| **Towards Energy-Preserving Natural Language**  **Understanding With Spiking Neural Networks** | Artificial neural networks have shown promising re-  sults in a variety of natural language understanding (NLU) tasks.  Despite their successes, conventional neural-based NLU models are  criticized for high energy consumption, making them laborious to  be widely applied in low-power electronics, such as smartphones  and intelligent terminals. In this paper, we introduce a potential  direction to alleviate this bottleneck by proposing a spiking encoder.  The core of our model is bi-directional spiking neural network  (SNN) which transforms numeric values into discrete spiking sig-  nals and replaces massive multiplications with much cheaper addi-  tive operations. We examine our model on sentiment classification  and machine translation tasks. Experimental results reveal that our  model achieves comparable classification and translation accuracy  to advanced TRANSFORMER baseline, whereas significantly reduces  the required computational energy to 0.82%. | Recent neural networks rely on massive multiplication op-  erations on float values during inference time. For exam-  ple, TEXTCNN [15] uses cross-correlation operation to compute  the similarity of two inputs, and TRANSFORMER [1] model  conducts the scaled dot-product attention for alignments. Sev-  eral model compression approaches have been proposed to  reduce the computational complexity via decreasing the model  throughput, i.e. dimensionality. He et al. [9] suggested remov-  ing redundant features in hidden states for eliminating useless  calculations. Li et al. [13] proposed to binarize gates in the  recurrent neural networks to accelerate the model inference. As a  representative method, Hinton et al. [8] introduced a knowledge  distillation (KD) scheme, which transfers useful information  from a heavy teacher network to a portable student model.  Jiao et al. [6] successfully exploited KD to pre-train a smaller  language model for downstream tasks. Nevertheless, all those  techniques compress existing models and restrict the model  throughput, rather than directly solve the root cause of the  problem – massive multiplications. In this paper, we aim to  explore an efficient architecture that uses less computational  complexity and lower energy consumption.  SNN is introduced to mimic the human brain by incorporating  spikes into neural models [5]. Recent studies demonstrate that  SNN is able to achieve promising performances and significantly  reduce the energy consumption on object recognition, detection,  and tracking tasks [12], [16]. For example, Kim et al. [12]  propose SPIKING-YOLO which applies the deep SNN to the  object detection task. To expand the applicability of SNN, Yang  et al. [16] introduce a hybrid paradigm – DASHNET, to combine  the advantages of vanilla neural network and SNN in a single  model. As far as we know, those studies are mainly examined  on computer vision and speech recognition tasks. Little work is  arranged to explore the feasibility of SNN application on NLP  tasks. |
| --- | --- | --- |

[1] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.

[5] A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, and A. Maida, “Deep learning in spiking neural networks,” Neural Netw., 2019,vol. 111, pp. 47–63.

[6] X. Q. Jiao et al., “TinyBERT: Distilling BERT for natural language understanding,” in Proc. Findings, Empirical Methods Natural Lang. Process., 2020, pp. 4163–4174.

[8] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” 2015, arXiv:1503.02531.

[9] Y. He, X. Zhang, and J. Sun, “Channel pruning for accelerating very deep neural networks,” in Proc. Int. Conf. Comput. Vis., 2017, pp. 1389–1397.

[12] S. Kim, S. Park, B. Na, and S. Yoon, “Spiking-YOLO: Spiking neural network for real-time object detection,” Proc. AAAI Conf. Artif. Intell., vol. 34, no. 07, pp. 11270–11277, 2020.

[13] Z. Li et al., “Towards binary-valued gates for robust LSTM training,” in Proc. Int. Conf. Mach. Learn., 2018, vol. 80, pp. 3001–3010.

[15] Y. Kim, “Convolutional neural networks for sentence classification,” in Proc. Empirical Methods Natural Lang. Process., 2014, pp. 1746–1751.

[16] Z. Yang et al., “DashNet: A hybrid artificial and spiking neural network for high-speed object tracking,” 2019, arXiv:1909.12942.