KALEIDOSCOPE: AN EFFICIENT, LEARNABLE REPRESENTATION FOR ALL STRUCTURED LINEAR MAPS

Постановка проблемы

- Для ускорения обучения в современном машинном обучении используются structured linear mapping.

- Применяется много различных классов, в каждом из которых есть компромиссы (точность, скорость, сходимость).

- Для каждой конкретной задачи исследователям приходится тратить время и силы на выбор подходящего класса.

- Хотим получить универсальное представление отображений

Критерии параметризации

1. Достаточно жестко определенное время работы
2. Алгоритм матричного умножения близок к оптимальному
3. Покрытие важных классов структуры
4. Дифференцируемая
5. Должны работать эффективные алгоритмы обучения.

K-matrices

- Представляют из себя произведение butterfly matrix
- Любое линейное преобразование (проводимое за s < n^2) может быть представлено K-matrix
- Полностью дифференцируемые, при обучении можно использовать стандартные алгоритмы оптимизации, как SGD
- Благодаря простоте и структуре, легко имплементируются и используются.

Butterfly matrix

Definition 2.1. *A **butterfly factor** of size $k \geq 2$ (denoted as $\mathbf{B}_k$) is a matrix of the form $\mathbf{B}_k = \begin{bmatrix} \mathbf{D}_1 & \mathbf{D}_2 \\ \mathbf{D}_3 & \mathbf{D}_4 \end{bmatrix}$ where each $\mathbf{D}_i$ is a $\frac{k}{2} \times \frac{k}{2}$ diagonal matrix. We restrict k to be a power of 2.*

Definition 2.2. *A **butterfly factor matrix** of size n with block size k (denoted as $\mathbf{B}_k^{(n)}$) is a block diagonal matrix of $\frac{n}{k}$ (possibly different) butterfly factors of size k:*

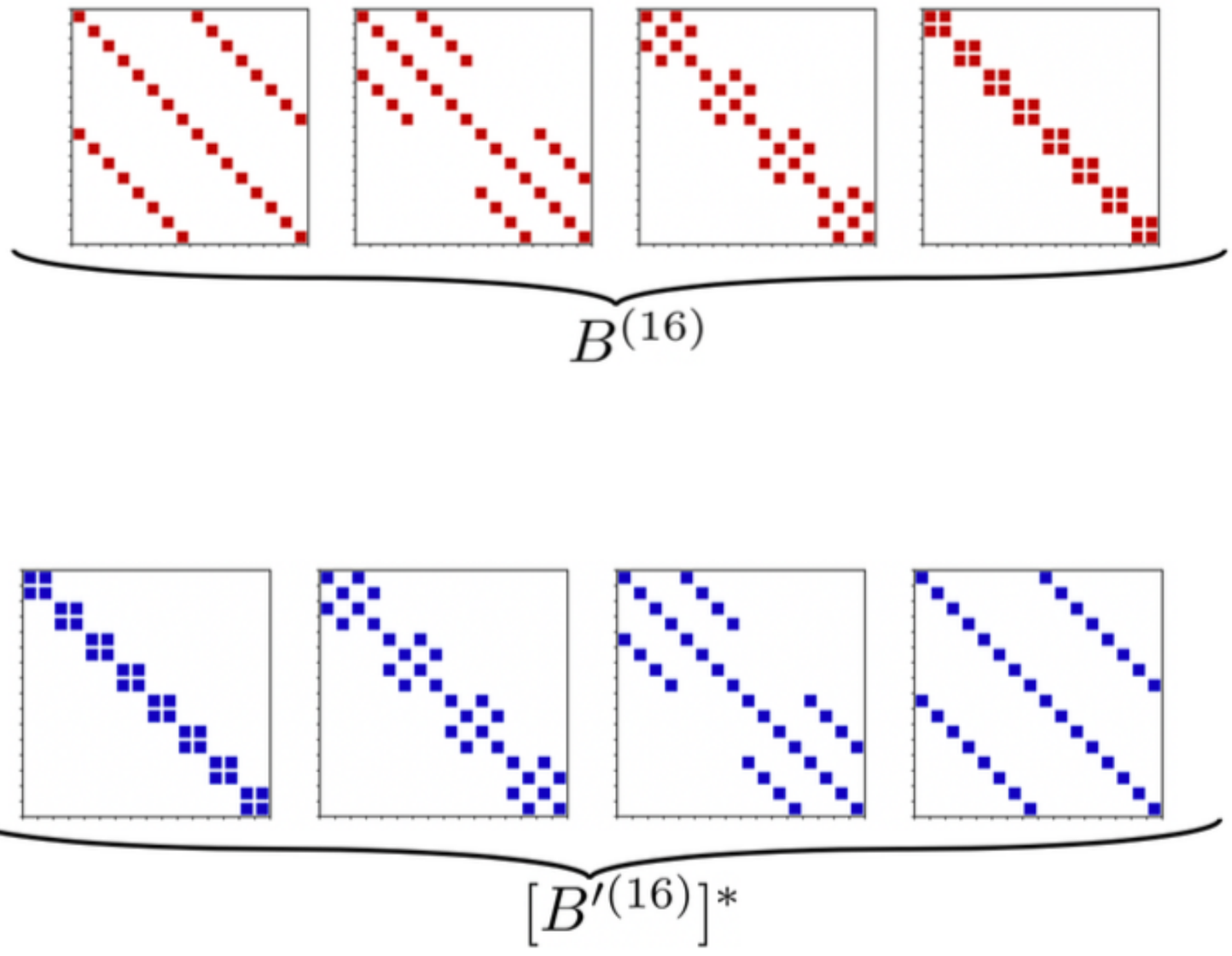
$$\mathbf{B}_k^{(n)} = \operatorname{diag}\left([\mathbf{B}_k]_1, [\mathbf{B}_k]_2, \ldots, [\mathbf{B}_k]_{\frac{n}{k}} \right)$$

