Case study - Nichita Utiu 233

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. *Learning deep structured semantic models for web search using clickthrough data*[1]

Mathematics Subject Classification: 68T05 [**Artificial Intelligence: Learning**], 68P20 [**Information Storage and Retrieval**]
Keywords: Deep Learning, Semantic Model, Clickthrough Data, Web Search

The paper presents a way of performing *Latent Semantic Analysis* (LSA) using deep learning models. It starts from a set of textual documents, corresponding queries and the clickthrough rate at which the documents were chosen by users based on these queries. The problem of discovering a *Latent Semantic Model* (LSM) is then approached as a machine learning problem of minimizing the difference between the estimated clickthrough rate given the *condensed semantic representation* and the gold standard rates.

From a mathematical standpoint, LSA aims to find a function mapping from the domain of the documents and queries to a space of low dimensional vectors so that a given similarity metric is maximized for semantically similar texts. The authors note that this problem is often approached as an unsupervised learning one; the paper improving upon this by taking a supervised approach with the aid of the clickthrough data. They cite *Salakhutdinov and Hinton's*[2] paper as the primary method they improve upon.

Each query and document is given as a high-dimensional term vectors($\mathbb{R}^{5*10^5}$) where each value of the vector corresponds to the frequency of a specific term in the document. The dimensions of these vectors are first reduced through a process named *word hashing* to vectors of size $3 * 10^4$. Afterwards, a non-linear projection is applied to the resulting vectors further reducing their size to condensed representations of 128 elements. Given the input vector x after the word hashing process, the non-linear projection denoted y is obtained as follows:

$$l_1 = W_1 x$$
$$l_i = f(W_i l_{i-1} + b_i), i = 2, ..., N$$
$$y = l_N$$

Here, f is defined as the element-wise $tanh$ function and $W_i \in \mathbb{R}^{s_i \times s_{i-1}}$ where s_i represents the size of the i-th vector. A bias term b_i is added to every linear combination to increase model capacity.

Given the above formulas, the semantic similarity between a document D and a query Q is measured as the cosine similarity of their corresponding projections:

$$R(Q, D) = cosine(y_Q, y_D) = \frac{{y_Q}^T y_D}{\|y_Q\| \, \|y_D\|}$$

The similarities between a query and all the documents are fed through a *softmax* function to get the posterior probabilities $P(D|Q)$ and the problem becomes one of minimizing the cross entropy between the actual clickthrough rates and the predicted probabilities. Therefore, the authors, then optimize the parameters through a simple gradient-based approach.

In sections 4 and the authors present the evaluation method and results. The evaluation is done in a 2-fold cross validated manner to tune certain hyperparameters. The datasets consist of 16,510 English language queries and 15 corresponding documents.

The performance is evaluated in terms of Normalized Discounted Cumulative Gain (NDCG)[3]. The method is evaluated against 2 baseline models and 6 contemporary ones. When using the non-linear projection, their model outperforms the current state-of-the-art by 0.19 in terms of score. This performance made this method the current state-of-the-art for LSA at the time of writing the paper.

The bibliography of the paper contains 26 titles: [1], [4], [7], [13], [18], [19], [22], [23], [25] presenting deep learning methods for information extraction, the others being about general semantic representation extraction and most notably, [22] being Hinton's paper on deep autoencoders for semantic analysis as the main source of inspiration.

References

[1] Huang, Po-Sen, et al. *"Learning deep structured semantic models for web search using clickthrough data"*. Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013.

[2] Salakhutdinov, Ruslan, and Geoffrey Hinton. *"Semantic hashing"*. International Journal of Approximate Reasoning 50.7 (2009): 969-978.

[3] Jrvelin, Kalervo, and Jaana Keklinen. *"IR evaluation methods for retrieving highly relevant documents"*. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2000.