uses the 815,233 cause-effect pairs extracted from the Annotated Gigaword and Wikipedia dataset, and an equal number of random relation pairs that are not causal as negative samples. The second causal graph is extracted from the web by (Heindorf et al., 2020b), who use a bootstrapping approach with the initial pattern of “A causes B” and apply it to the ClueWeb12 web crawl dataset with 733,019,372 English web pages, between February and May 2012. From this web crawl, they provide a causal graph with 80,223 concept nodes and 199,803 causal links between the nodes. This graph has been sampled and validated by human annotators with over 96% precision. For our indirect evaluation based on downstream question answering tasks, we use the 3031 causal questions from Yahoo! Answers corpus (Sharp et al., 2016). These questions are of the form “What causes X?”, and we use our faithful embeddings as a drop-in replacement for this causal QA task.

4.2 Metrics

Evaluating embeddings intrinsically has often led to varying leaderboards (Jastrzebski et al., 2017), hence we evaluate our embeddings based on their ability to map to the cause-effect relationship directly. We measure the faithfulness of the trained embeddings, using 3 metrics, one per property as per Eqns 4, 5, 6. For the neighborhood condition, we measure the area under the precision-recall curve as we choose multiple thresholds to define the neighborhood in the embedding space to correspondingly identify the relevant neighbors in the causal graph. For the uniformity condition, we measure the means of the per-dimension values of the word embeddings and compute the 1^{st} Wasserstein (Olkin and Pukelsheim, 1982) distance from the expected centroid of zero. We also perform a statistical test for uniform distribution, which measures the mean Kolmogorov-Smirnov (K-S) test statistic (Daniel, 1990) by bucketing embedding each dimension into 10 buckets. Since each dimension's test statistic can either pass or fail the test based on the significance level, we present the total number of dimensions that pass the test at $\alpha = 0.05$ significance level. Finally, to measure the distance correlation property, we report the Pearson correlation coefficient between distances in the causal graph and the embeddings on a held-out part of the causal graph. For the QA task, we report the precision-at-one (P@1), the fraction of test samples where the highest ranked answer is relevant and the mean reciprocal rank (MRR) (Manning et al., 2008), the inverse of the position of the correct answer in our ranking on the held-out question set provided by (Sharp et al., 2015).

4.3 Baselines

We evaluate our faithful embeddings by comparing them against two state-of-the-art approaches described in (Sharp et al., 2016) and (Veitch et al., 2020). cEmbedBi uses a bi-directional model, with the task of predicting the masked cause and effect word tokens. This approach uses separate embeddings for words used as causes and effects. Causal-{BERT,RoBERTa} (Veitch et al., 2020) uses the fine-tuning technique for the binary classification of edge detection, similar to ours, on the pre-trained large-uncased model. We can thus compare the gains we get by incorporating faithfulness conditions on the embeddings in downstream tasks.

Embedding	Distance Correlation			Neighborhood
	Euclidean	Cosine	Quasi-Pseudo	AUC-PR
Gigaword Causal Graph				
cEmbedBi	0.33	0.48	0.52	0.67
Causal-BERT	0.40	0.55	0.61	0.71
Causal-RoBERTa	0.41	0.61	0.66	0.76
Faithful-BERT	0.42	0.63	0.78	0.88
Faithful-RoBERTa	**0.45**	**0.67**	**0.81**	**0.89**
CauseNet from ClueWeb12 web crawl				
cEmbedBi	0.23	0.37	0.34	0.54
Causal-BERT	0.25	0.38	0.39	0.56
Causal-RoBERTa	0.28	0.36	0.47	0.59
Faithful-BERT	0.31	0.41	0.55	0.68
Faithful-RoBERTa	**0.37**	**0.43**	**0.58**	**0.71**

Table 1: Correlation and Neighborhood faithfulness measures of the embeddings trained for both the Gigaword causal graph and ClueWeb12 CauseNet graph.

Embedding	1^{st}-Wasserstein	Mean K-S statistic	Uniform dimensions (1024)
cEmbedBi	0.54	0.54	205
Causal-BERT	0.45	0.43	348
Causal-RoBERTa	0.39	0.38	385
Faithful-BERT	0.31	0.21	541
Faithful-RoBERTa	**0.30**	**0.18**	**574**

Table 2: Uniformity measures on the embeddings learnt for Gigaword Causal Graph.

