

AI浪潮下的學術圖書館

黃明居 | 教授，圖書館館長

112. 09. 21



國立陽明交通大學圖書館

National Yang Ming Chiao Tung University Library



目錄

1.	AI時代的大問題
2.	AI and Machine Learning in Libraries
3.	AI對圖書館的影響
4.	AI應用於圖書館工具 Semantic Scholar & Connected Papers
5.	NYCU Library Future & Conclusion



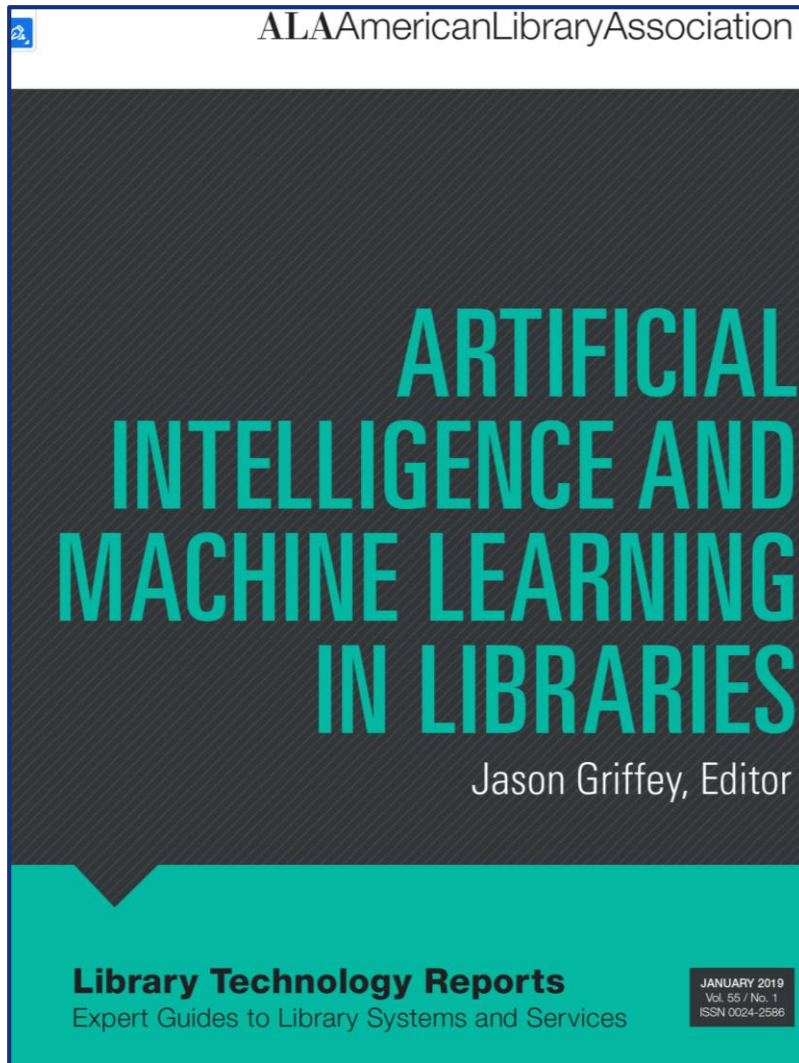


AI時代的大問題

- 當機器能讀取~~(讀懂?)~~圖書館所有資料時，圖書館和館員會發生什麼事（影響）？

“What Happens to Libraries and Librarians When Machines Can Read All the Books?” (Chris Bourg, 2017)

- 圖書館可以提供怎樣的資訊服務？
(Smart Library Services?)
- 圖書館與圖書館館員應如何因應？



AI and ML in Libraries (2019)

- ✓ Introduction
- ✓ HAMLET
- ✓ AI and Creating the First Multidisciplinary AI Lab
- ✓ An Exploration of Machine Learning

