Nested State Clouds: Distilling Knowledge Graphs from Contextual Embeddings

Bachelor's Project Thesis

Paul Bricman, s3908194, p.a.bricman@student.rug.nl,
Supervisors: Prof. Dr. Herbert Jaeger

Abstract: Interpretability techniques help ensure the safe deployment of ML models into production by providing practitioners with diverse debugging tools, yet the inner workings of large models remain elusive. In this work, we propose a novel interpretability technique which can be used to distill sparse knowledge graphs from a model's high-dimensional embeddings. This technique, termed Nested State Clouds (NSC), takes advantage of the relative spatial layouts of state clouds in latent space (e.g. "fruit" contextual embeddings appear to engulf "apple" ones). We successfully apply NSC to BERT, and recover an ontology of concepts grounded in the model's latent space.

1 Introduction

In the past years, ML models have been claimed to reach human parity on a range of tasks which were deemed challenging only years prior. For instance, machine translation is deemed on par with human translators on popular language pairs, scientific ML models are more accurate than explicit hand-crafted ones in a growing range of fields, while RL agents have outperformed professionals in multiple video games and industrial control applications.

The resounding success of recent ML work has in large part been attributed to the newly-gained ability of ML models to automatically extract relevant features from input data, rather than making use of hand-crafted features. This has proven an instrumental goal in advancing the state-of-the-art on the vast majority of ML tasks. Concretely, the automatically-derived features are represented through the specific activation patterns of hidden layers as information propagates through the ML model. Besides early layers being able to already extract surface input features, a recurring finding has been the fact that input representations become increasingly abstract with each sequential layer, before collapsing again to the concrete particularities of the output towards the final layers. When such representations are continuous and dense, they are referred to as embeddings.

In the context of our increased reliance on ML models on a societal level, the field of explainable AI (XAI) investigates methods for interpreting the inner workings of such models, which otherwise lack clear internal rules due to being largely trained on raw data. Techniques of this kind help researchers debug such ML systems, ensuring their safe use in practice. A pervasive trade-off in XAI, however, is the conflict between formulating explanations which (1) accurately reflect the actual processing performed by ML models (i.e. functionally-grounded), yet (2) are highly comprehensible and intelligible for humans (i.e. human-grounded). This echoes a constant conflict faced in machine translation, where models should (1) adequately preserve input meaning (i.e. adequacy), while (2) producing a coherent translation in itself (i.e. fluency). Framing explainable AI as a translation task from machine to human representations has proven a useful lens for understanding the obstacles faced by the current work.

Given the central role of input representations in recent ML models, a large body of advances in XAI has focused on distilling high-dimensional embeddings into a form which is cognitively ergonomic for humans. As the embeddings themselves arguably lack meaning when separated from input or output data, the vast majority of such work has specifically attempted to highlight how a certain ML model relates inputs and outputs by means of embeddings.

For instance, prior work has highlighted toxic gender biases in word embeddings (e.g. ”woman” being represented as closer in meaning to ”nurse” than ”programmer” by an ML model), which promptly led to debiasing techniques being developed by the NLP community. Moreover, methods have been developed to construct ontologies from embeddings by means of hierarchical clustering, in order to better understand how the underlying ML model groups concepts together. As another line of work, behavioral and structural probes have been employed to locate the ”site” of various computations (e.g. part-of-speech tagging in NLP), by means of relating embeddings from different layers to cruder external representations (e.g. part-of-speech). Additionally, methods have been suggested to explicitly represent part-whole relations using embedding ”columns”, hinting at future ML models which are partially explainable themselves, even before making use of post-hoc XAI tools.

However, even when interpretability techniques focus on directly relating inputs to outputs, embeddings are often involved as mediators. For instance, input feature explanations highlight which particular aspects of the input data have been most influential in yielding the output. Alternatively, techniques based on counterfactuals and adversarial examples – often confounded in the literature – aim to find marginally different inputs which cause massive changes in output. As a particularly ergonomic family of explanations, methods have also been developed to incentivize models to directly explain themselves in natural language. Finally, another effective technique is based on extracting knowledge graphs by inferring entity-relationship-entity triples directly via (masked) language modeling.

