Convoluted Graph Network Modelling of Natural Language Semantics: An Effort Towards AGI

Harry Prabowo and 13517094[1]
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
[1]13517094@std.stei.itb.ac.id

Approaches towards generally-intelligent digital agent are popularly hypothesized to be an effort towards whole brain emulation of some kind – an extremely processing-power-thirsty process, thus, in many scenarios, is an unrealistic solution. Studies of infant cognitive development indicated language acquisition as a major contributor, hence a very likely candidate for the cognitive framework of a generally intelligent agent. Methods of Natural Language Processing span statistical inference which relies on machine learning to generate robust semantic network. Graph theoretic methods – an elegant representation of entities and relationships – has been widely used in such neural-like computations. Thus theoretically, dynamic integrations of lexical semantics as a linguistic acquisition module for artificial intelligence will have potential to underlie more effective approaches towards a generally intelligent agent.

***Keywords*—graph convoluted network, neural network, lexical semantic, natural language processing, artificial intelligence**

I. Introduction

Popular for being the most complex object in universe, the human brain raises a common opinion among the informatics community which deems any effort of its complete reimaging or modelling to be unrealistic and improbable. Hence, any likely advancement towards AGI has yet to be progressive.

A very common analysis of knowledge acquisition is the observation of infant cognitive development. Through related studies, early stages of language acquisition are deemed crucial a factor in deciding a child's cognitive development, as cognitive perception and conception of early childhood are done through the mimicking and eliciting of linguistic inputs and responses in the form of either verbal or non-verbal communications[1]. Initial effort towards a Natural Language Processing (NLP) agent has seen many inspirations from this phenomenon.

Among the coarse divisions of NLP tasks are the syntax (the structure of language) and semantics (the meaning of words). For the purpose of achieving a general intelligence[4] based on linguistic acquisition, lexical semantics should logically serve as the best feed for a credible curation of semantic rules and ease of usage in computational applications with its lexical form factor. This allows the machine learning agent to look at how meaning of the lexical units correlate with the language syntax, theoretically resulting in a capable syntax-semantic interface[5].

Much of NLP methods, ever since the "statistical revolution"[2], has relied heavily on the power of machine learning. The machine learning paradigm would automatically deduce NLP tasks using statistical inference[3]. However, ever since the increasingly intensive development of deep neural networks, interest has substantially shifted towards newer, fuzzier approaches.

Graphs are ubiquitous in Natural Language Processing (NLP); they are pretty obvious when imagining words in a lexical resource or syntactic relations. They are less obvious, however, when faced with general intelligence problem, such as word disambiguation, sentiment analysis, machine translation, or structure inference and generation.[6] A solid example would be predictive keyboards that come with smartphones. Common workaround for general intelligence problems is to employ an artificial intelligence (AI) in order for a more contextual and predictive analysis. Albeit an AI in name, its implementation is limited to statistics inference of word usage and sentence structure curated online or from user usage data. This means that focus is actually shifted to cleaning poor knowledge representation filled to the brim with mostly inaccurate semantic interface, completely eliminating chances of building any general intelligence.

The nature of lexical form factor generated by NLP through statistical inference is that it is *automatically* a syntax-semantic interface as a product. Understanding this, it is possible to implement said lexical semantics as a cognitive module on which an AGI's neural network can refer to as knowledge representation.

II. Theory

2.1. Natural Language Processing (NLP)

A subfield of computer science which has since its emergence penetrated into its yet increasingly more relevant artificial intelligence brethren, modern NLP widely adopts representation learning and deep neural-network style machine learning methods as of 2010.[8] This is mainly caused by the "statistical revolution" in the late 1980s to mid-1990s.[2] Benefitting from the flurry results that is the characteristics of such methods, NLP can achieve state-of-the-art results in many natural language tasks, like language modelling, parsing, etc.

In some areas, this shift has entailed substantial changes in how NLP systems are designed compared to its early generations which were designed by hand-coding sets of rules (e.g. by writing grammars or devising heuristic rules). Increasingly more NLP applications relies on deep neural network, which emphasizes that it may be viewed as a new paradigm distinct from statistical natural language processing, for example *neural machine translation* (NMT) – as opposed to *static machine translation* (SMT).

Though natural language processing tasks are closely intertwined, they are frequently subdivided into categories for convenience. A coarse division is given below.

a) Syntax (e.g. part-of-speech classification, parsing, etc.)
b) Semantics (e.g. word vectors, lexical network, etc.)
c) Speech (e.g. speech recognition/segmentation, etc.)