Introduction

- in the 1940s, people began to speculate what it would mean for a computer to be “intelligent”
- AI 從符號人工智慧轉向機器學習
 - 人工定義規則 vs. 資料產生規則
 - Ex: AlphaGo (2016)
 - 透過人工神經網路、深度學習進行訓練
- AI & ML 成為現代科技體驗的一部分
 - 人們常常沒有意識到他們正在體驗的是機器學習系統
- 人工智慧優缺點
 - AI的好壞取決於它的訓練數據以及在學習決策時賦予系統的權重
 - 任務自動化、提高效率 ([Homecourt](#)) [HC2](#)
 - 道德、偏見、社會影響

For a given neural net, and a given training set, and a given query, one could build a statistical model of the likelihood of outcomes, but **not predict with certainty what that outcome might be**. This means that **when biases are present in training data**, the effects they might have on queries and outcomes may not be directly predictable.

HAMLET

How About Machine Learning Enhancing
Theses? - a pilot discovery project (MIT, 2017)

<https://hamlet.andromedayelton.com/about/>

- 透過 **機器學習 (Word2vec)** 加強圖書館探索服務
 - 透過 **MIT 研究生論文進行訓練**，用於分析比較文字
 - 允許使用者上傳文件並接收類似論文推薦
 - 透過相關論文建議引用文章來源，協助進行文獻回顧
- 受限於訓練所用資訊
 - 主要來源：MIT 的 STEM 論文
 - 可能不適合其他研究領域，如藝術史或舞蹈
 - 透過相似度與概率分析，準確度或相關性可能有疑慮
- **使用者仍應意識到工具的侷限性**，多元參考其他研究方法

MIT Libraries MLS

How About Machine Learning Enhancing
Theses? - a pilot discovery project (MIT, 2017)

MIT Libraries Machine Learning Studio

Visualizing a department: Physics

If you liked previous posts on concept clusters in aero-astro and chemistry theses, perhaps you will also like to see the physics department!

What do you think the labels for these clusters should be?

click a cluster to see its component theses



Latest Posts

- [Visualizing a department: Physics](#)
- [Visualizing a department: Chemistry](#)
- [Visualizing A Department](#)
- [Six Ways Of Looking At Oxygen](#)
- [Hello Gensim](#)

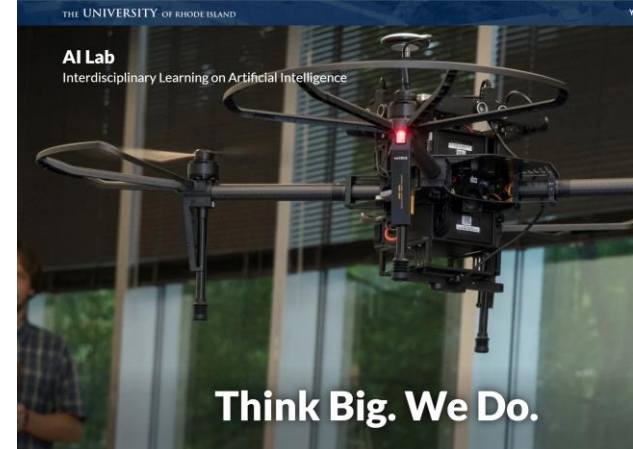
[Archive](#)
[Resources](#)
[Projects](#)

AI and Creating the First Multidisciplinary AI Lab

[AI Lab at the Univ. of Rhode Island](#)