In this context, we extend the existing toolkit of interpretability techniques with a novel approach called Nested State Clouds (NSC). This technique can be used to distill highly-comprehensible knowledge graphs directly from sets of high-dimensional embeddings, with no modality constraints. In other words, given an ML model and an auxiliary dataset, NSC can be used to automatically organize concepts in a part-whole hierarchy which, in large part, reflects the model’s internalized knowledge. As investigated in this paper, this novel interpretability technique can be used to surface learned ontologies from sequence-to-sequence models (e.g. BERT, GPT). However, NSC can also be applied to arbitrary classification models (e.g. ViT), by creating state clouds from individual class member embeddings, as mentioned in the discussion. Given this, an important benefit of NSC is its modality-agnostic nature, in stark contrast to methods which only focus on distilling knowledge graphs from text. By operating with spatial layouts of embeddings, NSC can distill high-dimensional representations which have been abstracted away from particular modalities.

NSC works by first generating a state cloud of contextual embeddings for each given entity, representing the different meanings of the entity in different contexts. For instance, the concept ”fruit” can refer to a host of different objects, depending on the context of use – variation which is captured by distinct contextual embeddings. Next, NSC compactly represents the overarching shape of the resulting state cloud as a high-dimensional ellipsoid, instead of a set of embeddings, which often results in orders-of-magnitude lower memory footprint. While the ellipsoids resemble PCA and SVD outputs, we opted for using conceptors as compact high-dimensional objects due to existing literature investigating meaningful ways of relating them to each other. Specifically, it has been posited – yet never before tested empirically, to the best of our knowledge – that conceptors benefit from an inherent abstraction ordering, a means of comparing two such objects in terms of their level of abstraction. Loosely speaking, a conceptor which spatially engulfs another can be said to be more abstract, as it encompasses a broader region of space than the other one.

After generating state clouds, representing them as conceptor objects, and conducting pairwise comparisons of abstraction, a search algorithm is employed to find a directed graph which accurately represents the estimated relations of abstraction. In contrast to simply constructing a directed graph by adding a new arc for each positive abstraction relation, a search algorithm allows us to better deal with noise. Additionally, the search framing enables us to specify additional ”nice-to-have” properties of the desired output graph. For instance, we penalize high numbers of arcs, parents per node, and children per node, in an attempt to keep the output explanations sparse and highly intelligible. In this, the objective function of the search algorithm essentially provides a "slider" between functionally-grounded and human-grounded explanations.

Attempting to place NSC in the existing landscape of interpretability techniques, we note that our approach can generate global explanations (i.e. holistically describing the model's processing across inferences) which are provided post-hoc (i.e. after training the model). This is in contrast to those XAI techniques which yield local explanations (i.e. describing the way a particular inference unfolds across layers) and those which are provided during the actual training of inherently interpretable models. Beyond this general placement of NSC in the XAI literature, we point out relevant similarities and differences to prior art throughout the paper.

Our contributions are as follows:

- We provide qualitative evidence highlighting the connection between the spatial layout of nested state clouds and the abstraction relation of the concepts they represent;
- We formulate an novel algorithm for flexibly distilling a set of high-dimensional state clouds into a compact directed graph which depicts part-whole relations;
- We draw evidence-based observations on the way individual symbols relate to concepts.

2 Methods

2.1 Model

As the object of our interpretation technique, we chose a pretrained BERT model, due to its widespread use in industry applications and its large number of derivative models (e.g. RoBERTa, ALBERT, distilBERT, etc.). BERT takes in a sequence of subword tokens as input and reconstructs it as output, while generating a unique contextual embedding for each token. Crucially, the same token can be attributed different embeddings in different contexts (e.g. "she" referring to different people). In practice, the contextual embeddings of individual tokens are mean-pooled together to yield an overarching document embedding which enables information retrieval. However, here we focus only on the contextual embeddings of individual tokens or at most short sequences of them which form a noun phrase (e.g. "orange juice").

2.2 Data

As NSC requires an auxiliary dataset for generating state clouds of contextual embeddings, we employ one of the datasets which have been used for training the BERT model, namely BookCorpus. This dataset consists of a variety of public domain books across different genres, and provides many different contexts for tokens to appear in.

