A Literature Review and Comparative Analysis of Deep Belief Networks

Haress, Ashraf  
Artificial Intelligence Major,  
Faculty of Informatics and Computer Sciences, The British University in EgyptCairo, Egypt  
[ashraf196280@bue.edu.eg](mailto:ashraf196280@bue.edu.eg)

*Abstract*: This paper re-introduces the Deep Belief Network (DBN) model by stating its history, previous work that leads up to its inception, its mechanism, and its applications. The paper then compares different DBN models on a specific application, in order to identify the more suitable model for said application.

Keywords: Deep Belief Networks, Restricted Boltzmann Machines, Explanation, Comparative Analysis

# Introduction

In 1983, Hinton & Sejnowski proposed Boltzmann (BM) & Restricted Boltzmann Machine (RBM) [1]. A Boltzmann Machine is a network model where the nodes are connected in a full mesh topology [2]. It is used to learn important aspects of an unknown probability distribution based on samples from this distribution [3]. Since there are edges (where n is the number of nodes) which is inefficient, the Restricted Boltzmann Machine (RBM) was proposed which excludes edges on the same layer in the network. Figure 1 demonstrates the difference between BM and RBM model.

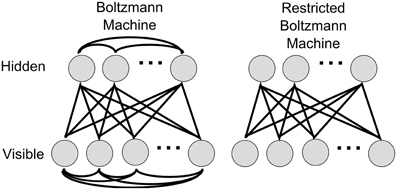


Fig. 1. BM vs RBM regarding connectivity [4]

BMs and RBMs have the same 2 layers: a visible layer, where each neuron corresponds to a component of an observation (i.e. input), and a hidden layer, where each neuron represents the dependency between the components of an observation. Hidden layers are also considered to be non-linear feature detectors [3].

The structure of BM and RBM is an undirected graphical model (UGM), which is also referred as Markov Network, or Markov Random Field (MRF). An MRF is a type of probabilistic graphical model (PGM). The concept of PGM relates to graph and probability theories. In MRF, the random variables represent nodes, and the undirected edges represent the variables which interact (correlate) in some way [5]. Therefore, one can calculate the joint probability of events of a random variable “X” as:

And by assuming we have an undirected graph where is the set of nodes and E is the set of undirected edges, then a clique can be defined as a subset of in which all the nodes of that subset are pairwise connected [6]. For example, in Figure 2, both and are cliques of the set of cliques , but only the latter subset is a maximal clique (which means no node can be added to the latter subset such that it will still remain a clique). Therefore, the latter subset is included in the set of maximal cliques related to graph (expressed as ).

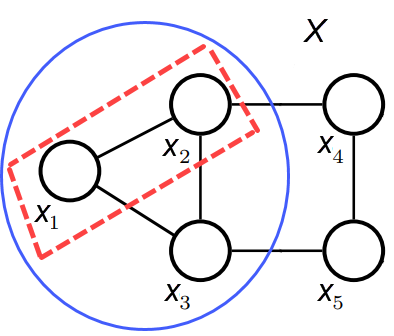


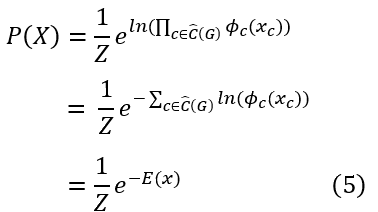
Fig. 2. Clique vs maximal clique. Adapted from [6]

Continuing with the explanation of equation (1), is the potential function applied on each maximal clique that includes an event such that . This potential function does not output the conditional probabilities of events included in a maximal clique , but rather it outputs a non-negative value which represents a “compatibility score” of how much the events of maximal clique correlate. Moreover, these non-negative values do not have a specific range. Therefore, they must be normalised to be in the range [0, 1], and as such (1) can represent a valid probability distribution model. For that normalization to occur, the product of the potential function is divided by — called the partition function, or the normalising constant — which is the sum of the potential values calculated for maximal cliques of the entire network (i.e., for all events in all of . Succinctly [6]:

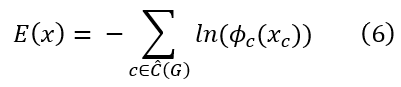
Moreover, since the output of the potential function isn’t restricted to a specific range, the product of such potential functions could be a very large value. Recalling the product property:

And the exponent raised to rule:

