

基于图论的关系抽取任务

网络空间安全 赵正一

摘要

关系抽取是从一段文本中，在两个命名实体之间，抽取某种预定义好的关系类别的一种工作。关系的主客体双方必须为出现在同一空间内的命名实体，这一空间可以是句子级、段落级甚至是文档级。当给定的文本很长的時候，从这一文本中提取出特定的有关联的主客体以及确定他们之间的关系就成为了这一任务的一大难点。近些年来，关系抽取工作主要聚焦于使用预训练的语言模型如 BERT 模型。这样的工作往往是序列的标注工作，将文本当作长序列，顺序学习提取出命名实体和他们之间的关系。本文通过介绍由图论启发的关系抽取模型，来探索使用图结构在关系抽取工作中的作用。对异构图进行编码，可以将实体之间的关系和实体本身同时进行特征抽取，避免了错误传播。通过实验验证，基于图论的关系抽取模型有较好的效果。

关键字：关系抽取；图论；对话级；BERT 模型

Abstract

Relation extraction is a work of extracting a certain predefined relationship category between two named entities from a piece of text. The subject and object of the relationship must be named entities that appear in the same space, which can be sentence-level, paragraph-level, or even document-level. When the given text is very long, extracting specific related subjects and objects from this text and determining the relationship between them becomes a major difficulty in this task. In recent years, relation extraction work has mainly focused on the use of pre-trained language models such as the BERT model. Such work is often sequence labeling work, which treats the text as a long sequence, and sequentially learns to extract the named entities and the relationship between them. This paper introduces the relation extraction model inspired by graph theory to explore the role of graph structure

in relation extraction work. By encoding heterogeneous graphs, the relationship between entities and the entities themselves can be simultaneously feature extracted, avoiding error propagation. It is verified by experiments that the relationship extraction model based on graph theory has a good effect.

Keywords: Relation Extraction; Graph Theory; Dialog-Level; BERT-based model

目录

1	背景介绍	1
2	相关工作	2
2.1	基于循环神经网络的关系抽取.....	2
2.2	基于图结构的关系抽取.....	2
2.3	基于 BERT 模型的关系抽取	2
3	模型介绍	3
3.1	预备知识	3
3.1.1	形式化表示.....	3
3.1.2	多视图图模型	3
3.1.3	信号与系统.....	3
3.2	BERT 模块	4
3.3	图卷积模块.....	4
4	实验验证	6
4.1	实验设计	6
4.2	结果分析	7
5	总结和展望.....	7

1 背景介绍

给定一段文本和两个命名实体，通过理解文本含义推断出两个命名实体之间的语义关系的工作就被称作是关系抽取。文本段落可以是一句话、一段话或者是一篇文章，甚至也可以是一段对话。

这里使用 TACRED[1] 数据集中的例子来解释关系抽取任务。值得一提的是，在关系抽取任务中，我们关心两个命名实体之间的关系类别，这个类别往往是预定义好的。比如说下面这个例子中：(*Cathleen P. Black*, *chairwoman*) 这两个实体的关系应当为 (*per:title*)。原句为： *Carey will succeed Cathleen P. Black, who held the position for 15 years and will take on a new role as chairwoman of Hearst Magazines, the company said*。如果两个命名实体之间的关系不在预定义好的类别中，那么判定他们之间的关系为无关。

再从 DialogRE 数据集中举例进行说明。DialogRE 数据集是第一个人工标注的适用于对话级的关系抽取数据集，如表 1 所示。对话级关系抽取工作的目标旨在从对话中抽取多轮对话的实体关联信息。

S1: Hey Pheeb.			
S2: Hey!			
S1: Any sign of your brother ?			
S2: No, but he is always late.			
S1: I thought you only met him once?			
S2: Yeah, I did. I think it sounds y'know big sistery, y'know, 'Frank's always late.'			
S1: Well relax, he'll be here.			
	Argument Pair	Trigger	Relation type
R1	(Frank, S2)	brother	per:siblings
R2	(S2, Pheeb)	none	per:alternate names

表 1: DialogRE 数据集示例 [2]

