

# 推理系统-模型小型化

# Transformer 小型化



# ZOMI



BUILDING A BETTER CONNECTED WORLD

Ascend

[www.hiascend.com](http://www.hiascend.com)

# Talk Overview

## 1. 推理系统介绍

- 推理系统与推理引擎区别
- 推理工作流程
- 推理系统介绍
- 推理引擎介绍

## 2. 模型小型化

- 基础参数概念
- CNN小型化结构
- Transform小型化结构

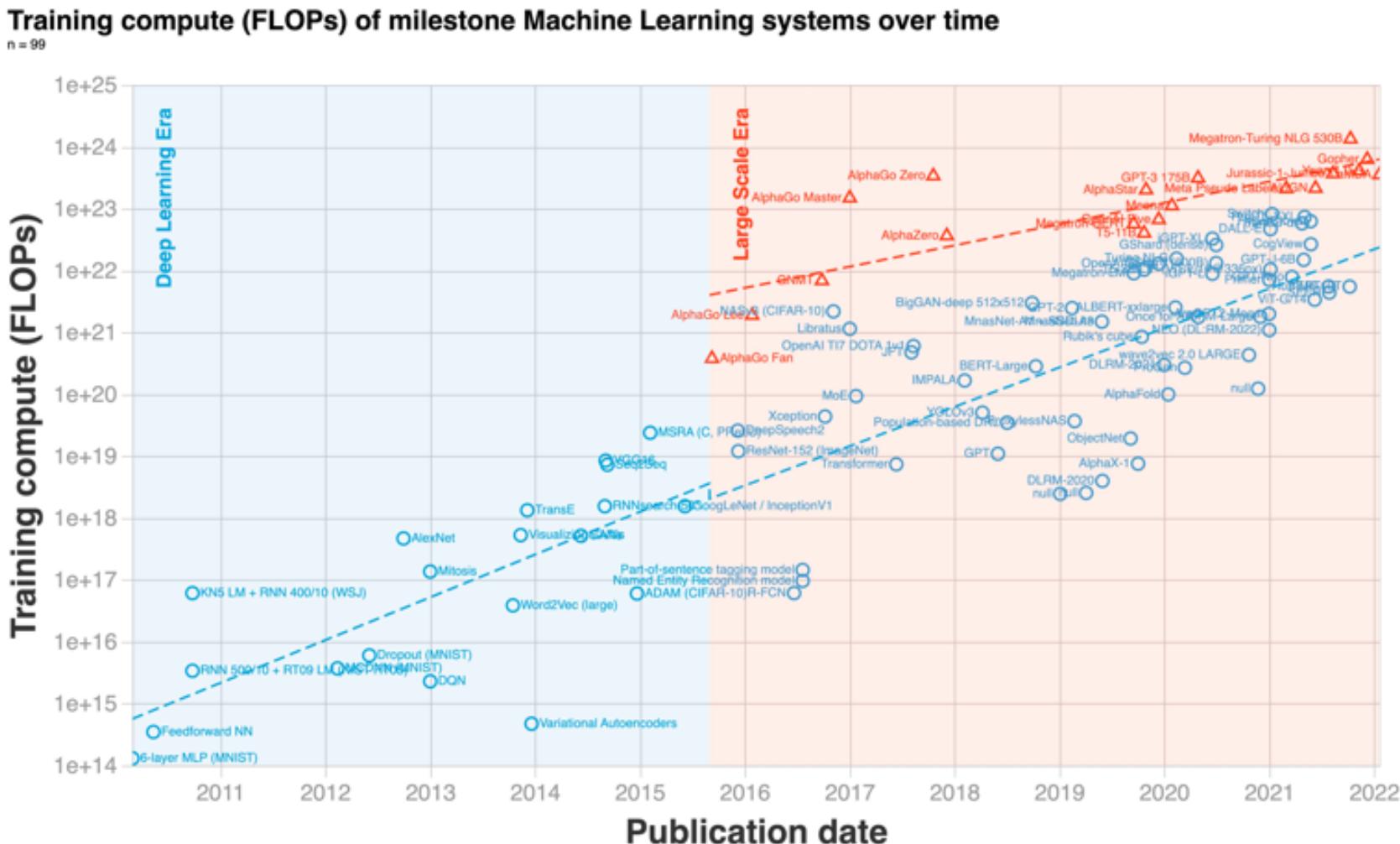
## 3. 离线优化压缩

- 低比特量化
- 二值化网络
- 模型剪枝
- 模型蒸馏

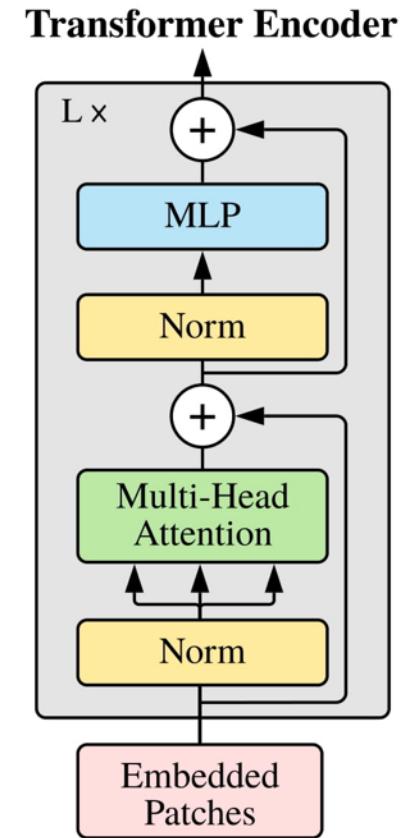
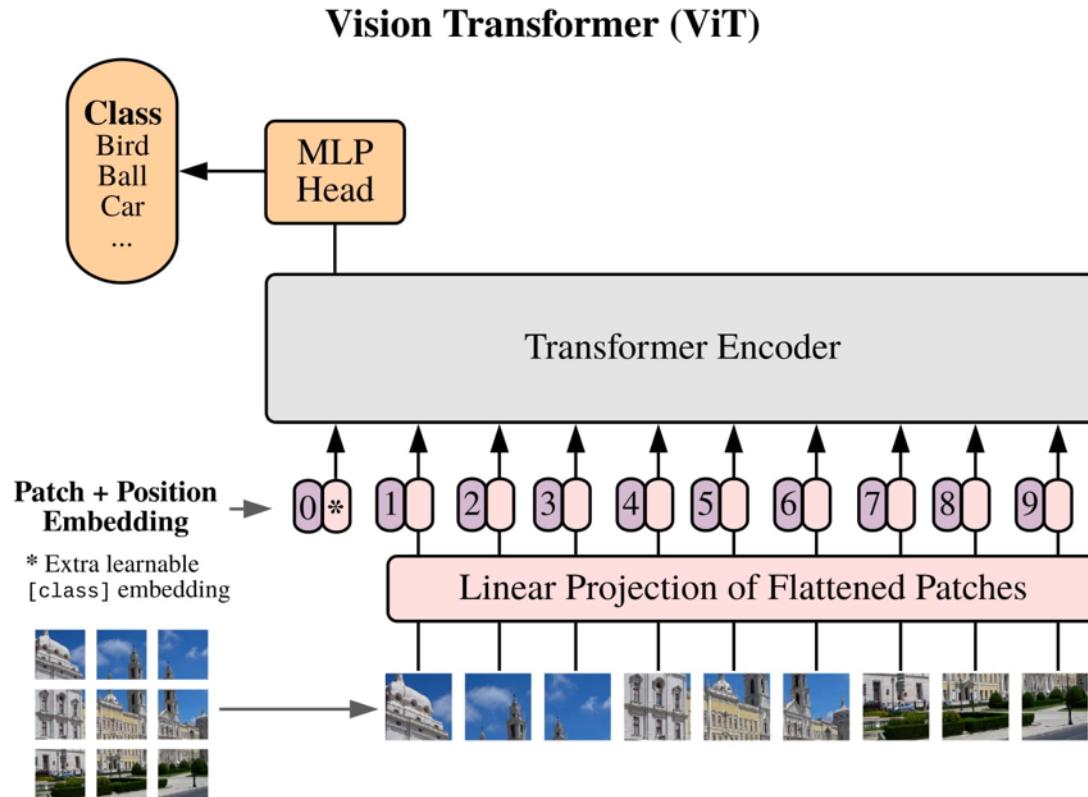
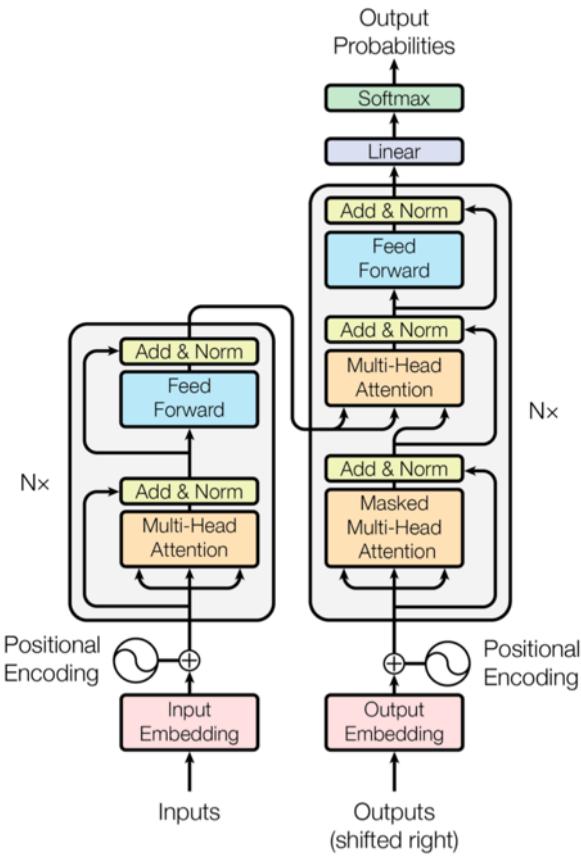
## 4. 部署和运行优化

