Università
degli Studi
di Padova

Exploring SqueezeBERT

Lorenzo Baietti	ID:2130676
Francesco Carlesso	ID:2125806
Noemi Cicala	ID:2105377
Matteo Mazzini	ID:2107797
Lisa Tassinari	ID:2121469

MSc Data Science

Natural Language Processing

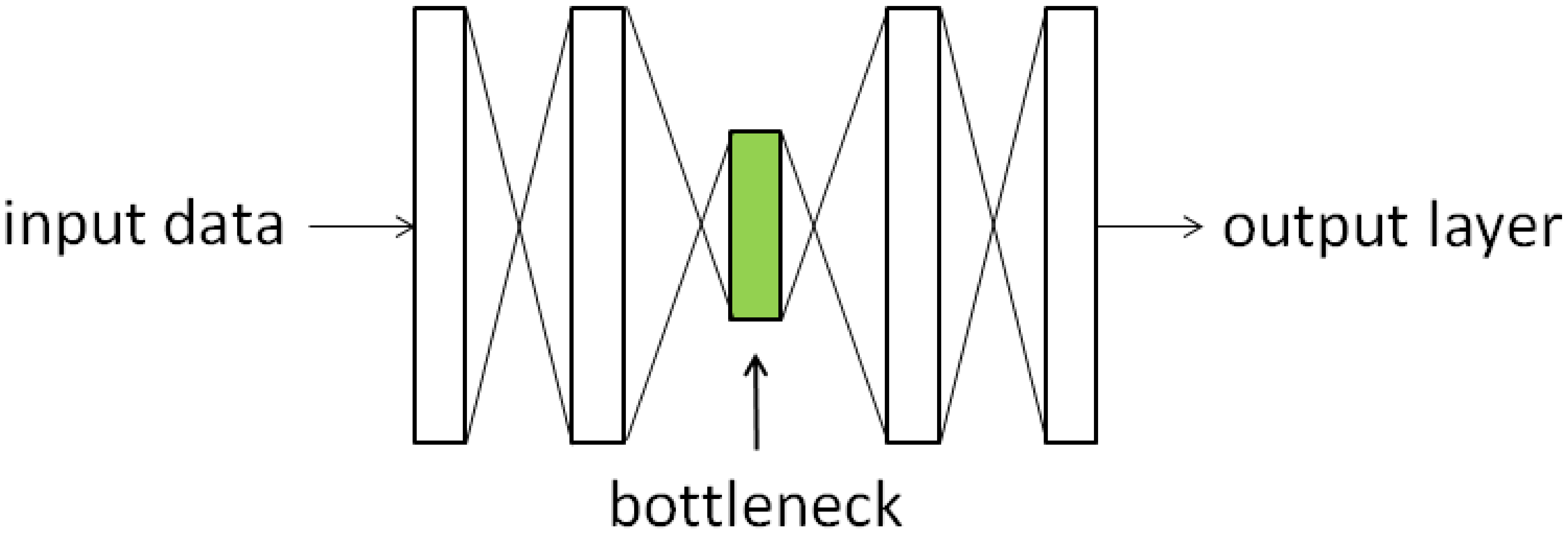