5 Results

5.1 Faithfulness

As shown in Tables 1 and 2, our Faithful-RoBERTa model outperforms Causal-{BERT, RoBERTa} and cEmbedBi (Sharp et al., 2016) on each of the three properties of faithfulness, namely the neighborhood, uniformity, and distance correlation, by more than 30%. Additionally, we report the correlation for Euclidean and Cosine similarity, despite not using it to optimize at training time. Faithful versions of the BERT and RoBERTa models increase the area under the curve of the precision-recall curve in detecting neighboring nodes of the Gigaword and CauseNet causal graphs by 21-23% and 17-20% respectively. In Figure 2, we present the precision-recall curve when we use the models for ranking causal pairs above non-causal pairs on the SemEval Task 8 tuples (Hendrickx et al., 2007) by varying the distance threshold in the embedding space which outlines the boundary of the neighboring nodes in the causal graph. This increase in accuracy for neighborhood detection indicates that incorporating the constraints during training time with our asymmetric causal embedding distance provides benefits in aligning the contextual embeddings as per the causal graph.

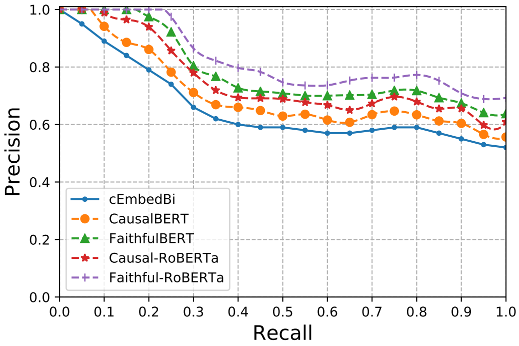

Figure 2: Precision-Recall to detect neighboring nodes in causal graph from the embeddings by applying threshold on distance measure

5.2 QA task

To evaluate if learning faithful embeddings is useful for causal aligned downstream tasks, we evaluate the fine-tuned embeddings to be directly used for question answering. As used in (Fried et al., 2015), we use the maximum, minimum, average distance between words of the question and answer words and the overall distance between the composite question and answer vectors from the embedding. Note that since both cEmbedBi and Causal-{BERT, RoBERTa} are trained with cosine similarity in mind, we use the cosine similarity, but for our Faithful-{BERT, RoBERTa} models, the distance measure used to rank is the quasi-pseudo metric defined in Def 2. We use these 4 features to train an SVM ranker to re-rank candidate answers provided by the candidate retrieval tool (Jansen et al., 2014). We see in Table 3 that Faithful-RoBERTa increases both the precision of the first answer predicted by 10.2%, and the mean reciprocal rank by 10.8%. This means that not only is the first ranked answer more causally correct, but the retrieval of the correct answer in the top-k positions has improved. This improvement in an out-of-domain QA task by aligning the embeddings to an externally available causal graph demonstrates that benefits of faithfulness transfer to downstream tasks.

5.3 Re-alignment towards causation

To understand the reason behind the improved performance, we perform a qualitative inspection of 100 randomly sampled word pairs from the Gigaword causal graph (in Appendix) that are at varying distances in the original pre-trained embedding and trace how they have re-aligned after fine-tuning with the faithfulness objective. We annotate each of these word-pairs as being either causal or not as shown in the confusion matrix with examples in Table 4. In Figure 3, we see re-alignment of these word pairs from association based RoBERTa embeddings to the causally aligned Faithful-RoBERTa embedding space, that is, causal word pairs (blue and orange) move closer, and non-causal word pairs (green and red) move further based on the quasi-pseudo metric d_M. Specifically, the associative but non-causal word pairs (green) have moved further in Faithful-RoBERTa, while the non-associative but causal word pairs (orange) have moved closer. We see that in the cosine-similarity based RoBERTa, the causal word pairs had a mean distance of 0.48, while in the quasi-pseudo metric based Faithful-RoBERTa, the mean distance between the causal word pairs *reduced to 0.28*. The distances are normalized between 0 and 1 based on the maximum and minimum values of distances (cosine or d_M) in the sampled word-pairs.