随着技术的发展，图结构的神经网络模型因为它优异的性能而被广泛使用。通常，图中的每一个节点都表示给定文本中的一个命名实体。边的构造方法有很多，通常研究依赖于外部的句法解析树，将文本序列转换成从属的树结构来初始化图结构。也正是因为这个原因，由于句法树的错误提取，可能会导致错误传播的问题发生。

本文介绍了一个基于图论的关系抽取模型用来抽取实体之间的关系。在没有外部知识的情况下，构造一个傅立叶域的图结构，在通过傅立叶域的变换操作进行神经网络学习，最后将产出的结果再通过一个反向的傅立叶变换函数转换到空间域中。其中，图在傅立叶域中的卷积操作的结果在空间域中也得以保留。通过实验验证，基于图卷积神经

网络的模型有着不错的效果。

2 相关工作

本文从基于长短期记忆模块的循环神经网络模型 (LSTM-RNN)、基于图结构模型和基于 BERT 模型的关系抽取方法对相关工作进行了简要综述。

2.1 基于循环神经网络的关系抽取

早期关系抽取的工作高度依赖于人工执行的特征抽取工作 [6][7], 通过摘取不同的特征来表示命名实体。模型高度依赖人工提取的特征的有效性。当前工作主要聚集于学习模型的表示, 比如将循环神经网络 (RNN) 应用到关系抽取工作中去。前人提出的双向长短期记忆模块 (LSTM) [8] 极大的增强了 RNN 对于长距离依赖问题难以解决的困扰。在此基础上, 又有人提出了 PA-LSTM[1] 来额外编码序列中的位置信息, 增强了关系抽取模型的性能。

2.2 基于图结构的关系抽取

基于图结构的模型由于在关系推理中的有效性优势, 在关系抽取工作中得到了广泛应用。有人利用图卷积神经网络 (GCN) [9] 来捕捉远距离依赖的结构信息。有人提出了一个注意力导向的 GCN (AGGCN) 模型 [10], 通过增加自注意力机制来增强图结构表示学习的能力。AGGCN 在句子级的关系抽取工作中表现优异, 但是仍有不少缺点。比如它高度依赖于外部的句法知识信息。错误的句法信息在后续的图表示学习中可能引起错误传播。为了消除这一问题的影响, 有人提出了隐性图学习模型 [11][12]。使用一个端到端的学习方式来学习图结构中的隐藏信息, 这样的学习方式不再需要外部的句法树作为辅助工具。除此之外, 有人还尝试把这一结构看作是一整个变量 [13], 通过对这一变量的学习调整来获取图的特征表示。

2.3 基于 BERT 模型的关系抽取

近年来, 大规模的预训练语言模型在很多任务上都达到了最佳的表现, 其中比较典型的预训练语言模型就是 BERT[17]。基于 BERT 的模型比起基于 RNN 的模型和基于图结构的模型都在不同程度上表现优异 [14][15][2]。基于此, 有人提出了 SpanBERT 模型来学习更好的表示信息 [15], 并且在 TACRED 数据集上表现达到了最优 [1]。TACRED 数据集是一个句子级的关系抽取数据集。对于对话级的关系抽取任务来说, 有人认为将

对话人的个人信息融入到对话本身可以增强 BERT 的表现能力，并提出了 BERTs 模型 [2]。本文也使用了 BERT 模型来提取多视图图结构的节点特征表示。

3 模型介绍

3.1 预备知识

3.1.1 形式化表示

假设 $X = \{x_1, x_2, \dots, x_T\}$ 是一个输入序列，其中 x_t 是序列中的第 t^{th} 个字/词， T 是句子中字词的总数，也是句子的长度。对于句子级别的关系抽取工作而言， X 是输入的句子；对于对话级别的关系抽取工作而言， X 是整段对话。也就是说，对于本文提出的模型而言，不明显区分句子和对话。为了预测两实体之间的关系，本文还定义了两个实体，主体 X_s 和客体 X_o 。主体和客体都是序列中的子序列。一个实体可能含有多个字/词。给定主体、客体和全体输入序列，我们的任务是提取出主体客体之间的关系 $r \in R$ ，其中 R 是所有与定义好的关系类别。