Note that, as per named, NLP algorithms are limited to only data in the form of language and its components. However, through the following implementation, the goal is to construct intelligence based on language acquisition as cognitive framework, enabling relational knowledge – on a conceptual level – of data beyond the boundary of language, including non-verbal communication, computer vision, even emotional intelligence & decision making, etc.

2.2. Basic Graph Theory

A graph $G = (V, E)$ is a structure consisting of a set of vertices (or nodes) $V = \{v_i | i = 1, n\}$, some of which are connected through a set of edges $E = \{(v_i, v_j) | v_i, v_j \in V\}$. In a weighted graph $G_w = (V, E, W)$, edges have associated a weight or cost w_{ij}:

$$W = \left\{ w_{ij} \middle| \begin{matrix} w_{ij} \text{ is the weight/cost associated with} \\ \text{edge } (v_i, v_j), w_i, j \in R \end{matrix} \right\}.$$

Edges can be directed or undirected.

When nodes and edges are arranged into a complex graph, what often emerges is a complex community structure. A single graph has varying distinctive features, such as treelike properties, islands, highly clustered neighborhoods, and *highly connected hubs*. These will then be the so-called network topology. Commonly used signature statistical properties are the *average node degree*, the *average path length*, and the *clustering coefficient*. The average node degree, a measure of graph density, is the average number of edges per node. It is calculated by dividing the number of edges by the number of nodes, and then multiplying it by two.

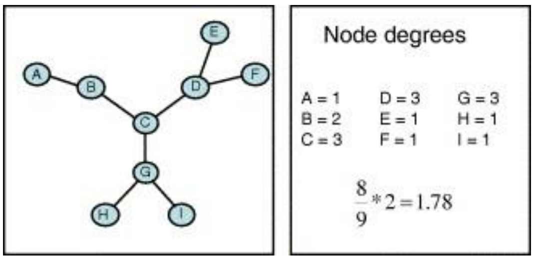

Fig. 1. Average node degree.[16]

The average path length or *average shortest path*, refers to the average distance between two nodes. A simple algorithm is used to determine the minimum distance between any node. An average is then calculated based on these values.

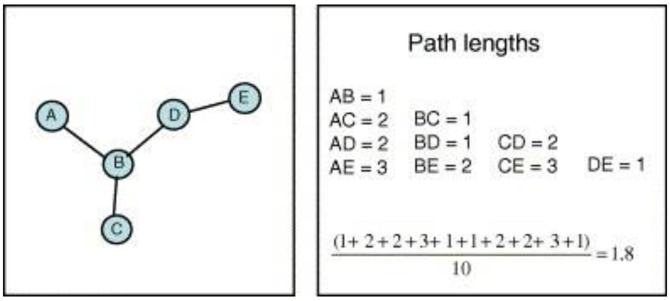

Fig. 2. Average path length.[16]

A common method to calculate the clustering in a graph (at a local level) is to calculate the clustering coefficient for a given node by counting the number of edges between the node's neighbors, and then dividing the value by all their possible edges. This results in a value either 0 or 1, which is then averaged over all nodes in a graph. In a fully connected graph, or the clustering coefficient is 1.

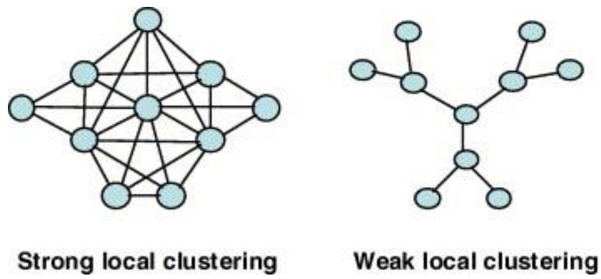

Fig. 3. Components of graphs with strong and weak clustering.[16]

The nodes and vertices, depending on the NLP application, may represent a variety of language-related units and links. Vertices can represent units of different characteristics and quantity, e.g. words, sentences, or even articles. Edges can encode relationships such as co-occurrence, collocations, syntactic structure (e.g. parent and child in a syntactic dependency), or lexical similarity (e.g. cosine between the vector representations of two vertices of the same cluster (e.g. two sentences)).[6]

As the goal is to construct a neural network, it's logical to opt for *heterogenous graph*¸ which vertices are able to correspond to different types of entities, and the edges to different types of links between vertices of the same or different types:

$$V = V_1 \cup V_2 \cup \cdots \cup V_t,$$

with each V_i the set of nodes representing one type of entity.

