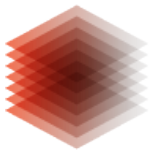
Task 4 @ SimpleText Track 2024:
SOTA? Tracking the State-of-the-Art in Scholarly Publications

Jennifer D'Souza **Salomon Kabongo** **Hamed Babaei Giglou**
Yue Zhang **Sören Auer**



CLEF - **September 21, 2023**

Background

- Leaderboards are scoreboards in AI and related fields, showing the best results achieved by models on specific tasks, datasets, and evaluation metrics. While they have traditionally been community-curated, see the benchmarks feature https://orkg.org/benchmarks on the Open Research Knowledge Graph, their construction could be greatly expedited using text mining.

Application - ORKG benchmarks

Language Modeling with Gated Convolutional Networks

Task

Dataset

Metric

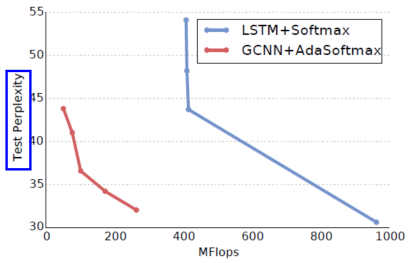
Yann N. Dauphin [1] Angela Fan [1] Michael Auli [1] David Grangier [1]

Abstract

The pre-dominant approach to language modeling to date is based on recurrent neural networks. Their success on this task is often linked to their ability to capture unbounded context. In this paper we develop a finite context approach through stacked convolutions, which can be more efficient since they allow parallelization over sequential tokens. We propose a novel simplified gating mechanism that outperforms Oord et al. (2016b) and investigate the impact of key architectural decisions. The proposed approach achieves state-of-the-art on the WikiText-103 benchmark, even though it features long-term dependencies, as well as competitive results on the Google Billion Words benchmark. Our model reduces the latency to score a sentence by an order of magnitude compared to a recurrent baseline. To our knowledge, this is the first time a non-recurrent approach is competitive with strong recurrent models on these large scale language tasks.

2. Approach

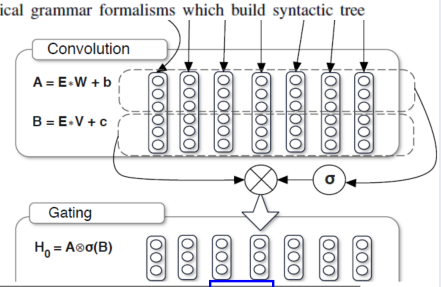
In this paper we introduce a new neural language model that replaces the recurrent connections typically used in recurrent networks with gated temporal convolutions. Neural language models (Bengio et al., 2003) produce a representation $\mathbf{H} = [\mathbf{h}_0, \ldots, \mathbf{h}_N]$ of the context for each word $w_0, \ldots, w_N$ to predict the next word $P(w_i|\mathbf{h}_i)$. Recurrent neural networks f compute $\mathbf{H}$ through a recurrent function $\mathbf{h}_i = f(\mathbf{h}_{i-1}, w_{i-1})$ which is an inherently sequential process that cannot be parallelized over i.[1]



outperform classical *n-gram* language models (Kneser & Ney, 1995; Chen & Goodman, 1996). These classical models suffer from data sparsity, which makes it difficult to represent large contexts and thus, long-range dependencies. Neural language models tackle this issue by embedding words in continuous space over which a neural network is applied. The current state of the art for language modeling is based on long short term memory networks (LSTM; Hochreiter et al., 1997) which can theoretically model arbitrarily long dependencies.

In this paper, we introduce new gated convolutional networks and apply them to language modeling. Convolutional networks can be stacked to represent large context sizes and extract hierarchical features over larger and larger contexts with more abstractive features (LeCun & Bengio, 1995). This allows them to model long-term dependencies by applying $\mathcal{O}(\frac{N}{k})$ operations over a context of size N and kernel width k. In contrast, recurrent networks view the input as a chain structure and therefore require a linear number $\mathcal{O}(N)$ of operations.