➤ The idea is similar to that of a makerspace in the library

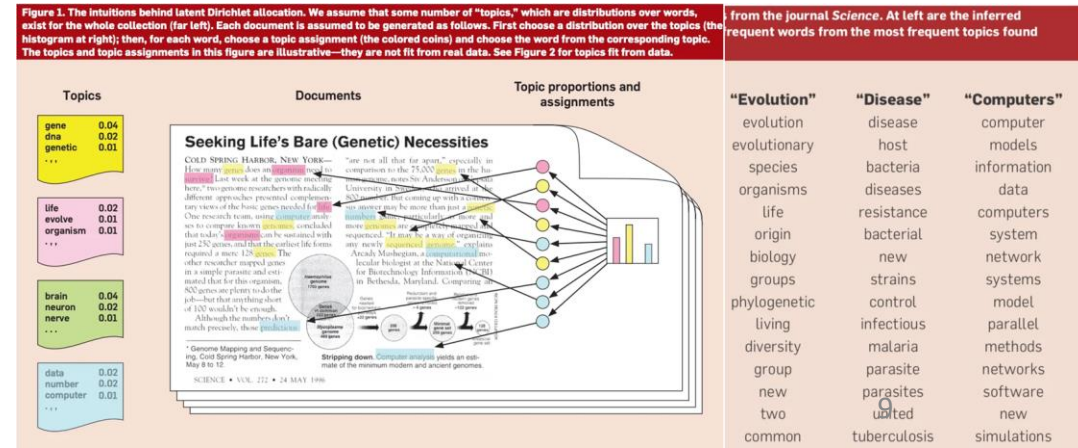
- 核心任務
 - AI 功能教育
 - 促進 AI 研究跨學科合作
 - 促進 AI 社會影響的積極討論
- 推出意義重大
 - 透過產生研究人員間協同作用，讓學校站在發展前沿
 - 與各學科互動合作，讓學生實際使用深度學習演算法來增強穿戴式裝置功能，深入了解如何利用腦電波控制機器人並參與人機關係相關議題討論
- 成為想法與靈感來源、AI 新課程產生器
 - 透過活動提高大眾對 AI 的認識，促進教育推廣
- 透過圖書館的研究中心角色，服務使用者
 - 圖書館的中立性、可訪問性



An Exploration of Machine Learning

Craig Boman

- 研究目標
 - 探索 LDA 技術，產生 [圖書館主題標目](#) (Library Subject Headings (LSH))
 - [Latent Dirichlet Allocation](#)
- 資料處理
 - 以 [Gutenberg 計畫電子書](#) 為基礎
 - 透過 bash 命令檢索、提取電子書資料
 - 透過 bash 和 Python 進行資料轉換與載入
- 透過 [統計學原理](#) 延伸的機器學習預測



AI and ML in Libraries

- 電子書籍或期刊的**提供商**(擁有大量數字化文本語料庫)
 - ✓ 嘗試以AI和ML為基礎的**新索引與搜尋服務**
- 機器學習系統經過訓練，創建Metadata的潛力非常高 (**人工智慧的編目系統**)
 - ✓ 更加注重**培訓數據的準備和產出的評估**，而不是直接創建描述 (**館員的重要角色**)
- **個人化服務**：隨著系統根據讀者的行為進行自我訓練，隨時間的推移，系統會繼續學習
- **研究人員和學生將擁有AI系統**，協助他們尋找資訊、總結資訊並建立個人參考書目等
- 如何對待這些系統的**知識產權**將對圖書館未來如何使用、蒐集、共享和保存等將會有長期影響
- **AI&ML系統值得圖書館和圖書館員密切關注!**
- 當機器人能夠寫出與人類所寫的論文毫無區別時，**教育將如何改變？**

AI對圖書館的影響

- ✓ How AI can improve information organization, accessibility, user services, and library analytics.
- ✓ It also emphasizes the importance of AI literacy for both librarians and patrons in today's society.