- 图转换优化（算子融合/重排/替换）
- 并发执行与内存分配
- 动态batch与bin Packing

# 深度学习模型发展

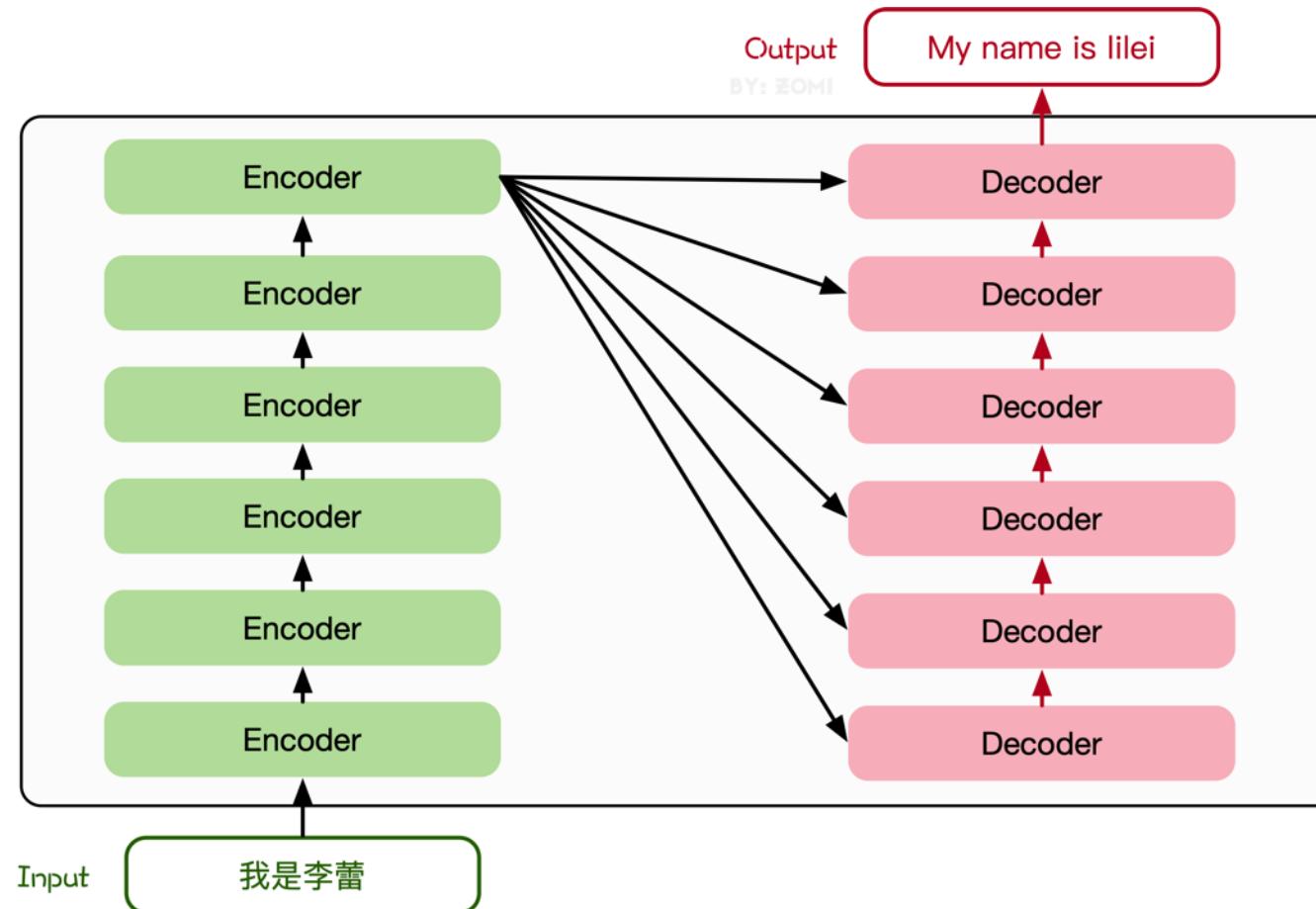


# Transformer, Attention is all you need



Vaswani, Transformer, 2017

# Transformer, Attention is all you need



# Transformer, Attention is all you need

Transformer solves Seq2Seq problem, replaces LSTM/RNN with a full attention structure:

- One step calculation to solve the long-term dependency problem;
- The computational complexity of each layer is better;
- Point multiplication results can be directly calculated;

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	<b>41.29</b>	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	<b><math>3.3 \cdot 10^{18}</math></b>	
Transformer (big)	<b>28.4</b>	<b>41.0</b>	$2.3 \cdot 10^{19}$	

Vaswani, Transformer, 2017

# Inference

1. Q8BERT: Quantized 8Bit BERT – 2019.10
2. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter – 2019.10
3. TinyBERT: Distilling BERT for Natural Language Understanding – 2019.09
4. Training data-efficient image transformers & distillation through attention – 2021.3
5. MiniViT: Compressing Vision Transformers with Weight Multiplexing – 2022.04
6. TinyViT: Fast Pretraining Distillation for Small Vision Transformers – 2022.06
7. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification - 2022.10
8. Compressing Visual-linguistic Model via Knowledge Distillation – 2021.03
9. MiniVLM: A Smaller and Faster Vision-Language Model – 2021.09

# 轻量级模型

1. MobileViT ( 2021 )
2. Mobile-Former ( 2021 )
3. EfficientFormer ( 2022 )

# 轻量化网络总结

## 如何选择轻量化网络：

1. 不同网络架构，即使 FLOPs 相同，但其 MAC 也可能差异巨大
2. FLOPs 低不等于 latency 低，结合具硬件架构具体分析
3. 多数时候加速芯片算力的瓶颈在于访存带宽
4. 不同硬件平台部署轻量级模型需要根据具体业务选择对应指标

# Inference

1. Q8BERT: Quantized 8Bit BERT – 2019.10
2. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter – 2019.10
3. TinyBERT: Distilling BERT for Natural Language Understanding – 2019.09
  
4. Training data-efficient image transformers & distillation through attention – 2021.3
5. MiniViT: Compressing Vision Transformers with Weight Multiplexing – 2022.04
6. TinyViT: Fast Pretraining Distillation for Small Vision Transformers – 2022.06
7. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification - 2022.10

8. Compressing Visual-linguistic Model via Knowledge Distillation – 2021.03

9. MiniVLM: A Smaller and Faster Vision-Language Model – 2021.09



BUILDING A BETTER CONNECTED WORLD

THANK YOU

Copyright©2014 Huawei Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.