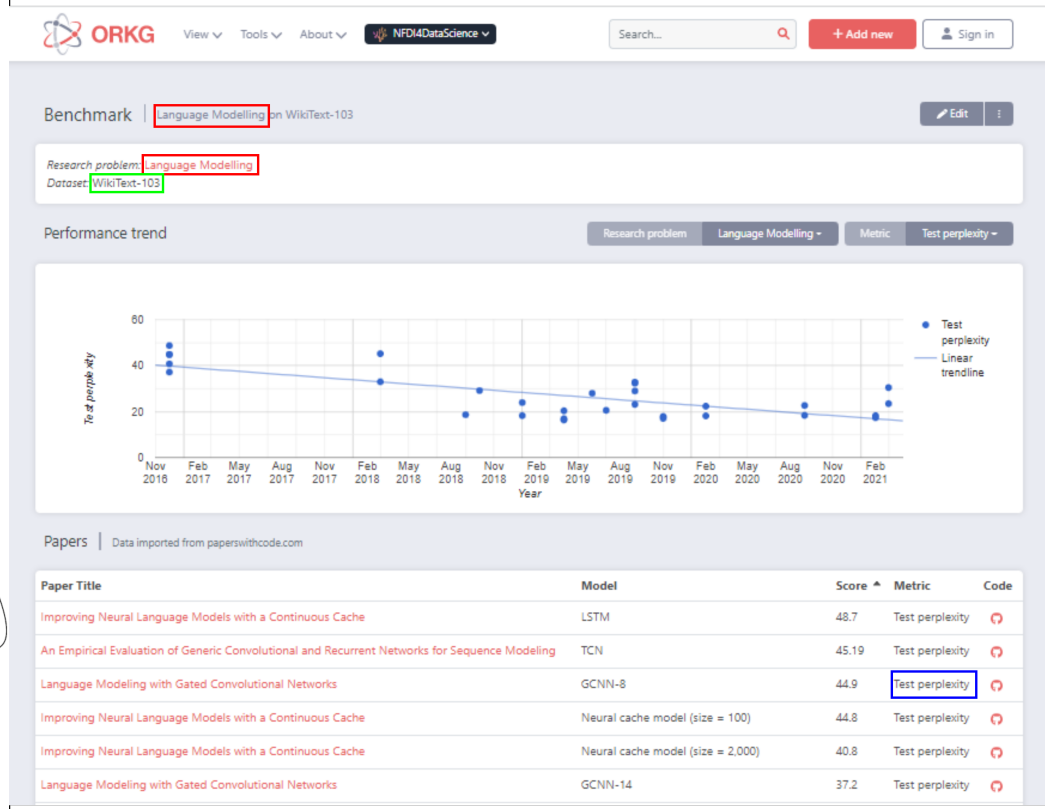
Analyzing the input hierarchically bears resemblance to classical grammar formalisms which build syntactic tree



Model	Test PPL	Hardware
LSTM-1024 (Grave et al., 2016b)	48.7	1 GPU
GCNN-8	44.9	1 GPU
GCNN-14	37.2	4 GPUs

Table 3. Results for single models on the WikiText-103 dataset.

lion Word, the average sentence length is quite short — only 20 words. We evaluate on WikiText-103 to determine if the model can perform well on a dataset where much larger contexts are available. On WikiText-103 an input sequence is an entire Wikipedia article instead of an individual sentence - increasing the average length to 4000 words.



Goals

- We want to create a shift from the traditional method of community curation of Leaderboards, alternatively the state-of-the-art or SOTA. We instead want to establish Leaderboard curation as an automated text mining task based on machine learning.
- The SOTA task itself has the following goals:
  - given an article, the model should first determine whether the article reports a leaderboard or not; and
  - for articles reporting a leaderboard, the model extracts all related (Task, Dataset, Metric, Score) tuples

Our Prior Work: Recognizing Textual Entailment

DocTAET Context Representation Feature

Title:

Deep Recurrent Generative Decoder for Abstractive Text **Summarization** *

Abstract:

We propose a new framework for abstractive text summarization based on a sequence-to-sequence oriented encoder-decoder model equipped with a deep recurrent generative decoder (DRGN). Latent structure information implied in the target summaries is learned based on a recurrent latent random model for improving the **summarization** quality. Neural variational inference is employed to address the intractable posterior inference for the recurrent latent variables. Abstractive summaries are generated based on both the generative latent variables and the discriminative deterministic states. Extensive experiments on some benchmark datasets in different languages show that DRGN achieves improvements over the state-of-the-art methods.