5 Ways Artificial Intelligence Impacts Libraries



Information Professionals



Library Operations
Robotic Process Automation
Smart Libraries



User Services



Data and AI Literacy



Library Analytics



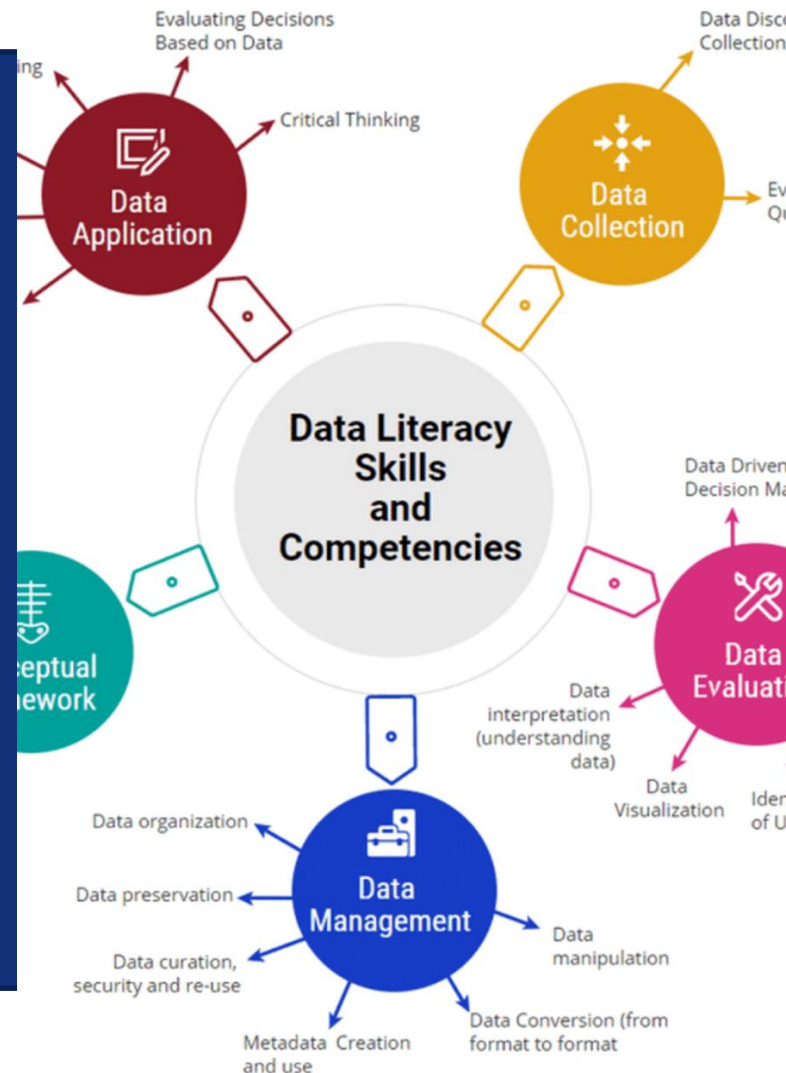
AI對圖書館的影響 (續1)

Information professionals

- ✓ librarians are **improving** the precision of search and recall efforts.
- ✓ librarians will to educate the public on **how to locate and interact with these AI tools.**

Library operations

- Robotic process automation
- Smart Libraries



Data and AI literacy

- ✓ Libraries and library professionals takes aim at **data literacy** and **AI literacy.**
- ✓ Data literacy deals with learning **how to locate, understand, and think critically about data**
- ✓ AI literacy entails an **understanding of its function, logic and limitations, and potential impacts.**

AI對圖書館的影響（續2）

- Library analytics
- ✓ AI can be used to **analyze library data to identify patterns and trends.**
- ✓ This information can be used **to improve library services and make better decisions** about collection development and staffing.
- ✓ **Ex:?**



- User services
- ✓ Providing **reliable and valuable services** tailored to unique user groups
- ✓ AI tools within their library services
- ✓ Ensure more **personalized** and intuitive services.
- ✓ ...



