Contribution ID: **4** Type: **Poster**

DCT-Former: Efficient Self-Attention with Discrete Cosine Transform

The **Trasformer** family of Deep-Learning models is emerging as the dominating paradigm for both natural language processing and, more recently, computer vision applications.
An intrinsic limitation of this family of "fully-attentive" architectures arises from the computation of the dot-product attention, which grows both in memory consumption and number of operations as $O(n^2)$ where n stands for the input sequence length, thus limiting the applications that require modeling very long sequences. Several approaches have been proposed so far in the literature to mitigate this issue, with varying degrees of success. Our idea takes inspiration from the world of *lossy* data compression to derive an approximation of the attention module by leveraging the properties of the **Discrete Cosine Transform**. An extensive experimental analysis shows that our method takes up less memory and computation for similar performance, drastically reducing inference times.
We aim that the results of our research might serve as a starting point for a broader class of deep neural models with reduced memory footprint.
The implementation is publicly available at https://github.com/cscribano/DCT-Former-Public.

Primary authors: SCRIBANO, Carmelo (University of Modena and Reggio Emilia); Dr FRANCHINI, Giorgia (University of Modena and Reggio Emilia)

Co-authors: Prof. PRATO, Marco (University of Modena and Reggio Emilia); Prof. BERTOGNA, Marko (University of Modena and Reggio Emilia)

Presenter: SCRIBANO, Carmelo (University of Modena and Reggio Emilia)

Session Classification: Poster