An example analogous to a heterogenous graph would be a graph consisting of articles, their authors and bibliographic references. Edges between nodes of authors could indicate *co-authorship/collaboration*, edges between nodes of authors and their papers correspond to *authorship*, and links between two papers could represent *citation/reference* relations.[6]

An extension on the notion of graph is the *hypergraph*, with edges – called *hyperedges* – that span an arbitrary number of vertices. $E = \{E_1, \ldots, E_m\}$ with $E_k \subseteq V, \forall_k = 1, m$. When $|E_k| = 2, \forall_k = 1, m$ the hypergraph is a standard graph[7]. The incidence matrix $A(n \times m) = [a_{ik}]$ of a hypergraph associates each row i with vertex v_i and each column k with hyperedge $E_k. a_{ik} = 1$ if $v_i \in E_k$.

As with any graph, a directed hypergraph has directed hyperedges. These are represented as ordered pairs $E_k = (X_k, Y_k)$, where X_k, Y_k are disjoint subsets of vertices, possibly empty. X_k is the *head* of E_k ($H(E_k)$), and Y_k is the *tail* ($\mathrm{T}(E_k)$). The incidence matrix of the hypergraph can encode directionality:

$$a_{ik} = \begin{cases} -1, & v_i \in H(E_k) \\ 1, & v_i \in \mathrm{T}(E_k) \\ 0, & \text{otherwise} \end{cases}$$

An example of a hypergraph in language is the grammar, where the nodes are nonterminal characters and words, and each hyperedge corresponds to a grammatical rule, with the left-hand side of the rule forming the head of the hyperedge, and the body of the rule the tail.[6] Following is the common look to a cluster of processed sentences.

SNo	ID	Text
1	d1s1	Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.
2	d2s1	Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.
3	d2s2	Ramadan told reporters in Baghdad that "Iraq cannot deal positively with whoever represents the Security Council unless there was a clear stance on the issue of lifting the blockade off of it.
4	d2s3	Baghdad had decided late last October to completely cease cooperating with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work became very limited since the fifth of August, and announced it will not resume its cooperation with the Commission even if it were subjected to a military operation.
5	d3s1	The Russian Foreign Minister, Igor Ivanov, warned today, Wednesday against using force against Iraq, which will destroy, according to him, seven years of difficult diplomatic work and will complicate the regional situation in the area.
6	d3s2	Ivanov contended that carrying out air strikes against Iraq, who refuses to cooperate with the United Nations inspectors, "will end the tremendous work achieved by the international group during the past seven years and will complicate the situation in the region."
7	d3s3	Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi weapons of mass destruction (UNSCOM).
8	d4s1	The Special Representative of the United Nations Secretary-General in Baghdad, Prakash Shah, announced today, Wednesday, after meeting with the Iraqi Deputy Prime Minister Tariq Aziz, that Iraq refuses to back down from its decision to cut off cooperation with the disarmament inspectors.
9	d5s1	British Prime Minister Tony Blair said today, Sunday, that the crisis between the international community and Iraq "did not end" and that Britain is still "ready, prepared, and able to strike Iraq."
10	d5s2	In a gathering with the press held at the Prime Minister's office, Blair contended that the crisis with Iraq "will not end until Iraq has absolutely and unconditionally respected its commitments" towards the United Nations.
11	d5s3	A spokesman for Tony Blair had indicated that the British Prime Minister gave permission to British Air Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.

Fig. 4. A cluster of 11 related sentences. [6]

2.3. Neural Network (NN)

Neural networks in this paper refers to Artificial Neural Network (ANN), which is inspired from the biological neural network that constitute a sentient brain – in this case, a brain capable of advanced linguistic ability, namely human brain.

An ANN is based on a collection of units called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like synapses in the brain, can transmit signal from one artificial neuron to another. An artificial neuron that receives signal can process it and proceeds to signal connected artificial neurons. A neuron with label j receiving an input $p_j(t)$ from predecessor neurons consists of the following components:[12]

- an *activation* $a_j(t)$, the neuron's state, depending on a discrete time parameter,
- possibly a *threshold* θ_j, which stays fixed unless changed by a learning function,
- an *activation function* f that computes the new activation at a given time $t+1$ from $a_j(t), \theta_j$ and the net input $p_j(t)$ giving rise to the relation
$$a_j(t+1) = f\left(a_j(t)\right), p_j(t), \theta_j),$$
- and an *output function* f_{out} computing the output from the activation
$$o_j(t) = f_{out}\left(a_j(t)\right).$$

Often the output function is simply the identity function.

An *input neuron* has no predecessor and instead serves as input interface for the whole network. On the other side, an *output neuron* also has no successor and thus serves as output.

