# Is Attention Better Than Matrix Decomposition?

Liam McDevitt

lm15ue@brocku.ca

Department of Computer Science
Brock University
COSC 5P77 Presentation

July 7, 2022

# Overview

Introduction

Matrix Decomposition

Hamburger

An Intuitive Illustration

Hams

One-step Gradient

Experiments & Results

Conclusions & Future Work

# Introduction [1]

- In the last few years self-attention has made quite a boom providing significant advantages at capturing long-range dependencies in convolution neural networks (CNN) and recurrent neural networks (RNN).
- An attention transformer can solely out perform RNNs and CNNs for these dependencies as well.
- The ability to capture these long-range dependencies has the utter most importance in a few key problems:
  - natural language processing and
  - computer vision.
- Self-attention along with many variants provided significant scientific advancements influencing more and more scientist to join in on the progress.

# Introduction [1] (cont.)

- However, it begs the question, when modeling the global context, is hand-crafted attention irreplaceable?
- The core focus of this paper is to show that matrix decomposition (MD), which was introduced over 20 years ago, does not only provide significant performance upgrades in comparison to self-attention but it also lowers the computation cost for encoding long-range dependencies when modeling the global context.
- They develop a counterpart by formulating the extraction of global information in the networks as acquiring a dictionary with corresponding codes capturing the inherent correlation.
- This context discovery is modeled from the input tensor as a low-rank completion problem and solved using matrix decomposition.

# Matrix Decomposition [1]

- Breaks a matrix down into sub-matrices through factorization.
- Decomposing a matrix and generating a matrix are inverses.

$$\underset{\xrightarrow[\textit{decomposition}]{}}{\overset{\xleftarrow{\textit{generation}}}{\boldsymbol{X} = \bar{\boldsymbol{X}} + \boldsymbol{E} = \boldsymbol{D}\boldsymbol{C} + \boldsymbol{E}}}$$

- Data: $X = [x_1, ..., x_n] \in \mathbb{R}^{d \times n}$
- Dict.: $D = [d_1, ..., d_r] \in \mathbb{R}^{d \times r}$
- Codes: $C = [c_1, ..., c_n] \in \mathbb{R}^{r \times n}$
- Low-rank: $\bar{X} \in \mathbb{R}^{d \times n}$
- Noise: $E \in \mathbb{R}^{d \times n}$

$$\operatorname{rank}(\bar{\boldsymbol{X}}) \leq \min(\operatorname{rank}(\boldsymbol{D}), \operatorname{rank}(\boldsymbol{C})) \leq r \ll \min(d, n)$$

# Hamburger [1]

- Hamburger is a global correlation block where the global context is modeled by optimizing a low-rank completion problem.
- Matrix decomposition factorizes the learned representation into sub-matrices to further restoring a clean low-rank embedding.

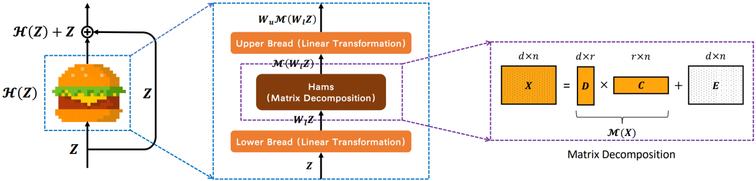


Figure 1: Hamburger's Architecture

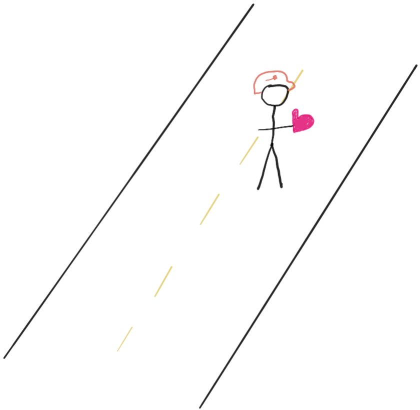
$$\min_{D,C} \mathcal{L}(X, DC) + \mathcal{R}_1(D) + \mathcal{R}_2(C)$$

$$\mathcal{H}(Z) = W_u\mathcal{M}(W_lZ)$$

$$Y = Z + \mathrm{BN}(\mathcal{H}(Z))$$

# An Intuitive Illustration [1]

- Low-rank assumption → inductive bias: *the low-level representations contain limited and much less high-level concepts than the scale of the representations themselves.*



# Hams [1]

- Hams is the process of discovering global information though an optimization problem of matrix decomposition (MD).

$$\mathcal{M}(\boldsymbol{X}) = \bar{\boldsymbol{X}} = \boldsymbol{DC}$$

- Throughout the paper and appendix three MD models are investigated:
  - Vector Quantization (VQ),
  - Non-negative Matrix Factorization (NMF), and
  - Concept Decomposition (CD).
- These three global context models are light and greatly efficient since they're mainly matrix multiplications.

# One-step Gradient [1]

- Integrating this into the network is difficult since we need to back-propagate the gradient through the iterative algorithm.
- This optimization is similar to RNNs, where normally Back-Propagation Through Time (BPTT) is used to differentiate the iterative process.
- BPTT was reviewed but ultimately it hindered Hamburger's performance.
- They attempt to remove some terms from the gradient while still keeping more dominate ones to help ensure its direction is proper for approximation.
- The one-step gradient gets its gradient by a linear approximation of the BPTT algorithm where the first term of the gradient is from the last step of the optimization.

# Experiments & Results [1]

- Ablation experiments for Hamburger were performed on the PASCAL VOC dataset for semantic segmentation.
- Important findings:
  - Hams is essential for Hamburger's performance.
  - The upper and lower bread contributes to significant performance improvements.
  - NMF outperforms VQ and CD.
  - $3 \backsim 6$ iterations $K$ for training and testing.

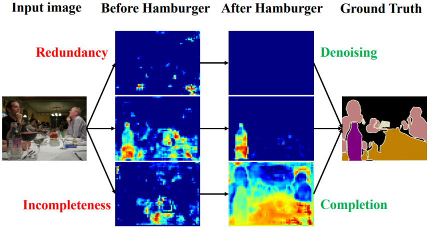


Figure 2: Visualization of feature maps

# Experiments & Results [1] (cont.)

- Hamburger was compared with Attention for both semantic segmentation and image generation.
- Results from semantic segmentation:
  - Hamburger with ResNet-101 outperformed numerous other state-of-the-art attention modules on the PASCAL VOC test set and the PASCAL Context validation set for mean intersection over union (mIoU) percentage.
- Results for image generation:
  - Hamburger versus Attention when used in GANs for image generation show Hamburger to perform 15% faster with a 5% boost in performance in Frechet Inception Distance (FID).

# Conclusions & Future Work [1]

- Conclusions:
  - The main focus of this paper was on modeling long-range dependencies in networks.
  - Self-attention is not irreplaceable as shown in this paper.
  - 20 year old matrix decomposition can compete with self-attention in difficult vision tasks proving to be fast, light, and memory efficient.
  - Introduced was a method, Hamburger, for learning the global context by formulating it as a low-rank completion problem.
- Future work:
  - Building upon the one-step gradient trick for differentiating matrix decompositions theoretically,
  - making new advanced matrix decomposition strategies,
  - and designing their very own decoder like Transformer for natural language processing.

# Bibliography

[1] Z. Geng, M.-H. Guo, H. Chen, X. Li, K. Wei, and Z. Lin, “Is attention better than matrix decomposition?” In *International Conference on Learning Representations*, 2021.