Definition 2.3. *A **butterfly matrix** of size n (denoted as $\mathbf{B}^{(n)}$) is a matrix that can be expressed as a product of butterfly factor matrices: $\mathbf{B}^{(n)} = \mathbf{B}_n^{(n)} \mathbf{B}_{\frac{n}{2}}^{(n)} \ldots \mathbf{B}_2^{(n)}$. Equivalently, we may define $\mathbf{B}^{(n)}$ recursively as a matrix that can be expressed in the following form:*

$$\mathbf{B}^{(n)} = \mathbf{B}_n^{(n)} \begin{bmatrix} [\mathbf{B}^{(\frac{n}{2})}]_1 & 0 \\ 0 & [\mathbf{B}^{(\frac{n}{2})}]_2 \end{bmatrix}$$

(Note that $[\mathbf{B}^{(\frac{n}{2})}]_1$ and $[\mathbf{B}^{(\frac{n}{2})}]_2$ may be different.)

Пример



Точки показывают возможное расположение ненулевых элементов для n = 16

THE KALEIDOSCOPE HIERARCHY

- *Define $\mathcal{B}$ as the set of all matrices that can be expressed as in the form $\mathbf{B}^{(n)}$ (for some n).*
- *Define $\mathcal{B}\mathcal{B}^*$ as the set of matrices $\mathbf{M}$ of the form $\mathbf{M} = \mathbf{M}_1\mathbf{M}_2^*$ for some $\mathbf{M}_1, \mathbf{M}_2 \in \mathcal{B}$.*
- *Define $(\mathcal{B}\mathcal{B}^*)^w$ as the set of matrices $\mathbf{M}$ that can be expressed as $\mathbf{M} = \mathbf{M}_w \dots \mathbf{M}_2\mathbf{M}_1$, with each $\mathbf{M}_i \in \mathcal{B}\mathcal{B}^*$ $(1 \leq i \leq w)$. (The notation w represents **width**.)*
- *Define $(\mathcal{B}\mathcal{B}^*)_e^w$ as the set of $n \times n$ matrices $\mathbf{M}$ that can be expressed as $\mathbf{M} = \mathbf{S}\mathbf{E}\mathbf{S}^T$ for some $en \times en$ matrix $\mathbf{E} \in (\mathcal{B}\mathcal{B}^*)^w$, where $\mathbf{S} \in \mathbb{F}^{n \times en} = [\mathbf{I}_n \quad 0 \quad \dots \quad 0]$ (i.e. $\mathbf{M}$ is the upper-left corner of $\mathbf{E}$). (The notation e represents **expansion** relative to n.)*

- $\mathbf{M}$ *is a **kaleidoscope matrix**, abbreviated as **K-matrix**, if $M \in (\mathcal{B}\mathcal{B}^*)_e^w$ for some w and e.*