AI應用於圖書館工具

- ✓ Connected Papers
- ✓ Semantic Scholar
- ✓ Hypothesis



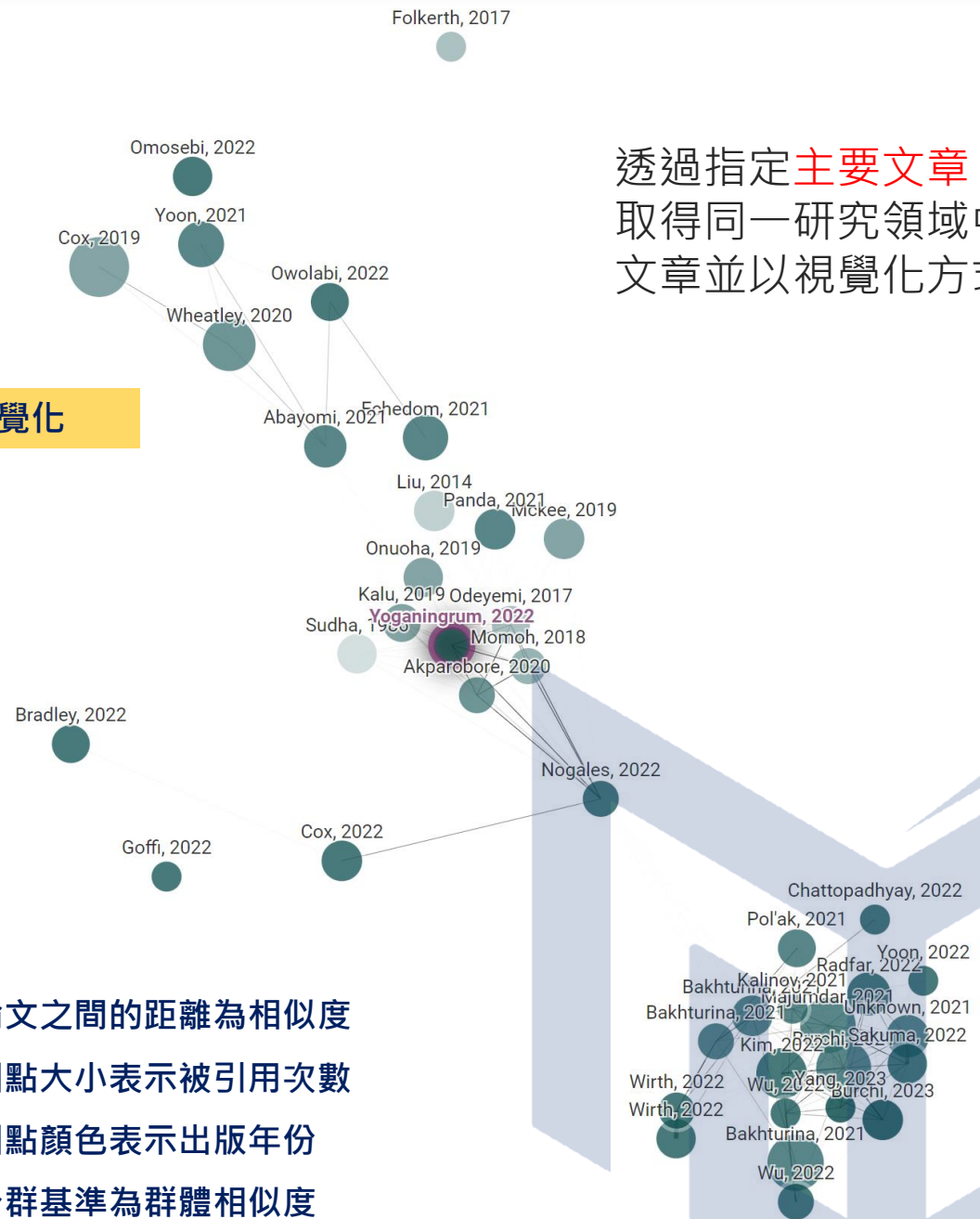
Connected Papers

<https://www.connectedpapers.com/>

學術領域視覺化

1. 論文之間的距離為相似度
2. 圓點大小表示被引用次數
3. 圓點顏色表示出版年份
4. 分群基準為群體相似度

透過指定**主要文章**，
取得同一研究領域中相似
文章並以視覺化方式呈現





Connected Papers

<https://www.connectedpapers.com/>



- 以**原始論文**為出發點，根據**研究領域、關鍵字、主題、引用關係**等基礎因子分析約**5萬**篇文章並選擇與原始論文具**最強關聯性**的數十篇文章並製成關係圖
- 透過同被引（**Co-citation**）與書目耦合（**Bibliographic Coupling**）計算兩兩論文之間的相似度
- 圖中論文以**相似度**進行位置排列，即便無引用關係的兩篇文章也互相連結且放得很近
- 透過力導向圖（**Force Directed Graph**）將相似論文分成一群，並將不太相似的論文推往遠方。與原論文最相近的論文相似度最高，將明顯標記
- 分析資料來源：**Semantic Scholar Paper Corpus**



Semantic Scholar



SEMANTIC SCHOLAR

A free, AI-powered research tool for scientific literature

Search 214,093,134 papers from all fields of science

Search 🔍

Try: [Douglas Thomas Bolger](#) • [Mica](#) • [Group Dynamics](#)

分析文章影響力

透過機器學習分析影響引用次數的因素以及文章之間的動態因子



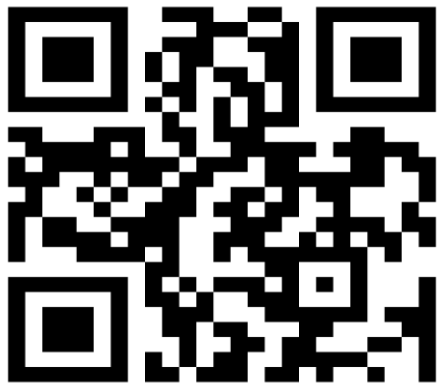
Semantic Scholar

Identifying Meaningful Citations

- ✓ 透過機器文字處理取得每篇文章中的以下資訊，確認**文獻的重要性**，進而計算影響
 - 直接引用總次數
 - 各段落直接引用次數：透過**章節**定義引用來源
 - **非直接**引用總次數、**各段落非直接**引用次數
 - **作者重複狀況**：文章與文獻出現相同作者時為 True
 - 文章認為有用：透過**特定片語**確認文獻有用時為 True
 - 文獻出現在表格或圖表說明中
 - 引用文獻數目倒數：引用文獻越少，單一文獻越重要
 - **直接**引用比例：單一文獻直接引用次數除以所有文獻直接引用次數
