Modified Attention with Non-Linear Kernels and its Impact on Few-Shot Learning

THE UNIVERSITY OF CHICAGO
DEPARTMENT OF COMPUTER SCIENCE

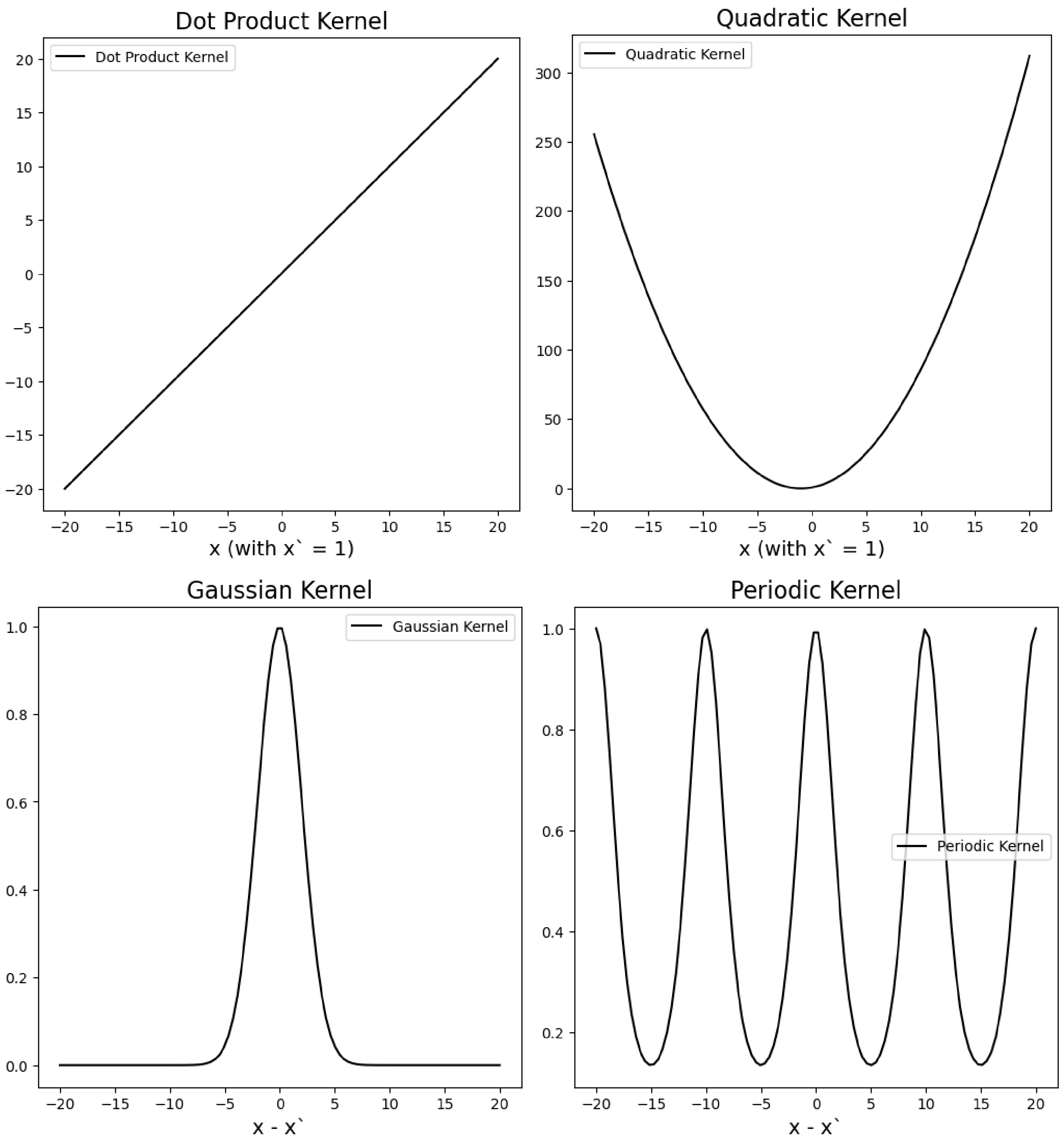
Jake Williams Bayard Walsh Chenfeng Li (williamsjl, bkwalsh, cfli@uchicago.edu)

The attention mechanism in transformers is centered around the interaction between three learned components. We replace the traditional dot product interaction with functions developed for kernel methods and fine tune models designed for natural language processing. We study these new transformers on different few-shot learning tasks designed to test different abilities of the models.