A.Y.. 2023 / 2024

Key Points

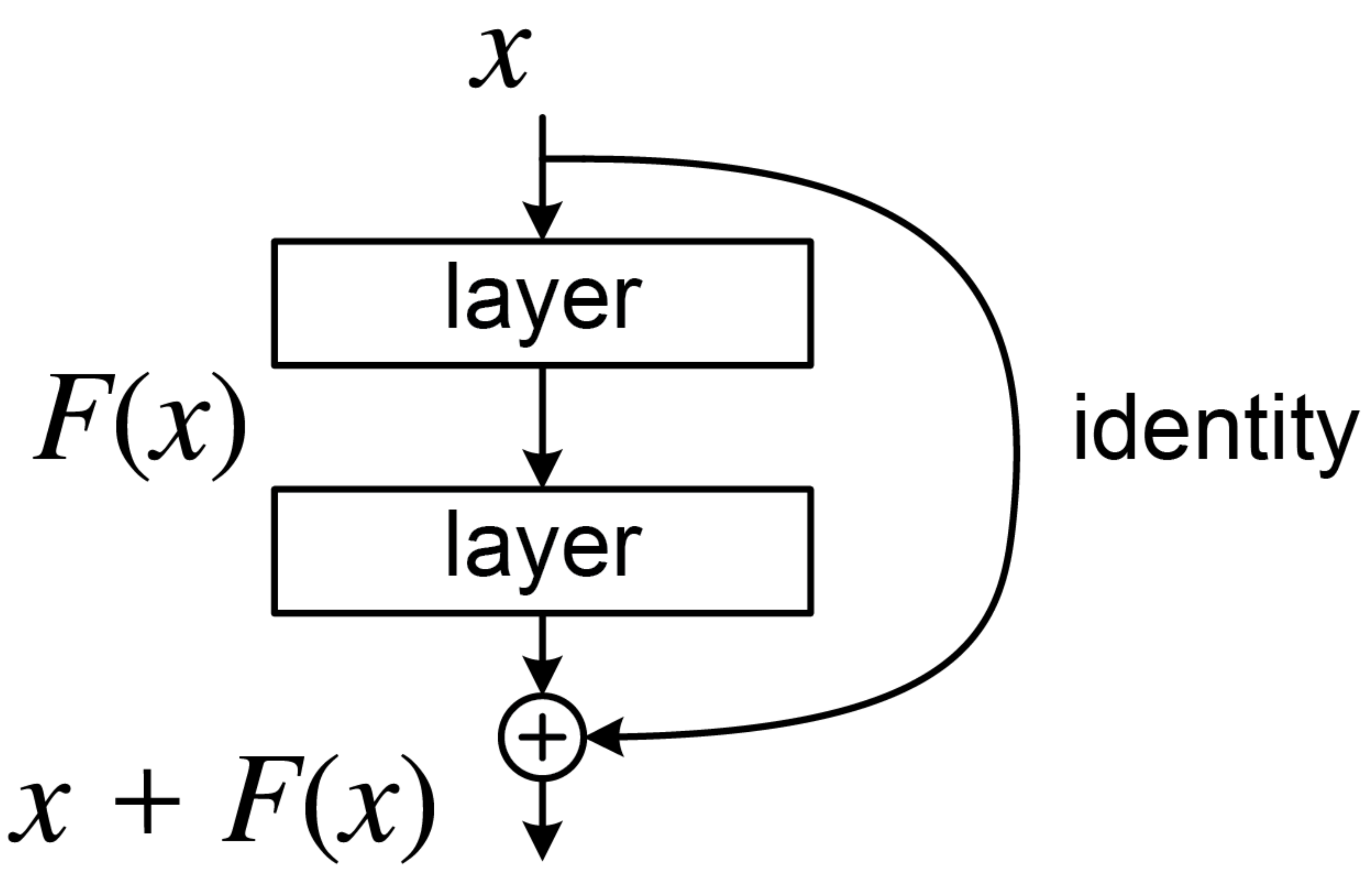
- Tailored for mobile devices
- Use of computer vision techniques
- Focus on computational efficiency (fewer parameters than other similar models)
- Accuracy vs. Inference speed trade-off