 - **計算文章與文獻摘要相似度**
 - 計算文獻 PageRank
 - 計算文獻之間**產生循環**之後，引用同篇文獻的文章數
 - 文獻研究領域



Semantic Scholar



自動產生文章摘要

- ✓ 透過專業背景知識與自然語言技術針對 6000 萬篇文章摘要，以利快速理解內容

AI 驅動的個人化文章推薦

- ✓ 透過 AI 學習個人感興趣文章及新進著作間關係，自動推薦最新內容，維持知識更新

整合式文章閱讀器

- ✓ 透過分析文章與關聯引用文獻，可直接閱讀引用文獻摘要、查閱文章目錄、放入個人收藏並檢視機器自動標記的重點



Semantic Scholar Reader

The screenshot displays the Semantic Reader interface. On the left is a dark sidebar with a 'Table Of Contents' section highlighted in yellow, listing sections from Introduction to References. The main content area shows a search result for the paper 'On the Properties of Neural Machine Translation: Encoder-Decoder Approaches' by Kyunghyun Cho, Bart van Merriënboer, and Yoshua Bengio. The paper's abstract and metadata are visible, with a yellow box highlighting the title and authors. To the right of the text is a diagram (Figure 2) comparing a basic RNN cell and an LSTM memory cell. The RNN cell shows a simple recurrent connection with a tanh activation function. The LSTM cell is more complex, featuring an input gate, a forget gate, and a cell state update mechanism. Below the diagram is its caption. At the bottom of the page, there is a section header '2 BACKGROUND: RECURRENT NETWORKS' followed by introductory text about RNNs and LSTM equations.