3.1.2 多视图图模型

对于一个多视图的图结构来说，对于同一对节点，存在多个不同的边，这些边并不是在同一张图内形成的。形式化来说，就是对于一个多视图的图 $G = \{V, A^1, A^2, \dots, A^N\}$ 来说， V 是全体节点的集合， A 是一个邻接矩阵， N 是视图的数目。

在本文中，多视图图模型用来用来构建给定文本中复杂的实体间关系。图中的每个节点都对应着一个字/词，这些字词有多种组合的方式，不同的组合方式构建成为不同的视图。通过对不同的视图进行分析编码，就可以最大程度上还原原文本中的语义关系。

3.1.3 信号与系统

信号可以看作是在 n 维空间里的一个向量。这个空间中的向量可以看作是由一组基向量通过合成产生的，也可以说是他们的线性组合：

$$\vec{A} = \sum_{k=1}^N a_k \hat{v}_k \quad (1)$$

当我们想要知道这个线性组合里面每一个不同的元素的 a_k 大小是多少的时候，就要用到信号的分析：

$$a_j = \vec{A} \cdot \hat{v}_j \quad (2)$$

现在我们假定我们的基向量是正定的，常见的基向量的组合就是由 \sin 和 \cos 函数组成的基向量，这一组基向量常用来进行傅立叶变换。现在我们对一个周期性的信号用傅立叶展开：

$$x(t) = \sum_{k=-\infty}^{\infty} a_k e^{jk\omega_0 t} = \sum_{k=-\infty}^{\infty} a_k \phi_k(t) \quad (3)$$

这个函数的基向量就是由 $e^{jk\omega_0 t}$ 组成的，他在方向上的长度由 a_k 决定。通过：

$$\begin{aligned} x(t) &= a_0 + 2 \sum_{k=1}^{\infty} A_k \cos(k\omega_0 t + \theta_k), a_k = A_k e^{j\theta_k} \\ &= a_0 + 2 \sum_{k=1}^{\infty} [B_k \cos k\omega_0 t - C_k \sin k\omega_0 t], a_k = B_k + jC_k \end{aligned} \quad (4)$$

可以求得 a_k 的大小。

同样，不同的基向量可以组成同一个信号，也就是说，同一个函数可以通过不同的方式展开。如果要使用傅立叶变换来展开的话，那么要遵循以下公式：

$$X(j\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt \quad (5)$$

其中， $e^{-j\omega t}$ 是它的基向量，整体可以看作是我们在做信号分析。

3.2 BERT 模块

本文使用 BERT 模型作为特征的编码器，用 BERT 的编码向量来进行后续的抽取工作。选用 BERT 是因为 BERT 在表示学习上的高效性 [15][2]。给定含有 T 个字/词的序列 X ，我们将序列 X 转换成 BERT 模型需要的输入结构 $X_{input} = \{x_0, x_1, x_2, \dots, x_T, x_{T+1}\}$ 。这里的 x_0 表示 $[CLS]$ ，代表着一句话的开始。这里我们所说的一句话并不是狭义上的一个句子，而是送入模型的一个批次 (Batch)。因为模型是一个批次一个批次来进行学习训练的，所以这里为了表述方便写作一句话。这里的 x_{T+1} 代表 $[SEP]$ ，代表着一句话的结束。与输入序列相对应的隐层结果表示为 $H = \{h_0, h_1, h_2, \dots, h_T, h_{T+1}\}$ 。其中， h_0 被当作是每句话的句子级特征向量，被直接送入 SoftMax 分类模型中进行预测类别。除此之外，本文将剩下的隐层向量送入图结构中，通过图节点和边的关系来建立更抽象层次的特征表示。最后，从图中产生的结果会被送入 SoftMax 层和 h_0 一起进行分类。

3.3 图卷积模块

现在给定一个由 V 和 E 来表示的图 $G = (V, E)$, $N = |V|$ ，并给出它的邻接矩阵 $A \in \mathbb{R}^{N \times N}$ ，度矩阵 $D \in \mathbb{R}^{N \times N}$ ，以及一个纯量向量 $f: V \rightarrow \mathbb{R}^N$ 。通过运算，我们可以得到这个图结构的拉普拉斯矩阵 $L = D - A$ ，对于无向图来说，这个拉普拉斯矩阵是一