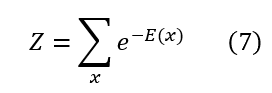
These properties could be utilized to change (1) to:



Such that:



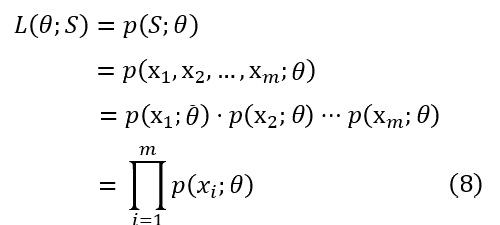
Where is the energy function. Noting that the negative sign was included as the sum of values is negative. Thus, equation (5) represents the probability distribution of MRF, which is also called Gibbs distribution [6]. In addition, we can use (3) and (4) to change (2) into:



Before explaining the RBM training process, it is important to note that the energy function does not have to be as stated in (6), rather it depends on the structure of the model and how we want represent correlation of data.

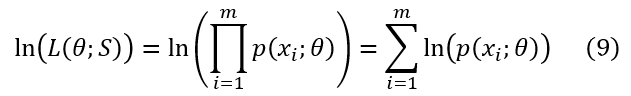
.**State that before we start talking about the RBM training process, the energy function is different for the RBM model than that of MRF such that INSERT EQUATION**.

Now, to understand the RBM training process, MRF unsupervised learning must be understood. Unsupervised learning means learning an unknown distribution based on sample data. Let be the parameters of the energy function that we want to learn in order to model a distribution on the examples of the sample data that resembles the data’s distribution, and let the training data and assumed to be independent and identically distributed (i.i.d.). To estimate the parameters , Maximum Liklihood Estimation (MLE) is used. In the context of MRFs, MLE means finding that will maximize the probability of under ; the MRF distribution [6]. If we denote the likelihood function, which is the joint probability density function of , as , then we have:



Where the semicolon “;” indicates that the symbols that appear after it are parameters of the probability distribution

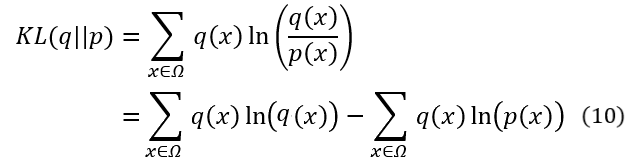
[[1]](#footnote-1). However, following the Gibbs distribution for MRF, it is generally **not possible** to find analytically using gradients [6] for (7) and its log-likelihood equivalent which is given by:



Therefore:

1. Gradient ascent is applied using Constrastive Divergence (CD).
2. CD uses Gibbs sampling to choose the examples (observations) that the CD will train upon from the data set.

Regarding the former point, to understand CD, we have to first discuss Kullback–Leibler (KL) divergence. if we calculate the distance (i.e. difference) between the unknown distribution of the data and the distribution of the MRF given a finite space , then one can say that minimizing that distance is equivalent to maximizing the likelihood of which is given by in (7). The KL divergence is given by:



Therefore, approximating the expectation of to resemble the expectation of by training samples from results in a higher log-likelihood of . That is why maximizing the log-likelihood corresponds to minimizng the KL divergence.

Now, Contrastive Divergence is defined

.**add Z=epower part above and change all numbers, continue with pdf until KL and updating theta, and stop before “Log-likelihood gradient of MRFs with latent variables” then check that , then how pot. Func differs and we’re using energy**.

we have Z which denotes the sum of probabilities of all nodes (x) in X,

To illustrate MN with an example, suppose you In 2006, Hinton tried to revive neural networks after it came to a halt due to the “vanishing gradient” problem, where he proposed a greedy learning method for RBM: Train every layer of the network using RBM training [5]. In other words, he treated the neural network as stacked layers of RBM, where this interpretation of the network is called the Deep Belief Network (DBN). DBN allowed good initialization of weights (using RBM training) which prevented the vanishing gradient problem.

# Ease of Use

## Selecting a Template (Heading 2)

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the Microsoft Word, Letter file.

## Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

# Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

## Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

## Units

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.

Identify applicable funding agency here. If none, delete this text box.

* Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (*bullet list*)

## Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

*a**b* 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”.
* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

# Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**

1. The first term of (7) is asking for the likelihood of the parameters θ while the data *S* are present, but the second term is asking for the probability of observing *S* while the parameters θ are present. [↑](#footnote-ref-1)