Table Of Contents

- 1 Introduction
- 2 Background: Recurrent Networks
- 3 Long-term Recurrent Convolutional Network (LRCN) model
- 4 Activity recognition
 - 4.1 Evaluation
- 5 Image captioning
 - 5.1 Evaluation
 - 5.1.1 Retrieval
 - 5.1.2 Generation
- 6 Video description
 - 6.1 Evaluation
- 7 Related Work
 - 7.1 Prior Work
 - 7.2 Contemporaneous and Subsequent Work
- 8 Conclusion
- References
- Biographies
 - Jeff Donahue
 - Lisa Anne Hendricks
 - Marcus Rohrbach's

On the Properties of Neural Machine Translation: Encoder-Decoder Approaches

Kyunghyun Cho, Bart van Merriënboer, +1 author Yoshua Bengio · SSST@EMNLP · 3 September 2014

TLDR It is shown that the neural machine translation performs relatively well on short sentences without unknown words, but its performance degrades rapidly as the length of the sentence and the number of unknown words increase. Expand

4,069 likes · 736 comments · Save To Library

Section 4): unlike existing labeled video activity datasets may not have actions or activities with particularly complex temporal dynamics, we nonetheless observe significant improvements on conventional benchmarks.

Second, we explore end-to-end trainable image to sentence mappings. Strong results for machine translation tasks have recently been reported [9], [10]; such models are encoder-decoder pairs based on LSTM networks. We propose a multimodal analog of this model, and describe an architecture which uses a visual convnet to encode a deep state vector, and an LSTM to decode the vector into a natural language string (Figure 3 middle; Section 5). The resulting model can be trained end-to-end on large-scale image and text datasets, and even with modest training provides competitive generation results compared to existing methods.

Finally, we show that LSTM decoders can be driven directly from conventional computer vision methods which predict higher-level discriminative labels, such as the se-

Figure 2. A diagram of a basic RNN cell (left) and an LSTM memory cell (right) used in this paper (from [13], a slight simplification of the architecture described in [14], which was derived from the LSTM initially proposed in [7]).

2 BACKGROUND: RECURRENT NETWORKS

urrent neural networks (RNNs, Figure 2, left) model temporal dynamics by mapping input sequences to outputs and hidden states to outputs via the following equations (Figure 2, left):

$$h_t = g(W_{xh}x_t + W_{hh}h_{t-1} + b_h)$$

$$z_t = g(W_{hz}h_t + b_z)$$

element-wise non-linearity, such as a sigmoid or hyperbolic tangent, x_t is the input, $h_t \in \mathbb{R}^N$ is the hidden state, z_t is the output at time t . Given an input sequence (x_1, x_2, \dots, x_T) , the hidden states are computed sequentially as h_1 (letting $h_0 = 0$), z_1, \dots, z_T .

RNNs have proven successful on tasks such as image captioning [15] and text generation [16]. It can be difficult to train them to learn long-term dynamics, in part due to the vanishing and exploding gradients problem that can result from propagating the gradients down through the many layers of the recurrent network, each corresponding to a particular time step. LSTMs provide a solution by incorporating memory units that explicitly allow the network to learn when to “forget” previous hidden states and when to update hidden states given new information. As research on LSTMs has progressed, hidden units with varying connections within the memory unit have been proposed. We use the LSTM unit as described in [11] (Figure 2, right), a slight simplification of the one described in [8], which was derived from the original LSTM unit proposed in [7]. Letting $\sigma(x) = (1 + e^{-x})^{-1}$ be the sigmoid non-linearity which squashes real-valued inputs to a $[0, 1]$ range, and letting $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = 2\sigma(2x) - 1$ be the hyperbolic tangent non-linearity, similarly squashing its inputs to a $[-1, 1]$ range, the LSTM updates for time step t given inputs x_t, h_{t-1} , and c_{t-1} are:



Hypothesis



Bidirectional Encoder Representations from Transformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

1 Introduction

Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018). These include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan and Brockett, 2005), which aim to predict the relationships between sentences by analyzing them holistically, as well as token-level tasks such as named entity recognition and question answering, where models are required to produce fine-grained output at the token level (Tjong Kim Sang and De Meulder, 2003; Rajpurkar et al., 2016).

feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning all pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks, and could be very harmful when applying fine-tuning based approaches to token-level tasks such as question answering, where it is crucial to incorporate context from both directions.