The network consists of connections, each connection transferring the output of a neuron i to the input of a neuron j. In this sense i is the predecessor of j and j is the successor of i. Each connection is assigned a weight w_{ij}.[12] Sometimes a bias term added to total weighted sum of inputs to serve as threshold to shift the activation function.[13]

The *propagation function* calculates the *input* $p_j(t)$ to the neuron j from the outputs $o_i(t)$ of predecessor neurons and is typically:[12]

$$p_j(t) = \sum_i o_i(t)\, w_{ij},$$

with a bias value, the above term changes to the following:

$$p_j(t) = \sum_i o_i(t)\, w_{ij} + w_{0j},$$

where w_{0j} is a bias.

That which modifies the parameters of the neural network is the *learning rule*. This is in order for a given input to the network to produce favorable output. This process typically amounts to modifying weights and thresholds of the variables of the network.[12]

In the case of theoretical graph modelling of ANNs for NLP-based AI, a *convolutional neural network* (CNN) is used, as it uses a variation of *multilayer perceptrons* (MLP – a feed forward ANN consisting of at least an input layer, a hidden layer, and an output layer[9]) that requires minimal preprocessing. Such NN utilizes a supervised learning version of backpropagation. In the convolutional layer, there are filters that are convolved with the input. Each filter is equivalent to a weights vector that has to be trained.

CNNs have also been explored in NLP, as it is very effective for various NLP problems and especially achieved great results in semantic parsing[10]. Such supervised deep learning methods were the first to achieve human-competitive performance on certain tasks[11].

Using the CNN as the framework for knowledge representation in our AI, the idea is to implement *clustering algorithm* in order to construct a semantics *graph convolutional network* (GCN) generated by the NLP module from fed data.

The goal is then to learn a function of signals/features on a graph which, with inputs of node features and adjacency matrix, would produce a node-level output.

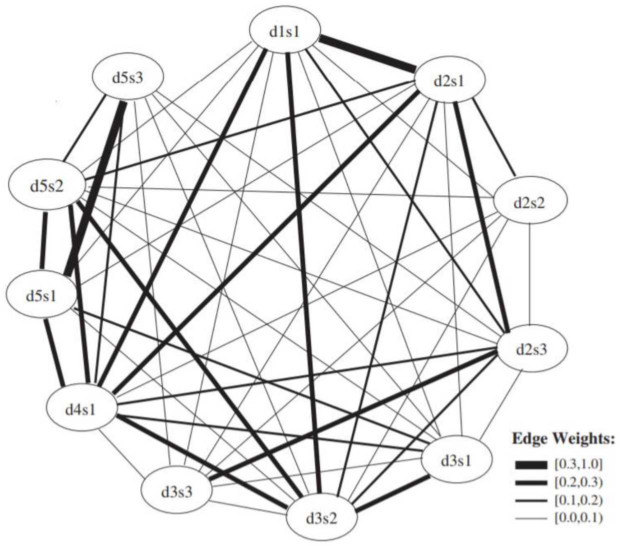

Fig. 5. Weighted cosine similarity graph for the cluster in Figure 4.[6]

2.4. Artificial General Intelligence (AGI)

General intelligence[4] refers to a human-like intelligence. An AGI should be able to perform any intellectual tasks a human being can – also commonly referred as "strong AI". Academic sources reserve the term to refer to machines capable of experiencing consciousness or is deemed to have its own sentience. However, sentience-capable AI is outside the scope of this paper and will not be discussed in detail. Other topics concerning social implications of AGI will also not be covered in this paper.

Albeit there exist out there no fixed criteria to date, but there *is* wide agreement among AI researches that intelligence is required to do the following:[14]

- reason, use strategy, make decisions under uncertainty;
- represent knowledge, including common sense;
- plan;
- learn;
- communicate in natural language;
- and integrate all these skills towards common goals.

Other indicators include ability to sense or act in the world where intelligent behavior is observable, imagination, and even autonomy. Theoretically, an AGI should pass the following.

a. The Turing Test (*Turing*)
b. The Robot College Student Test (*Goertzel*)
c. The Employment Test (*Nilsson*)

Common hypothesis of constructing a theoretical ANN is by complete modelling of the brain, and putting the network into action in a controlled environment – essentially an emulation. This is not only improbable, but also a totally unrealistic approach. Many researches in the field has turned to mimicking the neurons in more detail. This is done, among many ways, by modelling neural networks such as that of ANN after real brain neurons. This, however, ultimately produced not very favorable results. As it appears, modelling the biological structure of the brain does automatically not entail a biological thought process.