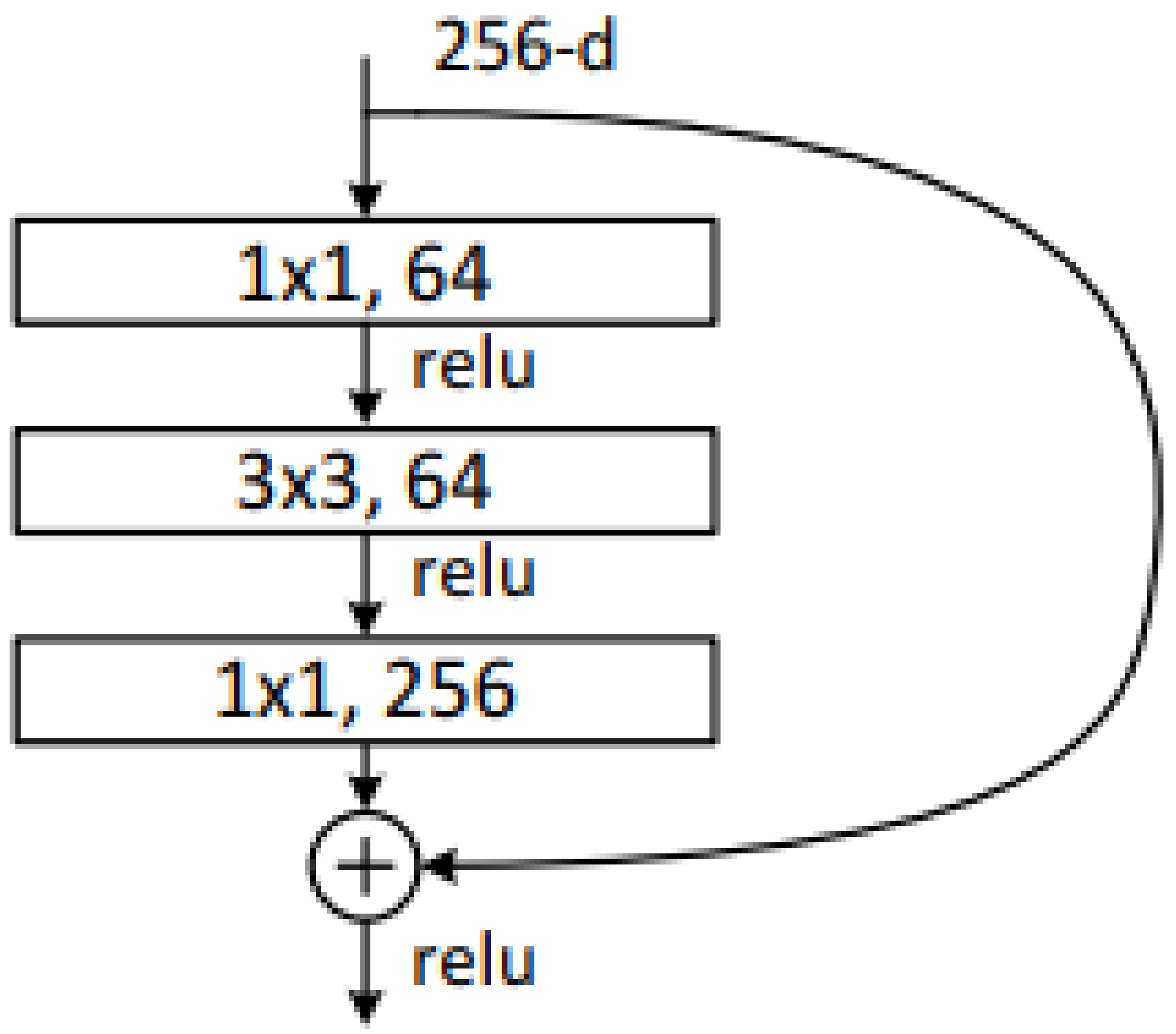
SqueezeBERT’s Architecture

Bottleneck Layers