We focus our investigation on relating a set of 100 hand-picked concepts to each other. For each concept, we extract all contexts in which they appear verbatim in the dataset. A context is defined as the span of text starting 300 characters before and ending 300 characters after the concept occurence. Additionally, we trim the incomplete beginning and ending sentences (i.e. trailing) from each context, leaving in only complete sentences surrounding the concept occurence.

For each context, we extract the contextual embedding of the concept occurence, obtaining a set of such embeddings for each concept. The cardinality of each set depends on the frequency of occurence of the concept in the dataset. We further filter our set of concepts based on the size of the set of contextual embeddings, eliminating concepts which had fewer occurences than the number of BERT embedding dimensions (i.e. 768). We cover difficulties in handling such noisy state clouds in the discussion.

2.3 Conceptors

From each remaining state cloud representing the set of concept nuances used in the dataset, we derived a conceptor based on the closed-form equation introduced in previous work. Each conceptor is represented through a square matrix whose dimensions match the dimensionality of BERT embeddings (i.e. 768), representing the PCA-like directions spanned by the state cloud, yet mediated by an additional aperture parameter. Additionally, we note that for rich state clouds based on relatively frequent concepts (e.g. "water"), the conceptor representation resulted in three orders-of-magnitude smaller memory footprint compared to storing the original set of contextual embeddings.

$$C(R, \alpha) = R(R + \alpha^{-2}I)^{-1}$$

2.4 Abstraction Ordering

For each pair of conceptors learned from state clouds, we estimate their relation of abstraction based on a heuristic. First, we substract one conceptor matrix from the other. Second, we compute the mean of the eigenvalues the difference matrix. Intuitively, all eigenvalues of the difference matrix are positive if the first conceptor spatially engulfs the other, having higher spread than the second across all dimensions. Conversely, all such eigenvalues are negative if the first conceptor is completely contained by the second across all dimensions. Inevitably, however, the two conceptors will exhibit one such relation across *some* of the 768 dimensions, while simultaneously exhibiting the opposite in other dimensions. Hence, we average the eigenvalues in an attempt to reach a "consensus" opinion on how the two conceptors are related to each other.

$$f(C_1, C_2) = \frac{1}{n} \sum_{i=1}^{n} \lambda_i (C_1 - C_2)$$

2.5 Graph Search

Given the pairwise estimates of abstraction ordering computed before, we conduct a graph search. We specifically note that we search *for* a graph in the space of possible graphs, rather than searching for a path through a given graph. All candidate graphs considered are directed acyclic graphs (DAG), while nodes are identified with concepts, and arcs indicate meronymous relations of abstraction.

We attempt to solve the graph search task through the local search algorithm of simulated annealing. As an initial candidate, we start with a completely disconnected graph, where no concepts are related to each other. This is represented through an adjacency matrix full of null values. Then, we randomly sample a new proposal by randomly mutating the current graph – removing a previous arc or adding a new one. The acceptance probability is informed by a temperature schedule which linearly decreases from one to zero over the course of the search process, encouraging heavy exploration in the first epochs while conducting using an increasingly conservative strategy towards the end.

Algorithm 2.1 Graph Search in NSC

```
$s \Leftarrow 0$ (fully-disconnected graph)
for $k = 0$ to $epochs$ do
    $T \Leftarrow 1 - \frac{k}{epochs}$
    $s_{new} \Leftarrow neighbor(s)$
    if $P(score(D, s), score(D, s_{new}), T) \geq random(0, 1)$ then
        $s \Leftarrow s_{new}$
    end if
end for
```

The objective function which the search algorithm attempts to maximize is a linear combination of the following:

- the number of arcs
- the number of parents per node
- the number of children per node
- the sum of abstraction estimates included in the candidate graph via arcs

$$\begin{aligned} score(D, A) = \\ \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha D_{ij} - \beta A_{ij}) \\ -\gamma \frac{1}{n} \sum_{i=1}^{n} \left| (\sum_{j=1}^{n} A_{ij}) - children \right| \\ -\delta \frac{1}{n} \sum_{j=1}^{n} \left| (\sum_{i=1}^{n} A_{ij}) - parents \right| \end{aligned}$$

The result of the graph search is the final output of NSC: a graph which indicates how the underlying ML model relates concepts by means of contextual embeddings.