Embedding	P@1	MRR
cEmbedBi	37.28	46.39
Causal-BERT	38.12	47.26
Causal-RoBERTa	38.74	49.01
Faithful-BERT	39.21	49.72
Faithful-RoBERTa	**41.07**	**51.42**
Ablation Study of Faithful-BERT		
w/o Neighborhood	38.55	48.67
w/o Uniformity	39.01	48.92
w/o Distance Correlation	38.28	48.04
Ablation Study of Faithful-RoBERTa		
w/o Neighborhood	39.69	49.39
w/o Uniformity	40.43	50.06
w/o Distance Correlation	39.50	49.28

Table 3: Performance on the QA task in Yahoo! Answers dataset using the Faithful versions of BERT and RoBERTa incorporating the Gigaword causal graph.

	Cause	**Non-cause**
Associated	rain → flood	accident → fog
Non-Associated	war → epidemic	earthquake → spring

Table 4: Examples of word-pairs chosen to inspect faithfulness over the Gigaword causal graph.

We further analyzed how these associative and causal re-alignments impacted the causal QA task by categorizing the word pairs into three types of variables - mediators, colliders and confounders. **Mediators**: For the question, “What causes a tornado?”, the answer involves “thunderstorms”, which is a mediator caused by “high pressure”. We see that “high pressure” is now much closer to “tornado” in Faithful-RoBERTa than baseline embeddings. **Colliders**: For the question, “What causes persistent cough?”, the colliders “smoking” and “asthma” have moved further based on d_M in

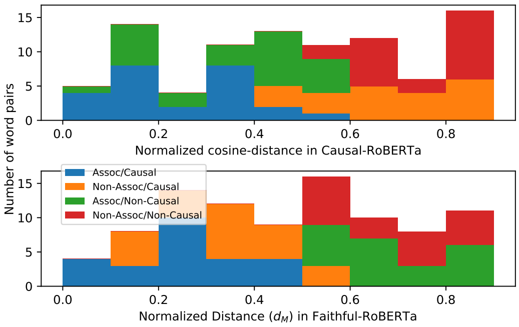

Figure 3: Re-alignment of word-pairs from the causal-RoBERTa embedding to our Faithful-RoBERTa (best viewed in color)

Faithful-RoBERTa. **Confounders**: For questions with confounders like, "What causes indigestion?", the confounding links "anxiety → indigestion", and "anxiety → insomnia" are near, but "insomnia → indigestion", is far. This further demonstrates the utility of incorporating faithfulness over multiple nodes of the graph, in addition to pairwise causal link prediction.

6 Conclusion

We show that the faithfulness of text embeddings to a causal graph is important for causal inference-aligned downstream tasks. By incorporating the three faithfulness properties of neighborhood, uniformity, and distance correlation through regularization constraints while learning embeddings, we improve the precision of the first ranked answer in the causal QA task by 10.2%. We show that this is due to causal re-alignment of embeddings as per an asymmetric pseudo-distance metric.

Acknowledgments

We thank Sam Bowman for his feedback to the draft version of this manuscript.

References

Sami Abu-El-Haija, Bryan Perozzi, and Rami Al-Rfou. 2017. Learning edge representations via low-rank asymmetric projections. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*.

Elias Bareinboim and Judea Pearl. 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352.

Luca Bombelli, Johan Noldus, and Julio Tafoya. 2013. Lorentzian manifolds and causal sets as partially ordered measure spaces.

Stephen Bonner and Flavian Vasile. 2017. Causal embeddings for recommendation. *CoRR*, abs/1706.07639.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. 2016. Multi-level cause-effect systems. In *Artificial Intelligence and Statistics*, pages 361–369.

Mo Chen, Qiong Yang, Xiaoou Tang, et al. 2007. Directed graph embedding. In *IJCAI*, pages 2707–2712.

Zhitang Chen, Kun Zhang, Laiwan Chan, and Bernhard Schölkopf. 2014. Causal discovery via reproducing kernel hilbert space embeddings. *Neural computation*, 26(7):1484–1517.