Besides data structures, focus should also be allocated to the communication of data inside the structure – in this case, network. Following are additional parameters in the effort to achieve general intelligence like that of a human being.

a. Graph clustering as conceptual intelligence
   (As explained in section 2.1 to 2.3 and shall be further elaborated in section 3).
b. "Forgetfulness" as memory management
   Generated (or generating) ANN will have its own way of "discarding" rarely used nodes either by adding an additional access frequency property to each node, or to have recurring network refresh that removes nodes with weak (read: few) connections and decrements overlapping vertices at each iteration. This mimics how human forgets, over time, trivial stuff easier than important ones.
c. Pattern recognition as a cache
   Among the many ways that exists by which human remembers a memory, pattern retracing is the most effective in most cases. A computational implementation would be an implementation of a separate 'memory' that caches memory access patterns according to access frequency (or possibly depend the caching on the intelligence's personality). This allows instant "reminiscence".

III. Strategy & Analysis

A. Convoluted Graph Network (CGN)

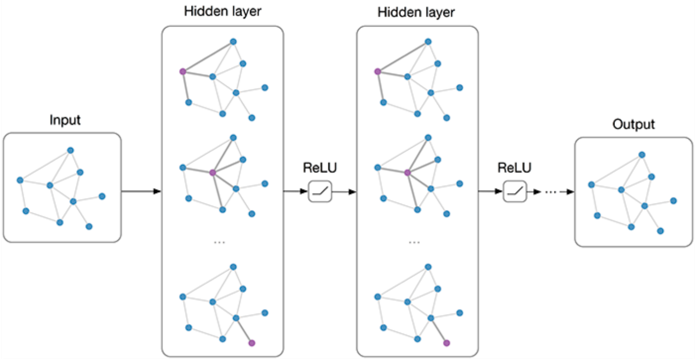

Fig. 6. Multi-layer Graph Convolutional Network (GCN) with first-order filter.
Source: https://tkipf.github.io/graph-convolutional-networks/

A graph $G = (V, E)$ which takes as input:

- A feature description x_i for every node i; summarized in a $N \times D$ feature matrix X (N: number of nodes, D: number of input features)
- A representative description of the graph structure in matrix form; typically in the form of an adjacency matrix A (or some function thereof)

and produces a node-level output Z (an N×F feature matrix, where F is the number of output features per node). Graph-level outputs can be modeled by introducing some form of pooling operation.

As one of the characteristics of CGN, outputs can be either node-level or graph-level. This allows for an expansion from classic CNNs.

Consider the following simple layer-wise propagation rule:

$$f\left(H^{(l)}, A\right) = \sigma\left(AH^{(l)}W^{(l)}\right),$$

where $W^{(l)}$ is a weight matrix for the l-th layer of neural network and $\sigma(.)$ is a non-linear activation function like the ReLU (rectifier linear unit). Despite its simplicity, this model is already powerful, as explained by the following.

Note that multiplication with A means summing up all the feature vectors of all neighboring nodes except the node itself (except for self-loops). The go-around is by enforcing self-loops in the graph, by adding the identity matrix to A.[15]

A bigger concern is that A with unstructured data as a source (especially in cases like NLP), multiplication with A will completely alter the scale of the feature vectors. Normalizing A such that all rows sum to one, i.e. $D^{-1}A$, where D is the diagonal node degree matrix, gets rid of this problem. Multiplying to $D^{-1}A$ corresponds to taking the average of neighboring node features. This is a simple example, as opposed to cases in practice, which dynamics would implement symmetric normalization, i.e. $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$, as this no longer amounts to mere averaging of neighboring nodes. With these two workarounds, we ultimately arrive at the propagation rule introduced in Kiph & Welling (ICLR 2017):[15]

$$f\left(H^{(l)}, A\right) = \sigma\left(\hat{D}^{-\frac{1}{2}}A\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right),$$

with $\hat{A} = A + I$, where I is the identity matrix and $\hat{D}$ is the diagonal node degree matrix of $\hat{A}$.