Kernel Methods

Kernel functions are a special class of function that can be written as the dot product between two mappings of its inputs, and so are often used as a measure of similarity between the inputs [1]. They are often used in classification and regression tasks, especially in support vector machines. A wide variety of kernel functions exist which capture different dependencies between the inputs, and selecting the right kernel is a meaningful and important part of using kernels in machine learning [2].

Common Kernels



Attention and Transformers

The attention mechanism is an algorithm intended for sequence based deep learning. For a set of N tokens, each with a corresponding vectors called the key (k_i), query (q_i) and value (v_i), the tokens each calculate their output as follows:

1. Construct a vector s, with $k_j = q_i*k_j$.
2. Set s equal to its own softmax to normalize it.
3. Set the output equal to s*V, where V is the collection of all value vectors into a matrix.

Transformers - the current state of the art in natural language processing and other sequence based deep learning - are a deep learning architecture that rely solely on attention and standard neural network practices [3].

Few Shot Learning

The testing methodology in this study utilizes Few-Shot learning to assess the performance of a modified kernel compared to the standard dot product. Few-shot learning involves testing a pre-trained model on a specific task after limited fine-tuning, following the approach of GPT-4 for benchmark evaluation [4,8]. Three key benchmarks—Massive Multitask Language Understanding (M-MLU)[5], AI2 Reasoning Challenge (ARC)[6], and Language Translation (English to French)[7] — were selected to evaluate the model's performance. MMLU focuses on various academic subjects, anticipating that a narrow Gaussian kernel may improve performance on subjects requiring memorization, ARC emphasizes logic and reasoning and may benefit from a periodic kernel, and translation evaluation from English to French, measured using BLEU, could benefit from a periodic kernel emphasizing connections between translation in the embedding space.

Modifying Attention with Kernels

In our work, we modify the standard attention mechanism by replacing the dot product in the firsts step of the attention algorithm with new kernel functions. We have selected three kernels to test:

1. Quadratic Kernel
2. Gaussian Kernel
3. Periodic Kernel

We expect each to have a different impact on the few shot learning, including performing differently on each task.

Results and Discussion

Due to time constraints, the final experiments are still ongoing, however the initial results indicate that the replaced kernels are performing well in fine-tuning. With comparable loss during training, these models should be capable of performing the few-shot learning tasks similarly to the original GPT-2. Final results for few-shot learning will be presented in a paper to come.

For future experiments, we would have liked to make the model much larger. Recent results indicate that model size is a large indicator for few-shot performance [8], so using GPT-3 as the base of our model would provide better opportunities for success. Similarly, we would be interested in the effects of training the model from scratch with each new kernel, rather than fine-tuning from the same starting point. This would give more flexibility in the learned model, although it would take much more time to train.

References

[1] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. 2008. Kernel methods in machine learning. The Annals of Statistics 36, 3 (2008), 1171 – 1220. https://doi.org/10.1214/009053607000000677

[2] David Kristjanson Duvenaud. 2014. Automatic Model Construction with Gaussian Processes. PhD thesis. Pembroke College.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL]

[4] Yaqing Wang and Quanming Yao. 2019. Few-shot Learning: A Survey. CoRR abs/1904.05046 (2019). arXiv:1904.05046.

[5] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring Massive Multitask Language Understanding. CoRR abs/2009.03300 (2020). arXiv:2009.03300 https://arxiv.org/abs/2009.03300

[6] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. CoRR abs/1803.05457 (2018). arXiv:1803.05457 http://arxiv.org/abs/1803.05457

[7] Dhruvil Dave. 2021. English-French Translation Dataset. https://doi.org/10.34740/KAGGLE/DSV/1926230

[8] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]