3 Results

In this paper, we designed a new interpretability technique which can be used to extract knowledge

Algorithm 2.2 Nested State Clouds

```
for $s$ in $symbols$ do
    $a \Leftarrow contexts(s)$
    $b \Leftarrow cloud(a)$
    $c \Leftarrow conceptor(b)$
end for
for $i, c_i$ in $conceptors$ do
    for $j, c_j$ in $conceptors$ do
        $D_{ij} \Leftarrow f(c_i, c_j)$
    end for
end for
$s_{output} \Leftarrow graph\_search(D)$
```

graphs from state clouds of contextual embeddings. We have noticed that NSC is able to successfully recover commonsense relations of abstraction from raw text data (e.g. ”apple” ISA ”fruit”, ”orange juice” ISA ”juice”). Additionally, we have found that for a limited number of concepts to relate, the graph search is robust with respect to the starting state. Moreover, the legibility terms included in the linear combination which comprise the search objective (e.g. arc count) successfully nudge the search towards relatively sparse outputs. Finally, the graph search history profile exhibits proper foraging behavior, with fast increases in solution quality in the beginning, followed by a more conservative strategy which ends in marginal improvements towards the move to heavy exploitation.

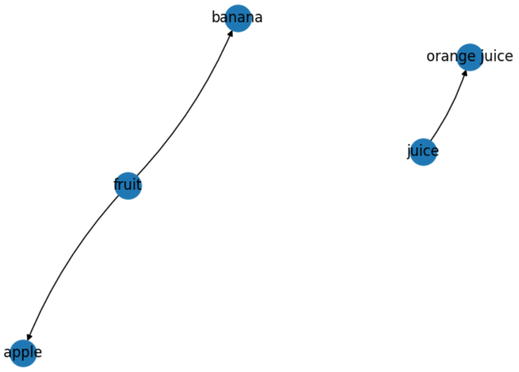

Figure 3.1: NSC output graph when applied to BERT using several symbols.

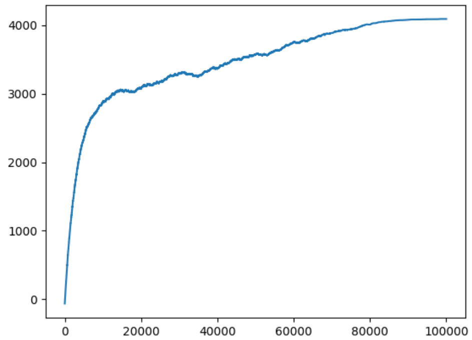

Figure 3.2: Candidate score by epoch during the local graph search.

4 Discussion

4.1 Potential issues

4.1.1 The same symbols can represent different concepts.

Upon inspecting low-dimensional state cloud projections, we observed the presence of distinctive clusters across latent space. For instance, the state cloud of the symbol ”plant” appears to be populated by at least three clusters. To investigate this, we ran a K-means clustering procedure (K=3) procedure on the ”plant” state cloud and surfaced the contexts which yielded contextual embeddings closest to the cluster centroids. Upon inspection of those contexts, we noticed that the context sets contained distinctive word senses, roughly corresponding to (1) plant objects, (2) the action of planting, and (3) factories.

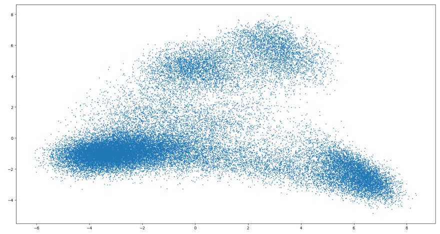

Figure 4.1: 2D PCA projection of the state cloud associated with the symbol ”plant”

This diversity of meanings assumed by the same symbol across the text corpus casts doubt on our assumption of there being a one-to-one correspondence between symbols and graph nodes. An intermediate clustering step might be effective in decoupling different word senses and produce different state clouds, though the issue of how many senses are there per symbol is non-trivial. Similar to how words themselves appear to discretely quantize the otherwise continuous semantic space, finite word senses as ”subsymbols” run into similar trade-offs between sparsity and accuracy.