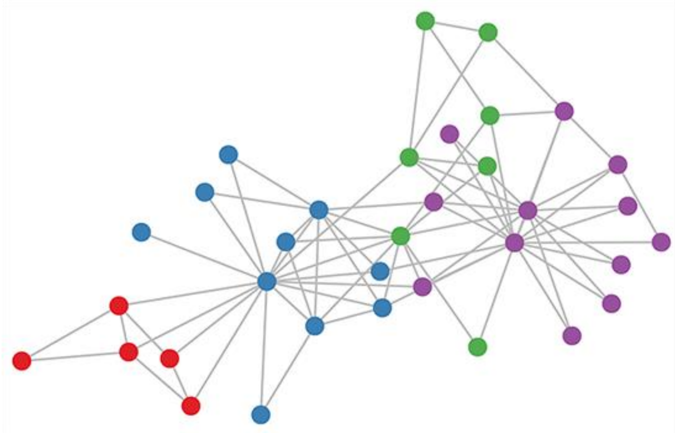

Fig. 7. Karate club graph, colors denote communities obtained via modularity-based clustering (Brandes et al., 2008).

Let's look at the application of this model to a well-known graph dataset: Zachary's karate club network. Taking 3-layer GCN with randomly initialized weights and inserting the adjacency matrix of the graph $X = I$ (assuming there isn't any node features), the model, performing three propagation steps during the forward pass, convolves the 3rd-order neighborhood of every node. This produces a remarkable result that closely resembles the community-structure of the graph (see Fig. 8 below), even with weights initialized completely randomly.

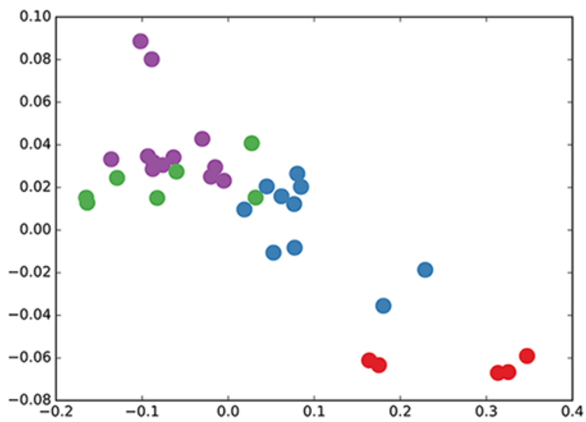

Fig. 8. GCN embedding (with random weights) for nodes in the karate club network.
Source: https://tkipf.github.io/graph-convolutional-networks/

A recent paper on the model DeepWalk (Perozzi et al., KDD 2014) showed that learning a very similar embedding is done through a complicated unsupervised training procedure. Basically, GCN produces results of the same caliber practically "for free" (as in without training the GCN model at all).

Explaining why, it's important to interpret the GCN model as a generalized version of the 1-dimensional Weisfeiler-Lehman algorithm, which works as follows:[17]

For all nodes $v_i \in G$:

- Get features (in the case of this algorithm, it's typically integers, commonly referred to as colors) $\{h_{vj}\}$ or neighboring nodes $\{v_j\}$
- Update node feature $h_{v_i} \leftarrow hash\left(\sum_j h_{v_j}\right)$, where $hash(.)$ is (ideally) an injective hash function.

Repeat k steps or until convergence.

In practice, the Weisfeiler-Lehman algorithm assigns to ever node a feature that uniquely describes its role in the graph. Exceptions are highly regular graphs (i.e. grids, chains, etc.). Otherwise, this feature assignment can be used as a check for graph isomorphism.

The Graph Convolutional layer-wise propagation rule, looking at Weisfeiler-Lehman algorithm (however in a vector format), is as follows:

$$h_{v_i}^{(l+1)} = \sigma\left(\sum_j \frac{1}{c_{ij}} h_{v_j}^{(l)} W^{(l)}\right),$$

where j indexes neighboring nodes of v_i. c_{ij} as a normalization constant for edge (v_i, v_j) which originates from using symmetrically normalized adjacency matrix $D^{-\left(\frac{1}{2}\right)}AD^{-\left(\frac{1}{2}\right)}$ in our GCN model. It is now apparent how this propagation rule is a differentiable and parameterized (with $W^{(l)}$) version of the hash function in Weisfeiler-Lehman algorithm. By choosing proper non—linearity and initialize the weight matrix such that it is orthogonal, this update rule can become stable in practice (partly because of the normalization by c_{ij}). According to this observation, it's possible to produce meaningful smooth embeddings where distance can be well-interpreted as similarity of local graph structures.[15]

B. Semi-supervised Learning

Since everything in the current model is differentiable and parameterized, it's possible to conduct labelling and observe how the embeddings react as the model trains. In the case of application on NLP, the networks produced by the NLP agent will become the feature inputs to the propagation rule, while the knowledge representation formed by the GCN will serve as dynamic neural network on which the NLP can depend on. Not only will the lexical semantic elements be dynamically intertwined with the GCN rule as the embeddings, but this model will also theoretically produce a 2-dimensional latent space we can visualize before it forms a more convoluted cognitive network on higher levels. At the same time, initial node features *could* be provided, which is similar to what's done in the experiments in the paper by Kipf & Welling, ICLR 2017.[15]

Note that while it's optimal to keep the NLP running throughout the learning process, it is important that GCN propagating is reiterated on every terminal input. Otherwise, the semantic network would not benefit from GCN as it provides incomplete relations in the knowledge representation, thus damaging it. This is analogous to the process of reading a sentence to a punctuation before interpreting the meaning behind it, except it's previously interpreted.