个对称矩阵。再将这个拉普拉斯矩阵通过对角化处理 $L = U\Lambda U^\top$ 得到它的特征向量矩阵 U 和特征值矩阵 Λ 。其中, $\Lambda = \text{diag}(\lambda_0, \dots, \lambda_{N-1}) \in \mathbb{R}^{N \times N}$ 。特别是, 拉普拉斯矩阵中特征向量之间是正交的。在图卷积模块这里, 我们把特征值称作是“频率”。

通过这一转换, 我们就可以把空间域的向量, 通过傅立叶变换转换到谱空间内, 也叫傅立叶域内。通过谱空间的运算法则, 比如增加自环、有向图变无向图等方式编码, 最终输入到神经网络中进行特征学习。运用图论, 传统的神经网络模型也可以对输入之间的邻居关系进行学习, 我们设置:

$$\begin{aligned}
 f^\top L f &= \sum_{v_i \in V} f(v_i) \sum_{v_j \in V} w_{i,j} (f(v_i) - f(v_j)) \\
 &= \sum_{v_i \in V} \sum_{v_j \in V} w_{i,j} (f^2(v_i) - f(v_i)f(v_j)) \\
 &= \frac{1}{2} \sum_{v_i \in V} \sum_{v_j \in V} w_{i,j} (f^2(v_i) - f(v_i)f(v_j) + f^2(j) - f(v_j)f(v_i)) \\
 &= \frac{1}{2} \sum_{v_i \in V} \sum_{v_j \in V} w_{i,j} (f(v_i) - f(v_j))^2
 \end{aligned} \tag{6}$$

代表两个节点之间的纯量向量差异值, 表示相邻节点的差别关系。

图卷积神经网络是在 ChebNet 的基础上发展来的, 通过简化 ChebNet 的运算法则来达成一个更快的训练过程。图卷积神经网络的目标函数为:

$$\begin{aligned}
 y &= g_{\theta'}(L)x = \sum_{k=0}^K \theta'_k T_k(\tilde{L})x, K = 1 \\
 y &= g_{\theta'}(L)x = \theta'_0 x + \theta'_1 \tilde{L}x \\
 &= \theta'_0 x + \theta'_1 \left(\frac{2L}{\lambda_{max}} - I \right) x \\
 &= \theta'_0 x + \theta'_1 (L - I)x \\
 &= \theta'_0 x - \theta'_1 (D^{-\frac{1}{2}} A D^{-\frac{1}{2}}) \\
 &= \theta (I + D^{-\frac{1}{2}} A D^{-\frac{1}{2}}) x \\
 &= \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)W^{(l)}})
 \end{aligned} \tag{7}$$

之后, 又有工作 [10][16] 探索了多视图训练的方法。使用稠密的图卷积神经网络来捕捉真实的图上结构信息。通过稠密的全连接层, 我们可以训练出一个更深的模型来捕捉局部和全局的信息。多视图的图卷积网络可以写作:

$$v_{n_i}^{(l)} = \rho \left(\sum_{j=1}^T A_{ij}^n W_n^{(l)} k_j^{(l)} + b_n^{(l)} \right) \tag{8}$$

其中, $W_n^{(l)}$ 和 $b_n^{(l)}$ 是训练参数 (权重和偏置值)。 ρ 是一个激活函数, $k_j^{(l)}$ 是全连接过程的初始节点表示向量, 这个向量包含了之前序列所有节点的信息。第一次图卷积工作输出的结果为 $V^1 = \{v_1^1, v_2^1, \dots, v_{T+1}^1\} = \{k_1^{(l)}, k_2^{(l)}, \dots, k_{T+1}^{(l)}\}$ 。

卷积层和池化层在神经网络中往往成对出现并不断循环, 常用的是 SAGPool 池化模型 [19], 计算公式如下:

$$s_{n_i} = \alpha \left(\sum_{j=1}^T A_{ij}^n W_{pool} v_j + b_{pool} \right) \quad (9)$$