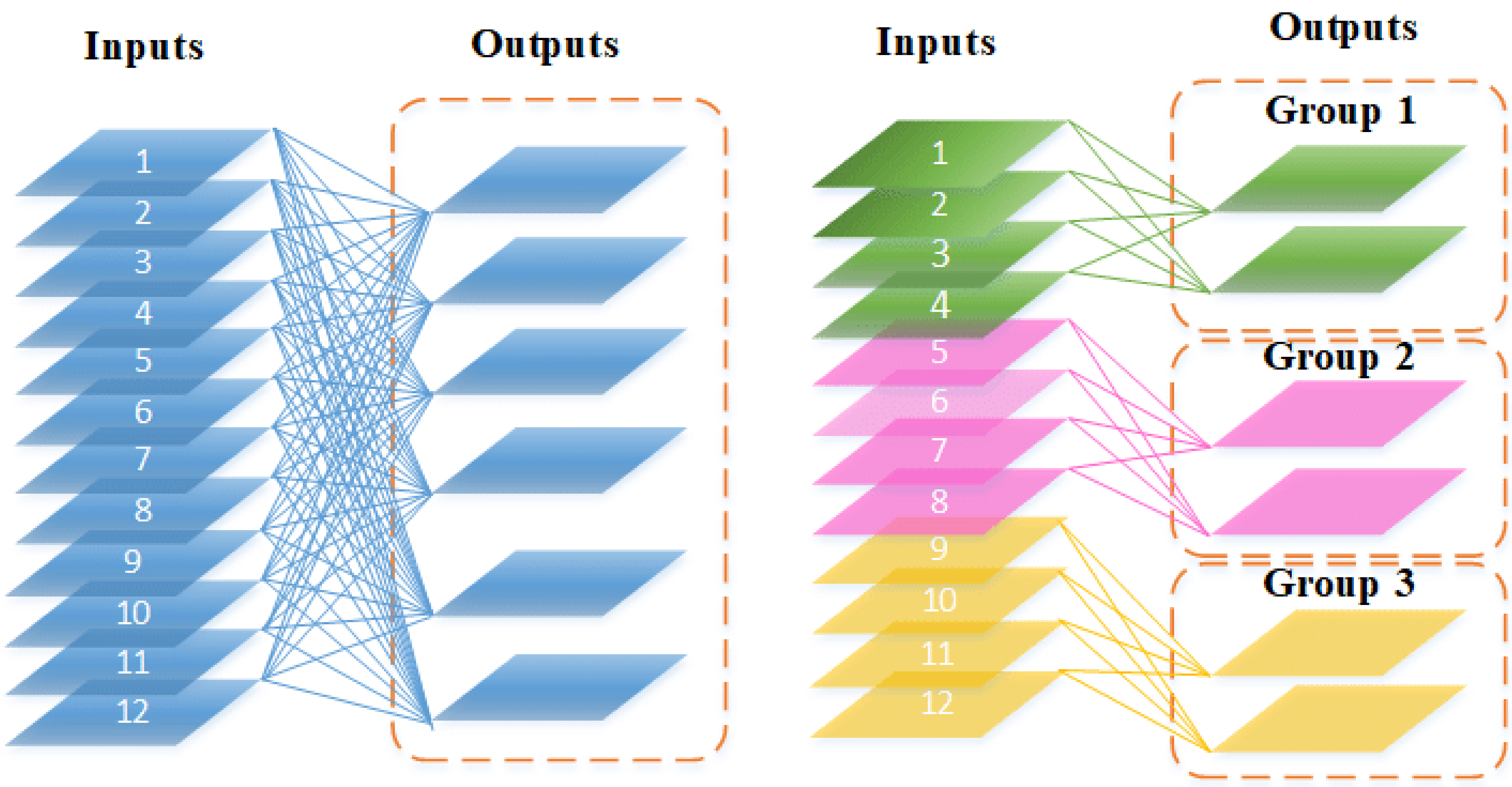


Residual Networks (ResNet)