C. NLP Liberation

A characteristic common to static statistics inference NLP is that it's unforgiving and rigid, causing an unrealistically steep learning curve, apparent in SMTs. On the other extreme side, neural-network-based NLP, like NMTs, tend to converge into highly convoluted syntax network that mostly make no sense, as its lexical semantics do not possess a knowledge representation. The only supervision control which can be asserted most effectively is by hand-coding rules and grammars, which do not help with the semantics at a fundamental level. This is explainable, as despite the dynamism of the curation method, the nature of the static cognitive product remains unchanged.

Running the GCN classification parallel to the NLP will theoretically produce a cognitive framework very close to that which is acquired biologically through linguistic acquisition as an infant, the supervisor being the "parents". Further control, however, still needs to be asserted upon the output by NLP, as explained by the following.

Notice that the illustrated GCN model is designed to cluster non-linear inputs exceptionally well even in an unsupervised state. This allows for more room for the NLP agent to be repetitive and less rigid. Now, the nature of neural-network-based NLP is actually the most favorable, as it allows for a richer variation of inputs, making way for the GCN to converge very meaningful embeddings.

Also, the nature of GCNs which doesn't heed linearity allows for more than just verbal language as inputs. This opens the opportunity for massively wide range of inputs, ranging from visual to musical, or even inputs outside of human's capability. This is also thanks to the self-evolving cognitive lexical semantic framework; the faulty syntactic interface is no longer relevant.

A very close implementation of this concept would be the semantic role labeling (SRL) by sentence encoding with GNC, a paper by Marcheggiani & Titov, (to appear in) EMNLP 2017, which resulted in best reported scores on the standard benchmark (CoNLL-2009) both for Chinese and English.[18]

D. Optimization Efforts to Approach General Intelligence

By eliminating the need to combine multiple instances of the same node in the NLP agent in different networks (which is normally a headache as the semantic side of things can take a turn for the worst), the GCN model can take advantage of that to even closer mimic real biological brain which in fact does allow this pattern of input. This actually allows for the embeddings to overall imply the property of *conceptual understanding* (the more repetitive the input, the more overlapping graphs are generated, the more CNN generated on the repeated inputs after clustering, the richer and more solid the information or "memory") by allowing highly variable networks to center around different instances of the same nodes. Fusion of identical graphs is also unnecessary as graphs that is in relation to each other will theoretically not have ends – it will, one way or another, combine to become one long connection of *neurons*, which in theory instantiates *reasoning capability*.

Undeniably, this learning process will also generate very large amount of unnecessary "memories" – junks. Fortunately, a similar characteristic is also observable in human learning process. The workaround to this junk management would be the implementation of human's inherent "forgetfulness", which would be in the form of a cluster analysis that runs periodically to locate locally weak clusters that has many loose ends and detaches them from the main strong cluster. In the long run, this allows for an *evolving thought process*, otherwise coarsely termed as *maturation*. On a side note, this allows for AGIs of this characteristic to be friendlier to supervise.

Further improvement is possible by constructing a caching system that stores cache information on access patterns. Note that this system *does not* replace the weights in the GCN if there's any. It is a completely detached system, much like muscle memory or short-term memory retrieval in human's brain. This allows for instant access to frequently used sub-graph. Implementation can be a simple caching system, or a different instance of machine learning may be employed.

IV. Discussion

5.1. Graph theoretic modelling is flexible and is widely used

The numerous publications and implementations are testaments to the multidisciplinary nature of graph modelling. Large statistical graphs such as surveys and analytics have been utilizing modelling techniques with favorable results. The same applies to the new but emerging field of biomedical informatics, in which graph modelling gets increasing attention and importance.

5.2. Graph modelling is intuitive for both humans and machines

As was apparent in the extremely popular implementation, the capability of graph modelling as a knowledge representation technique is undisputable. The connectionist principle of the graph theory is conceptually simple, unlike other forms of knowledge representation techniques. Curation of semantic network is also automatic.