4.1.2 State clouds are non-linear.

While we employ conceptors as compact elliptical objects which approximate high-dimensional state clouds of contextual embeddings, their limited expressivity might fail to capture the intricate non-linear layout of real-world embeddings. Low-dimensional PCA projections of several state clouds of BERT embeddings radically diverge from Gaussian distributions, bringing into question the suitability of elliptical conceptors to represent them. However, we note that state clouds of less ambiguous terms (i.e. limited number of word senses) appear more well-formed. Non-linearities might arise mainly from diverse word senses being assumed by the same symbols.

Table 4.1: Context samples by K-means cluster of ”plant” state cloud.

Cluster	Sample context
1	absently, i raised the blinds so that the plant was able to soak in the impromptu sunshine.
	i’ve brought you over a few macramé plant hangers to decorate your room.
2	i wanted to plant them myself.
	she’ll just plant new ones and start all over again.
3	the computers running the plant were all infected, of course.
	it was plant shutdown for two weeks.

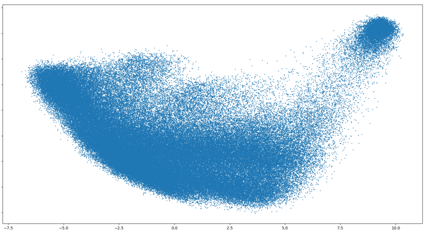

Figure 4.2: 2D PCA projection of the state cloud associated with the symbol ”earth”

4.1.3 NSC requires many exemplars.

The central role of state clouds in NSC means that the technique is highly dependent on a large number of occurences and contexts for each concept analyzed. This makes it difficult to interpret the model’s internal representations with respect to obscure tokens, as those are extremely rare in natural datasets. However, synthetic datasets might address this issue, provided the ability to synthesize a wide range of unique contexts for an arbitrary concept.

4.1.4 NSC output graph is heavily influenced by the graph search objective.

The graph search process employed to output a knowledge graph is highly sensitive to the search objective. Reaching a balance between the functional-grounded terms (e.g. faithfulness to detected abstraction ordering) and human-grounded terms (e.g. sparsity) is difficult to achieve manually. Often, one component of the linear combination tends dominates the others. This balance is especially difficult with higher number of concepts, bringing the scalability of NSC into question. However, normalizing the objective’s components based on the number of concepts being analyzed greatly improved robustness.

4.2 Future work

4.2.1 Non-linear conceptors

Non-linear variations of classical conceptors might be used to capture the internal structure of high-dimensional state clouds of contextual embeddings better than the original elliptical objects. It had

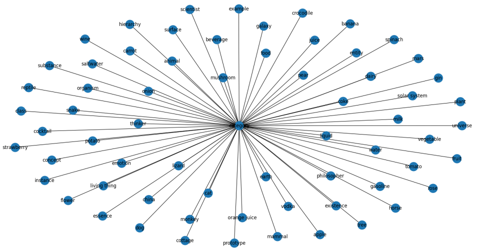

Figure 4.3: NSC output graph after undervaluing target children count per node.

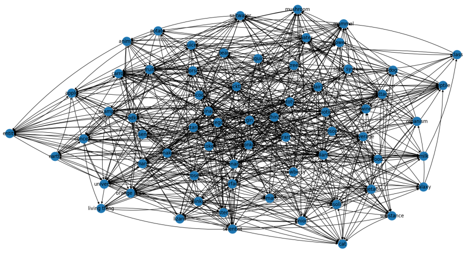

Figure 4.4: NSC output graph after undervaluing arc pruning.

been suggested that shallow MLPs could be employed for this task, yet this brings additional questions of interpretability. As we attempt to model high-dimensional representations ever more accurately, we might be forced to depart from sparse human-grounded explanations.

4.2.2 Beyond abstraction

While this work focuses solely on the meronymous ISA relationship of abstraction between to concepts, it's conceivable that the difference matrix computed in NSC as an intermediate step contains a rich representation of other relationships between concepts. This is reminiscent of the seemingly algebraic properties of prototype word embeddings (e.g. ”king” + ”woman” - ”man” ”queen”), which appear to encode diverse analogies and conceptual relationships. The relative spatial layout of entire state clouds might provide similar information.

References