Robert F. Cohen, Peter Eades, Tao Lin, and Frank Ruskey. 1995. Three-dimensional graph drawing. In *Graph Drawing*, pages 1–11, Berlin, Heidelberg. Springer Berlin Heidelberg.

Wayne W. Daniel. 1990. Kolmogorov-smirnov one-sample test. *Applied Nonparametric Statistics (2nd ed.)*, pages 319–330.

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.

Alessandro Epasto and Bryan Perozzi. 2019. Is a single embedding enough? learning node representations that capture multiple social contexts. *CoRR*, abs/1905.02138.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2020. Causalm: Causal model explanation through counterfactual language models. *CoRR*, abs/2005.13407.

Daniel Fried, Tamara Polajnar, and Stephen Clark. 2015. Low-rank tensors for verbs in compositional distributional semantics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 731–736, Beijing, China. Association for Computational Linguistics.

David Galles and Judea Pearl. 1998. An axiomatic characterization of causal counterfactuals. *Foundations of Science*, 3(1):151–182.

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling document-level causal structures for event causal relation identification. 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 1808–1817, Minneapolis, Minnesota. Association for Computational Linguistics.

Qiaozi Gao, Shaohua Yang, Joyce Chai, and Lucy Vanderwende. 2018. What action causes this? towards naive physical action-effect prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 934–945, Melbourne, Australia. Association for Computational Linguistics.

A. Gordo and F. Perronnin. 2011. Asymmetric distances for binary embeddings. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, page 729736, USA. IEEE Computer Society.

Tanya Goyal and Greg Durrett. 2019. Embedding time expressions for deep temporal ordering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4400–4406, Florence, Italy. Association for Computational Linguistics.

Christopher Grimsley, Elijah Mayfield, and Julia R.S. Bursten. 2020. Why attention is not explanation: Surgical intervention and causal reasoning about neural models. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1780–1790, Marseille, France. European Language Resources Association.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Representation learning on graphs: Methods and applications. *CoRR*, abs/1709.05584.

Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020a. Causenet: Towards a causality graph extracted from the web. In *Proceedings of the 29th ACM International Conference on Information amp; Knowledge Management*, CIKM '20, page 30233030, New York, NY, USA. Association for Computing Machinery.

Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020b. Causenet: Towards a causality graph extracted from the web. In *Proceedings of the 29th ACM International Conference on Information amp; Knowledge Management*, CIKM '20, page 30233030, New York, NY, USA. Association for Computing Machinery.

Iris Hendrickx, Roser Morante, Caroline Sporleder, and Antal van den Bosch. 2007. ILK: Machine learning of semantic relations with shallow features and almost no data. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 187–190, Prague, Czech Republic. Association for Computational Linguistics.

M. R. Hestenes. 1969. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4:303–320.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Peter Jansen, Mihai Surdeanu, and Peter Clark. 2014. Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 977–986, Baltimore, Maryland. Association for Computational Linguistics.

Stanisaw Jastrzebski, Damian Leniak, and Wojciech Marian Czarnecki. 2017. How to evaluate word embeddings? on importance of data efficiency and simple supervised tasks.

Katherine Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344, Online. Association for Computational Linguistics.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Roland Kuhn and Renato De Mori. 1990. Cache-based natural language model for speech recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12:570–583.

Daniel D. Lee and H. Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, page 535541, Cambridge, MA, USA. MIT Press.

Xiang Lisa Li and Jason Eisner. 2019. Specializing word embeddings (for parsing) by information bottleneck. *CoRR*, abs/1910.00163.

Yu Li, Ying Wang, Tingting Zhang, Jiawei Zhang, and Yi Chang. 2019. Learning network embedding with community structural information. In *IJCAI*, pages 2937–2943.

Yujia Li, Oriol Vinyals, Chris Dyer, Razvan Pascanu, and Peter Battaglia. 2018. Learning deep generative models of graphs.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2019a. K-BERT: enabling language representation with knowledge graph. *CoRR*, abs/1909.07606.

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

Zhibin Lu, Pan Du, and Jian-Yun Nie. 2020. Vgcn-bert: Augmenting bert with graph embedding for text classification.

Christopher D Manning, Hinrich Schütze, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Cambridge university press.