5.3. Network modelling reveals the global & local structure of language

Self-definition of lexical semantic network, a data-driven complex network models empirically displays actual relationship between system entities. Given the recent dramatic increase in computing power, it is possible to study even more complex with even less assumptions.

At a global level, many semantic network models exhibit scale-free and small-world architectures very close to a real-world phenomenon. At a local level, the profiles of network have been showed to be similar across several languages.

5.4. Linguistic acquisition is a promising framework for future endeavors in AGI

It's inherent that an AGI mimic the biological thought process of a human. Provided that cognitive development of an infant's is language-centric, a proper human-like conceptual thinking and reasoning capability is possible to be constructed. This also ensures that supervision isn't a totally foreign subject, as is with parenting

VII. Acknowledgment

This paper becomes a reality with the kind support and help from many individuals. I would like to extend my most sincere thanks to all of them.

Foremost, I want to offer this endeavor to God Almighty for the wisdom, strength, peace of mind, and health he bestowed upon me in order for me to finish writing this paper.

I'd like to express my gratitude towards my family for the encouragement and good wishes which greatly helped me in the completion of this paper.

Additionally, I'm highly indebted to Institut Teknologi Bandung and Sekolah Teknik Elektro & Informatika for their guidance as well as providing necessary information regarding this paper and also for their support in completing the endeavor.

I would to express my special gratitude to my Discrete Mathematics lecturer Dr. Ir. Rinaldi Munir, MT. for allowing me this opportunity and for imparting his knowledge and expertise in this study.

My appreciations also go to my colleagues and friends who have put in the effort to help me out with their abilities.

References

[1] Kim, Yaebin, Byington, Teresea, Infant Language Development, University of Nevada Cooperative Extension, 2015.

[2] M. Johnson, "How the Statistical Revolution Changes (Computational) Linguistics," Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics, 2009.

[3] G. Upton, I. Cook, Oxford Dictionary of Statistics, 2008, OUP. ISBN 978-0-19-954145-4.

[4] Newell & Simon, 1976. This is the term they use for "human-level" intelligence in the physical symbol system hypothesis.

[5] J. Pustejovsky, "The Generative Lexicon," *MIT Press*, 1995.

[6] V. Natase, R. Mihalcea, D. R. Radev, "A Survey of Graphs in Natural Language Processing," *Cambridge University Press*, 2015.

[7] G. Gallo, G. Longo, S. Nguyen, and S. Pallotino, *Directed hypergraphs and applications*. Discrete Applied Mathematics 42, 1993, pp. 177–201.

[8] Y. Goldberg, "A Primer on Neural Network Models for Natural Language Processing". *Journal of Artificial Intelligence Research 57* (2016) 345–420, 2016.

[9] F. Rosenblatt, x. "Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms". *Spartan Books*, Washington DC, 1961.

[10] E. Grefenstette, P. Blunsom, N. de Freitas, K. M. Hermann, "A Deep Architecture for Semantic Parsing," April 2014.

[11] D. Ciresan, U. Meier, J. Schmidhuber, *Multi-column deep neural networks for image classification*. 2012 IEEE Conference on Computer Vision and Pattern Recognition, June 2012, pp. 3642–3649.

[12] A. Zell, "chapter 5.2". *Simulation Neuronaler Netze [Simulation of Neural Networks]* (in German) (1st ed.). German: Addison-Wesley, 1994. ISBN 978-3-89319-554-1.

[13] F. Abbod, "Application of Artificial Intelligence to the Management of Urological Cancer". *The Journal of Urology*, 2007, ch. 178 (4) pp. 1150–1156.

[14] This list of intelligent traits is based on the topics covered by major AI textbooks, including: Russell & Norvig 2003, Luger & Stubblefield 2004, Poole, Mackworth & Goebel 1998 and Nilsson 1998.

[15] T. N. Kipf, M. Welling, *Semi-Supervised Classification with Graph Convolutional Networks*, Conference Paper at ICLR, 2017, arXiv:1609.02907v4 [cs.LG].

[16] M. E. Bales, S. B. Johnson, *Graph Theoretic Modeling of Large-scale Semantic Networks*, Department of Biomedical Informatics, Columbia University, New York, NY: 2005.

[17] B. L. Douglas, *The Weisfeiler-Lehman Method and Graph Isomorphism Testing*, arXiv:1101.5211v1 [math.CO], 2011.

[18] D. Marcheggiani, I. Titov, *Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling*, to appear in EMNLP 2017, arXiv:1703.04826v4 [cs.CL], 2017.

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 3 Desember 2017

Harry Prabowo | 13517094