Основное свойство K-matrix

- Все общие линейные преобразования содержатся в BB* иерархии
- Более формально: Пусть M - матрица n*n, такая что ее умножение с вектором v может быть представлено линейными преобразованиями глубиной d и числом гейтов s. Тогда $M \in (BB^*)^{O_d}_{O_{\frac{s}{n}}}$

Применение в распознавании речи

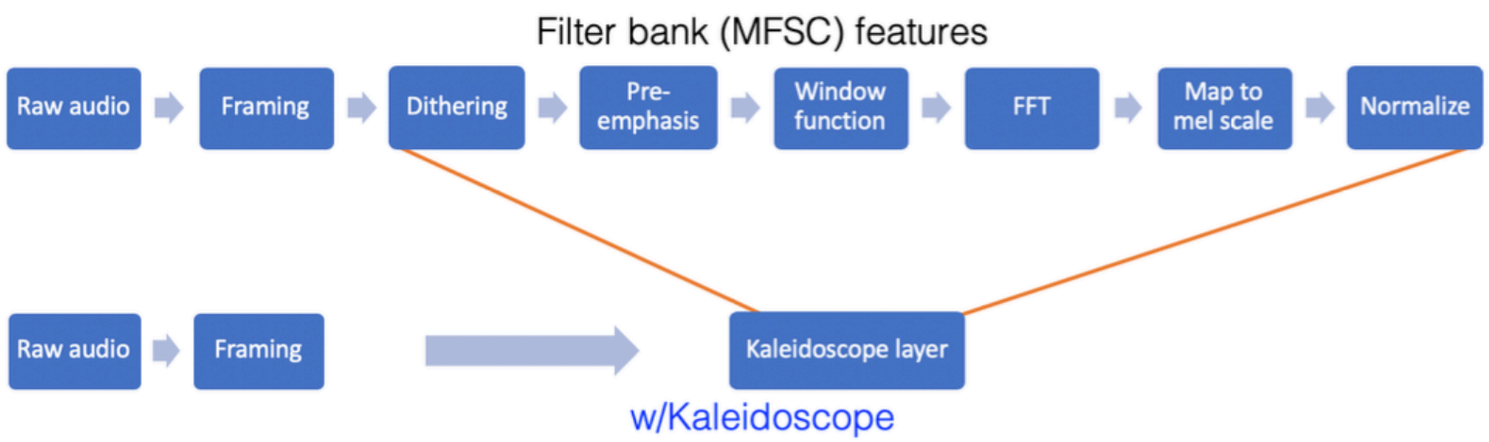


Figure 2: Comparison of the standard MFSC featurization pipeline with our "kaleidoscope" pipeline.

Method	Test set PER%	Raw audio input
MFSC features + LSTM	14.2	✗
SincNet (Ravanelli et al., 2019)	17.2	✓
Kaleidoscope + LSTM	14.6	✓

Применение в распознавание речи: сравнение числа параметров

Table 5: TIMIT phoneme error rate (PER%, ± standard deviation across random seeds).

Model	Test set PER%	# Parameters
Low rank + LSTM	23.6 ± 0.9	15.5M
Sparse + LSTM	21.8 ± 1.0	15.5M
Circulant + LSTM	23.6 ± 0.6	15.4M
Dense + LSTM	15.4 ± 0.6	15.9M
FFT + LSTM	15.7 ± 0.1	15.4M
Identity + LSTM	20.7 ± 0.3	15.4M
Kaleidoscope + LSTM	14.6 ± 0.3	15.4M
MFSC features + LSTM	14.2 ± 0.2	14.3M
SincNet (Ravanelli et al., 2019)	17.2	10.0M
LiGRU (Ravanelli et al., 2018)	13.8	12.3M

Применение в сверхточных сетях: Тестирование на ImageNet

	Shuffle	Hadamard	Kaleidoscope (K.)	K. vs. Shuffle
0.25 ShuffleNet g8	44.1% (0.46M)	43.9% (0.46M)	**49.2**% (0.51M)	+5.0% (+0.05M)
0.5 ShuffleNet g8	57.1% (1.0M)	56.2% (1.0M)	**59.5**% (1.1M)	+2.4% (+0.1M)
1.0 ShuffleNet g8	65.3% (2.5M)	65.0% (2.5M)	**66.5**% (2.8M)	+1.2% (+0.2M)

Shuffle architecture: 1x1 group conv → Batch norm, ReLU → Permutation → 3x3 depthwise conv → Batch norm → 1x1 group conv