Facundo Mémoli, Anastasios Sidiropoulos, and Vijay Sridhar. 2016. Quasimetric embeddings and their applications. *CoRR*, abs/1608.01396.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Bhavana Dalvi Mishra, Niket Tandon, Antoine Bosselut, Wen tau Yih, and Peter Clark. 2019. Everything happens for a reason: Discovering the purpose of actions in procedural text.

Seithuti P Moshokoa. 2005. On completeness of quasi-pseudometric spaces. *International journal of mathematics and mathematical sciences*, 2005(18):2933–2943.

Karthik Narasimhan, Tejas Kulkarni, and Regina Barzilay. 2015. Language understanding for text-based games using deep reinforcement learning. *arXiv preprint arXiv:1506.08941*.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics.

I. Olkin and F. Pukelsheim. 1982. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257 – 263.

Malte Ostendorff, Peter Bourgonje, Maria Berger, Julian Moreno-Schneider, Georg Rehm, and Bela Gipp. 2019. Enriching bert with knowledge graph embeddings for document classification.

Mingdong Ou, Peng Cui, Jian Pei, Ziwei Zhang, and Wenwu Zhu. 2016. Asymmetric transitivity preserving graph embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 11051114, New York, NY, USA. Association for Computing Machinery.

Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics Surveys*.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. *CoRR*, abs/1403.6652.

Margaret E. Roberts, Brandon M. Stewart, and Richard A. Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4):887–903.

Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. 2017. Causal consistency of structural equation models. *arXiv preprint arXiv:1707.00819*.

Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Bernhard Schölkopf. 2019. Causality for machine learning. *CoRR*, abs/1911.10500.

Rebecca Sharp, Peter Jansen, Mihai Surdeanu, and Peter Clark. 2015. Spinning straw into gold: Using free text to train monolingual alignment models for non-factoid question answering. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 231–237, Denver, Colorado. Association for Computational Linguistics.

Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. 2016. Creating causal embeddings for question answering with minimal supervision.

Peter Spirtes, Clark Glymour, and Richard Scheines. 1993. *Causation, Prediction, and Search*, volume 81.

Jiankai Sun, Bortik Bandyopadhyay, Armin Bashizade, Jiongqian Liang, P. Sadayappan, and Srinivasan Parthasarathy. 2018. ATP: directed graph embedding with asymmetric transitivity preservation. *CoRR*, abs/1811.00839.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: large-scale information network embedding. *CoRR*, abs/1503.03578.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily Fox. 2018. Neural granger causality for non-linear time series.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323.

Cunchao Tu, Weicheng Zhang, Zhiyuan Liu, and Maosong Sun. 2016. Max-margin deepwalk: Discriminative learning of network representation. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 38893895. AAAI Press.

Victor Veitch, Dhanya Sridhar, and David M. Blei. 2020. Adapting text embeddings for causal inference.

Petar Velikovi, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Li, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

S.V.N. Vishwanathan, Nicol N. Schraudolph, Risi Kondor, and Karsten M. Borgwardt. 2010. Graph kernels. *Journal of Machine Learning Research*, 11(40):1201–1242.

Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. 2017. Community preserving network embedding. In *AAAI*, volume 17, pages 3298239–3298270.

Zhipeng Xie and Feiteng Mu. 2019. Distributed representation of words in cause and effect spaces. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7330–7337.

Jiaxuan You, Rex Ying, Xiang Ren, William L. Hamilton, and Jure Leskovec. 2018. Graphrnn: A deep generative model for graphs. *CoRR*, abs/1802.08773.

Sendong Zhao, Quan Wang, Sean Massung, Bing Qin, Ting Liu, Bin Wang, and ChengXiang Zhai. 2017. Constructing and embedding abstract event causality networks from text snippets. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 335–344, New York, NY, USA. ACM.

Chang Zhou, Yuqiong Liu, Xiaofei Liu, Zhongyi Liu, and Jun Gao. 2017. Scalable graph embedding for asymmetric proximity.

Shijie Zhu, Jianxin Li, Hao Peng, Senzhang Wang, Philip S. Yu, and Lifang He. 2020. Adversarial directed graph embedding.