In this paper, we improve the fine-tuning based approaches by proposing BERT: Bidirectional Encoder Representations from Transformers. BERT alleviates the previously mentioned unidirectionality constraint by using a “masked language model” (MLM) pre-training objective, inspired by the Cloze task (Taylor, 1953). The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked

Annotations Settings X

Powered by Hypothesis.is

Public Search Share Help Profile

Showing 1 annotation (and 1 more) Show all (2)

BS0406 Public

we improve the fine-tuning based approaches by proposing BERT: Bidirectional Encoder Representations from [More](#)

B I ” Link Table Σ ≡ ? Preview

We might follow this part.

Add new tags

Post to Public X Cancel

© Annotations can be freely reused by anyone for any purpose.

BS0406 Public 1 min ago

we improve the fine-tuning based approaches by proposing BERT: Bidirectional Encoder Representations from [More](#)

21 ✎ 🗑 ↶ ↷

NYCU Library Value

Aggregation in Library

- 圖書館可檢索近2億筆館藏資源，如何建立平台，引導師生，提供更有系統、更具智慧的「個人化」「精準」資訊服務？
- 提供服務的過程中，同時可保存與再利用這些校園的智慧資產，為主要目的與核心概念

Active Collector and Gate-Opener

For Researchers in Research Life Cycle





讀者的需求

- 面對資訊爆炸的環境，看不完的資料與念不完的文章（資訊焦慮）
- 如何過濾不需要的資料？針對某主題作系統的分析與關聯？Learning Path?
- 是否能(主動)提供可能的研究趨勢？
- 針對主題核心的文章（或圖書）能夠具備「追蹤機制」？（主動提供最新資料）
- ...

How Much Information?

校園智慧資產

- ✓ 師生每天在校園中產生無數的**有形與無形**資料
- ✓ 教學、研究或活動的過程中產生的資料
- ✓ 各式各樣不同的類型資料
(文字, 照片, 影片...)
- ✓ ...



保存與再利用

- ✓ 校園智慧資產如何蒐集, 組織與保存? 如何再利用?
- ✓ 是否有很好的機制, 智慧化地主動蒐集?
- ✓ 從資訊生命週期 (供應鏈) 的概念出發, 建置一整合**智慧資產平台**

Fueling Research Life Cycle Excellence
with Comprehensive Library Services !



支援研究生命週期

服務

以研究生命週期觀點，整合既有資源，

以各種工具加值模式協助各階段工作，

透過平台紀錄校內有形與無形的研究成果與研究歷程



國立陽明交通大學圖書館
National Yang Ming Chiao Tung University Library

AI, all in one @ Research Life Cycle platform 具體範例





結論與討論

- ✓ 無論時代（科技）如何改變，圖書館與館員的**核心價值**仍然不變，圖書館館員仍扮演著重要的角色，但須**與時俱進**
- ✓ 館員是資料正確性的**守門員**，需要**教導讀者資料素養與AI素養**
- ✓ 館員更需**善用AI的工具**，積極改善自己的工作流程、完善規劃與決策品質
- ✓ 利用AI工具，提供讀者**更正確與便利的資訊服務**，協助讀者更有效率的學習與研究，縮短讀者的學習與研究時間



國立陽明交通大學圖書館

National Yang Ming Chiao Tung University Library

謝謝聆聽

Thank you for listening