Experimental Setup:

We use **ROUGE** score) as our evaluation metric with standard options For the experiments on the English dataset Gigawords, we set the dimension of word embeddings to 300, and the dimension of hidden states and latent variables to 500 For the dataset of LCSTS, the dimension of word embeddings is 350 The comparison results on the validation datasets of Gigawords and LCSTS are shown in Actually, the performance of the standard The results on the English datasets of Gigawords and DUC-2004 are shown in and respectively In fact, extracting all such features is a time consuming work, especially on large-scale datasets such as Gigawords The results on the Chinese dataset LCSTS are shown in

Table Info:

Table 1: **ROUGE** - F1 on validation sets R - **1** R - **2** R - **L** Table 1 : **ROUGE** - F1 on validation sets Table 2: **ROUGE**-F1 on Gigawords Table 3 : **ROUGE** - Recall on DUC2004 R - **1** R - **2** R - **L** Table 2 : **ROUGE** - F1 on Gigawords Table 3: **ROUGE**-Recall on DUC2004 Table 4 : **ROUGE** - F1 on LCSTS Table 3 : **ROUGE** - Recall on DUC2004 R - **1** R - **2** R - **L** Table 2 : **ROUGE** - F1 on Gigawords Table 4: **ROUGE**-F1 on LCSTS Table 4 : **ROUGE** - F1 on LCSTS R - **1** R - **2** Table 3 : **ROUGE** - Recall on DUC2004 R - **L'**

Leaderboard triples

(**Summarization**; DUC 2004 Task 1; **ROUGE-L**)
(**Summarization**; Gigaword; **ROUGE-1**)
(**Summarization**; DUC 2004 Task 1; **ROUGE-1**)
(**Summarization**; DUC 2004 Task 1; **ROUGE-2**)
(**Summarization**; Gigaword; **ROUGE-L**)
(**Summarization**; Gigaword; **ROUGE-2**)

Kabongo, S., D'Souza, J., Auer, S. (2021). Automated Mining of Leaderboards for Empirical AI Research. In: Ke, HR., Lee, C.S., Sugiyama, K. (eds) Towards Open and Trustworthy Digital Societies. ICADL 2021. Lecture Notes in Computer Science(), vol 13133. Springer, Cham. https://doi.org/10.1007/978-3-030-91669-5_35 Best Paper Award

Our Prior Work: Zero-shot Evaluation Results

- Zero-shot results for leaderboard extraction as an RTE task for the two best models, viz. ORKG-TDM$_{Bert}$ and ORKG-TDM$_{XLNet}$ given unseen leaderboards in training.

	Macro P	Macro R	Macro F1	Micro P	Micro R	Micro F1
			ORKG-TDM$_{Bert}$			
Fold-1	20.1	83.4	28.9	14.1	72.9	23.6
Fold-2	16.2	89	24.4	10.4	81.7	18.4
Average Fold 1 and Fold 2	18.2	86.2	26.7	12.3	77.3	21.0
			ORKG-TDM$_{XLNet}$			
Fold-1	14.3	86.6	21.9	9.2	78.1	16.5
Fold-2	14.9	86.4	22.7	10.1	76.8	17.8
Average Fold 1 and Fold 2	14.6	86.5	22.3	9.7	77.5	17.2

Kabongo S, D'Souza J, Auer S. Zero-shot Entailment of Leaderboards for Empirical AI Research. arXiv preprint arXiv:2303.16835. 2023 Mar 29. Accepted to JCDL 2023.

Moving Forward

- By releasing SOTA as a shared task, we hope to attract models that attempt the task via innovative and novel task formulations. E.g., as a sequence-to-sequence text generation task given Large Language Models (LLMs).

Planned Task Organization

- 1st Stage: Training Dataset Release.
  - Participants will be provided with approximately 5000 articles in xml format. A portion of the articles will be accompanied with (Task, Dataset, Metric, Score) tuple annotations. While another portion of the articles that do not report leaderboards will have no accompanying annotations.
- 2nd Stage: Test Dataset Releases w.r.t two Evaluation Settings.
  - **Few-shot.** A test dataset of scholarly articles will be released and participants will be expected to apply their models on this data. The (TDMS) annotations will be hidden from the participants and used in a blind evaluation. In this few-shot setting, the (TDMS)'s will be only those seen in training.
  - **Zero-shot.** Another unique test dataset of scholarly articles will be released. Again the (TDMS) annotations will be hidden from the participants. This test set will include articles (TDMS) with unseen T, D, or M in training.

Thank you !

Join us at SimpleText 2024 !

Website : https://simpletext-project.com

E-mail : contact@simpletext-project.com

Twitter : https://twitter.com/SimpletextW

Google group : https://groups.google.com/g/simpletext