之后通过迭代更新就可以用图结构实现关系的抽取。

4 实验验证

本文提介绍的基于图论的关系抽取模型可以应用于句子级和对话级的关系抽取任务。由于选用数据集的数据格式以及适用的基线模型是不同的, 本文通过分析两组实验验证了模型的有效性。以及对模型可能存在的修改空间进行了描述, 以应对不同的任务类别。

4.1 实验设计

DialogRE 数据集是第一个人工标注的适用于对话级的关系抽取数据集 [2], 它包含有 1788 段《老友记》中的对话文本。DialogRE 数据集在此基础上人工定义了 36 种关系类型。表一给出了 DialogRE 中的一个示例。

本文根据最近大火的 BERT 模型 [17], 对本文提出的基于图论的关系抽取模型进行了改进。BERT 是一个预训练的语言模型, 能够感知对话中说话人的信息, 并可以根据将说话人的信息融入到对话本身来提取特征。在对话级的关系抽取任务上有不错的表现。文本同时还同时分析了 CNN 模型, 基于 LSTM 的 RNN 模型, 以保证实验的完整性。

在试验参数的设置上, 采用了相同的输入数据。值得一提的是, 给定的 X , X_s 和 X_o 根据 BERT 的成文规则与 CLS 拼接在一起, 并在每句话的结尾添加 SEP 标签。最终我们得到的序列为:

$$[CLS] - X - [SEP] - X_s - [SEP] - X_o - [SEP]$$

抽取对话的触发词不作特别标注, 与原序列一起进行训练。

除此之外, 优化器选择了 Adam[18]。学习率设置为 $3e - 5$, 池化率设置为 0.7。使用了常见的 SAGPool 池化模型 [19] 作为池化层。由于池化层的不断累加, 最终图中的

节点数量很少，采用传统的 $F1$ 值作为评价指标不具备普适性，所以常采用 $F1c$ [2] 指标作为模型试验的评价指标。输入上，每批次仅将一组对话作为输入，而不是全部对话记录。

4.2 结果分析

表 2 展示了在 DialogRE 数据集上实验的结果。通过这个结果我们可以看出，基于 BERT 模型的方法比基于 CNN 模型的和基于 LSTM 的 RNN 模型的效果有明显的提高。融入了对话人身份信息的 BERT 模型比标准的 BERT 模型效果更好。基于这个想法，本文介绍的基于图论的关系抽取模型是在融入了对话人身份信息的 BERT 模型的基础上进行的扩展。实验结果表明，本文介绍的方法比 BERTs 的效果，在 $F1$ 值上高 5.9%，在 $F1c$ 值上高 1.7%。

Model	$F1$	$F1c$
CNN[20]	48.1%	46.7%
LSTM-RNN[21]	49.7%	46.2%
BERT[17]	66.8%	60.4%
BERTs[2]	73.0%	65.8%
Graph-based	78.9%	67.5%

表 2: 针对 DialogRE 数据集的对比实验

5 总结和展望

本文介绍了一个基于图论的关系抽取模型用来抽取对话级别的实体关系。模型很好地解决了在长序列文本中难以抽取出特定的实体以及实体之间的关系的的问题。在对话级别的关系抽取工作中，通过实验验证，能够达到不错的实验效果。特别的，本文介绍的模型包含有预训练语言模块——BERT、图结构模块——图卷积神经网络。通过图卷积神经网络，模型可以很好的模拟长序列中不同实体之间的可能的关系并做卷积学习。这种机制在处理长序列方面有很大的潜力。

在将来，应当更关注于这一模型在其他级别的数据上的泛化能力以及讨论模型的可解释性。除了关系抽取工作，命名实体识别、事件抽取以及共指解析工作同样是知识抽取工作的其他研究重点，讨论如何讲关系抽取模型泛化到其他三个子任务中也是将来值得探索的重点。