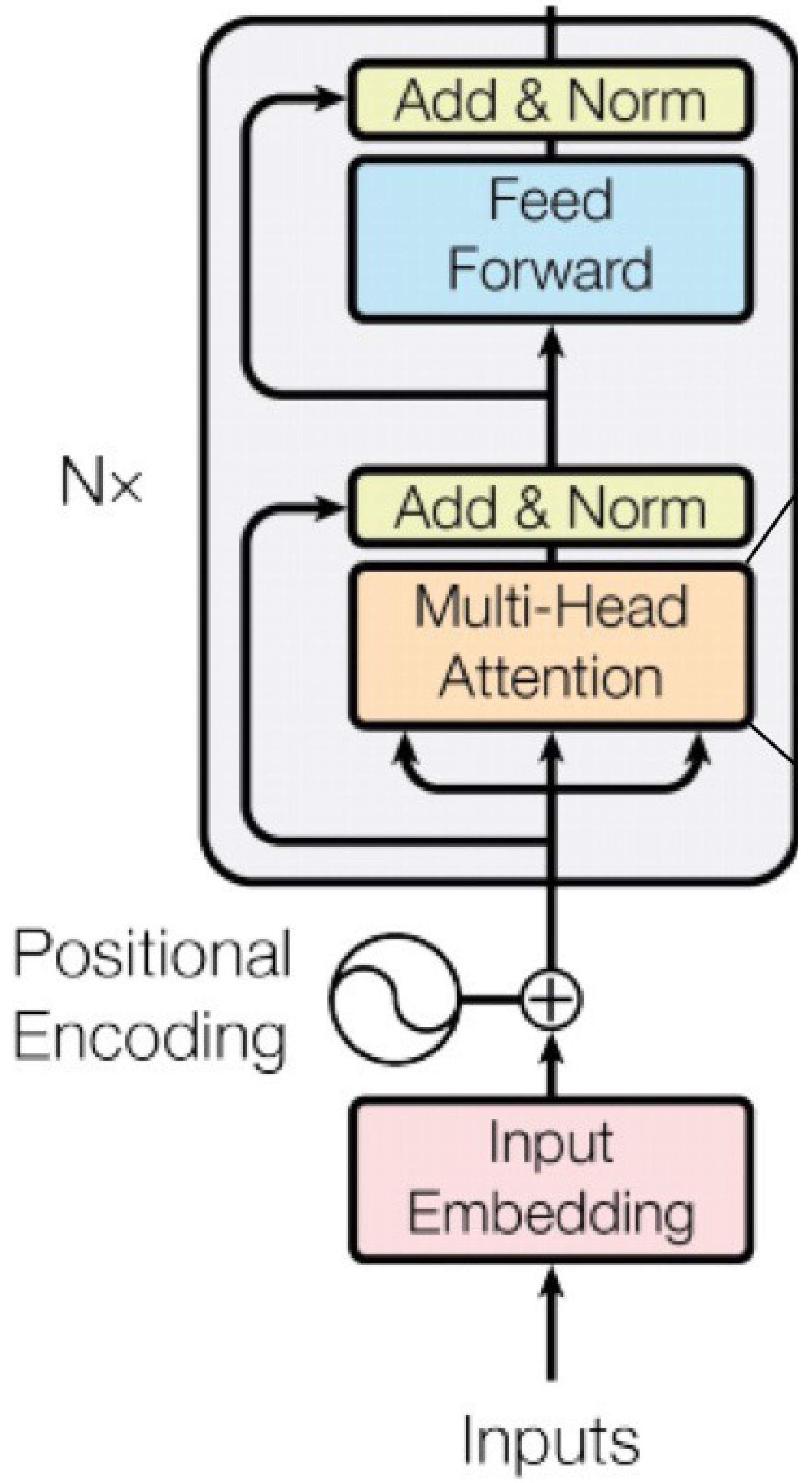




Grouped Convolutions



BERT-based Structures



- Embedding layer

  Transforms individual words into fixed-length vectors, followed by position encoding

- Encoder blocks

  Self-attention module with 3 Positionwise Fully-Connected (PFC) layers

  Three more PFCs known as Feed-Forward Network (FFN) layers

- Classifier

  Predicts the final output

SqueezeBERT Structure

- Serial connection approach with convolution before attention
- Replacing PFC layers with convolutions
- Grouped convolution to evenly distribute computational workload among FFN layers
- Similarites with BERT-base:

  768 of Embedding size; 12 Encoder blocks; 12 Heads per self-attention module; Word-piece tokenizer

Testing

Experimental Methodology

- SqueezeBERT and BERT-base comparison

- Three tasks:

  Masked Language Modeling (MLM); Text Classification; Token Classification

- Performance metrics:

  Average Cosine Similarity for MLM

  Accuracy for Text and Token Classification

Masked Language Modeling

Predicting a masked token in a sequence

Importance for Mobile Devices

- Improved understanding of context
- Multilingual applications and adaptability

Masked Language Modeling

Dataset: Improved version of the DailyDialog conversations dataset

Results

Model / Metrics	Average Cosine Similarity	CPU Time
SqueezeBERT	0.6972	115.056 sec
BERT-base	0.7820	174.756 sec

Text Classification

Assigning a sentence or document to an appropriate category

Importance for Mobile Devices

- Improved user experience
  - Spam Detection
  - Email Categorization
  - News Categorization

Text Classification

Dataset: News articles categorization

Results

Model / Metrics	Accuracy	CPU Time
SqueezeBERT	0.9463	62.666 sec
BERT-base	0.9705	99.316 sec

Token Classification

Named Entity Recognition (NER): Identifies specific entities within a text

Importance for Mobile Devices

- Contextual autocorrect and predictive text
- Accessibility features

Token Classification

Dataset: CoNLL-2003 dataset (english and german languages)

Results

Model / Metrics	Accuracy	CPU Time
SqueezeBERT	0.9674	172.922 sec
BERT-base	0.9756	300.034 sec

Conclusions

- SqueezeBERT on average 1.6 times faster than BERT-base
- Better results as tasks got easier
- Much less remarkable results than in the original paper
- Still a valid and efficient model for practical applications