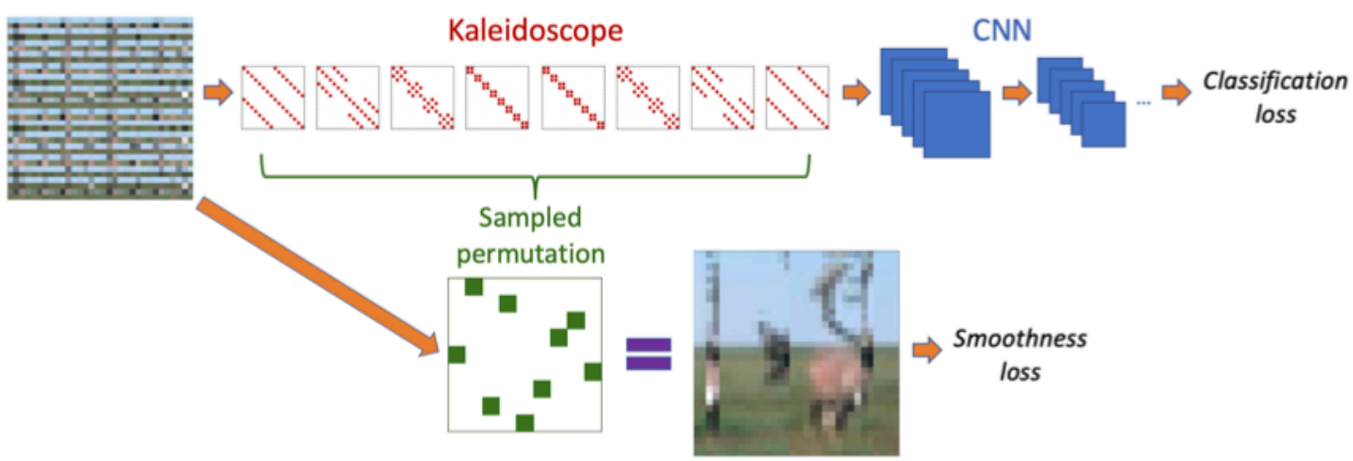
Hadamad: Hadamard → 1x1 group conv → Hadamard → Batch norm, ReLU → 3x3 depthwise conv → Batch norm → 1x1 group conv.

Kaleidoscope: K-matrix → 1x1 group conv → Batch norm, ReLU → K-matrix → 3x3 depthwise conv → Batch norm → 1x1 group conv.

Тестирование на CIFAR

Model	FC	RNN	CNN	Dense + CNN	K + CNN	Baseline CNN (unpermuted)
Accuracy	61.2	57.8	73.7	84.4	**92.5**	94.9

- FC: 3-layer MLP, with hidden size 1024 and ReLU nonlinearity in-between the fully connect layers
- Recurrent neural network (RNN): Gated recurrent unit (GRU) model (Cho et al., 2014), with hidden size 1024.
- CNN: The standard ResNet18 architecture, adapted to smaller image size of the CIFAR-10 dataset
- Dense + CNN: Additional linear layer (i.e. a dense matrix) 1024 × 1024 before the ResNet18 architecture.
- Baseline CNN: Standard ResNet18
- **K + CNN**



Применение в задаче перевода (Немецкий - Английский)

Тестирование скорости работы: как было показано в теоретической части умножение K-матрицы на вектор работает за O(nlogn)

Особенности архитектуры:
Замена dense матриц в линейных слоях декодера на K-matrix

Table 4: Inference speed on the IWSLT-14 German-English translation task (test set). Using K-matrices instead of dense matrices in the DynamicConv decoder linear layers results in 36% faster inference speed (measured on a single-threaded CPU with a batch size of 1 and beam size of 1).

Model	# params	BLEU	Sentences/sec	Tokens/sec
Transformer (Vaswani et al., 2017)	43M	34.4	3.0	66.4
DynamicConv Transformer (Wu et al., 2019)	39M	**35.2**	3.6	80.2
DynamicConv Transformer w/ K-matrices (ours)	**30M**	34.2	**4.9**	**103.4**

Выводы

- Авторы предложили решение проблемы ручного выбора типа линейного отображение введением универсального стандарта - каледоскопических матриц

- Математически доказали, что K-matrix могут представить любое структурированное линейное отображение

- В экспериментах продемонстрировали валидность такого подхода

- В будущем надеются на оптимизацию вычислений K-matrix и их широкое использование как универсального инструмента.