参考文献

- [1] Zhang, Y.; Zhong, V.; Chen, D.; Angeli, G.; and Manning, C. D. 2017. *Position-aware Attention and Supervised Data Improve Slot Filling*. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 35–45. Copenhagen, Denmark: Association for Computational Linguistics.
- [2] Yu, D.; Sun, K.; Cardie, C.; and Yu, D. 2020. *Dialogue-Based Relation Extraction*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4927–4940. Online: Association for Computational Linguistics.
- [3] Tarjan R E. *Graph theory and Gaussian elimination*[J]. Sparse Matrix Computations, 1976: 3-22.
- [4] Bagdanov A D, Worring M. *First order Gaussian graphs for efficient structure classification*[J]. Pattern Recognition, 2003, 36(6): 1311-1324.
- [5] Johnson D, Sinanovic S. *Symmetrizing the kullback-leibler distance*[J]. IEEE Transactions on Information Theory, 2001.
- [6] Miwa, M.; and Sasaki, Y. 2014. *Modeling Joint Entity and Relation Extraction with Table Representation*. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1858–1869. Doha, Qatar: Association for Computational Linguistics.
- [7] Gormley, M. R.; Yu, M.; and Dredze, M. 2015. *Improved Relation Extraction with Feature-Rich Compositional Embedding Models*. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 1774–1784. Lisbon, Portugal: Association for Computational Linguistics.
- [8] Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; and Xu, B. 2016. *Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification*. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 207–212. Berlin, Germany: Association for Computational Linguistics.
- [9] Zhang, Y.; Qi, P.; and Manning, C. D. 2018. *Graph Convolution over Pruned Dependency Trees Improves Relation Extraction*. In Proceedings of the 2018 Conference on Empirical

- Methods in Natural Language Processing, 2205–2215. Association for Computational Linguistics.
- [10] Guo, Z.; Zhang, Y.; and Lu, W. 2019. *Attention Guided Graph Convolutional Networks for Relation Extraction*. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 241–251. Florence, Italy: Association for Computational Linguistics.
- [11] Christopoulou, F.; Miwa, M.; and Ananiadou, S. 2018. *A Walk-based Model on Entity Graphs for Relation Extraction*. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 81–88. Melbourne, Australia: Association for Computational Linguistics.
- [12] Nan, G.; Guo, Z.; Sekulic, I.; and Lu, W. 2020. *Reasoning with Latent Structure Refinement for Document-Level Relation Extraction*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 1546–1557. Online: Association for Computational Linguistics.
- [13] Guo, Z.; Nan, G.; LU, W.; and Cohen, S. B. 2020. *Learning Latent Forests for Medical Relation Extraction*. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, 3651–3657. International Joint Conferences on Artificial Intelligence Organization.
- [14] Wu, S.; and He, Y. 2019. *Enriching Pre-Trained Language Model with Entity Information for Relation Classification*. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19, 2361–2364. New York, NY, USA: Association for Computing Machinery.
- [15] Joshi, M.; Chen, D.; Liu, Y.; Weld, D. S.; Zettlemoyer, L.; and Levy, O. 2020. *SpanBERT: Improving Pre-training by Representing and Predicting Spans*. Transactions of the Association for Computational Linguistics8: 64–77.
- [16] Guo, Z.; Zhang, Y.; Teng, Z.; and Lu, W. 2019. *Densely Connected Graph Convolutional Networks for Graph-to-Sequence Learning*. Transactions of the Association for Computational Linguistics7: 297–312.

- [17] Devlin J, Chang M W, Lee K, et al. *Bert: Pre-training of deep bidirectional transformers for language understanding*[J]. arXiv preprint arXiv:1810.04805, 2018.
- [18] Kingma, D. P.; and Ba, J. 2015. *Adam: A Method for Stochastic Optimization*. In International Conference on Learning Representations
- [19] Lee, J.; Lee, I.; and Kang, J. 2019. *Self-Attention Graph Pooling*. In Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, 3734–3743. Long Beach, California, USA: PMLR.
- [20] Lawrence, S.; Giles, C. L.; Ah Chung Tsoi; and Back, A. D. 1997. *Face recognition: a convolutional neural-network approach*. IEEE Transactions on Neural Networks 8(1): 98–113.
- [21] Hochreiter, S.; and Schmidhuber, J. 1997. *Long short-term memory*. Neural computation 9(8): 1735–1780.