arXiv:2302.04907v1 [cs.CL] 9 Feb 2023

二值化神经机器翻译

Yichi Zhang [1] Ank ush Garg [2] Yuan Cao [2] Łukasz Lew [2] Behrooz Ghorbani [2] Zhiru Zhang [1] Orhan Firat [2]

摘要

语言模型的快速扩展推动了低位宽量化研究。在本文中，我们提出了一种用于机器翻译的全新 Transformer 二值化技术（BMT），这是首创。我们发现并解决了使用一位权重和激活值时点积方差膨胀的问题。具体而言，BMT 利用额外的层归一化和残差连接来提升二值化质量。在 WMT 数据集上进行的实验表明，仅采用一位权重的 Transformer 能够达到与浮点模型相同的质量，同时体积缩小了 16× 。一位激活值会造成不同程度的质量下降，但所提出的架构改动能够缓解这一问题。我们还使用生产规模的翻译数据集开展了缩放定律研究，结果表明，一位权重 Transformer 在域内和域外环境下都能良好扩展并实现泛化。基于 JAX/Flax 的实现将开源。

1. 引言

神经语言模型正不断扩展，近期模型（如 GPT 系列）的参数数量大约每年增长 10× （Narayanan 等人，2021年）。Kaplan 等人 (2020) 的一项缩放定律研究表明，模型参数的持续增加与性能提升密切相关。近期大型模型的成功验证了这一趋势，例如拥有 5400 亿参数的 Pathways 语言模型（ PaLM ），在语言理解与生成方面取得了突破性表现（Chowdhery 等人，2022年）。拥有 5400 亿参数的 Minerva（Lewkowycz 等人，2022年）在 2021 年的波兰全国数学考试中也超过了全国平均水平；此前，语言模型在该考试中的表现远未达到人类水平。类似地，据报道，在神经机器翻译（机器翻译）领域，缩放定律同样成立，Ghorbani 等人 (2021年) 指出，随着模型大小增加，翻译质量也会提高。

这种激进的扩展趋势给模型服务带来了前所未有的挑战。具体而言：

推理成本呈指数级增长。 语言模型的大小和计算复杂度正在快速增加，每年模型大小大约 10× 增长，而每年的运算次数大约100× 增长（Hoffmann 等人，2022年）。然而，用于运行这些模型的硬件能效并未跟上这一速度。具体而言，FP32 运算所需的能耗在过去 11 年中（2007—2018年）仅提升了 2.5× ，制程节点从 45nm 发展到 7nm。在同一时期，动态随机存取存储器访问能耗仅提升了 6.3× （Jouppi 等人，2021年）。模型大小膨胀与硬件能效低下之间不断扩大的差距，正导致推理能耗呈指数级增长，并逐渐成为在数据中心运行语言模型成本的重要组成部分。

芯片间通信的开销变得不可忽视。对于如此大规模的模型，仅靠数据并行已不再够用，因为一次矩阵乘法无法在单个加速器芯片上完成。例如，PaLM 中的每个权重张量都分布在一个 pod 中的 3072 个 TPUv4 芯片上（Chowdhery 等人，2022年）。这导致在数据中心网络中传输权重和中间激活值的开销巨大。

对延迟敏感的应用如今几乎无法从参数缓存中获益。推理期间，将模型参数从动态随机存取存储器加载到片上加速器存储器通常非常耗时。过去，参数缓存是一种有效的延迟优化方法，因为它能重复利用模型权重，避免片外存储器传输。然而，针对边缘 TPU 的评估报告指出，该方法最适用于参数少于 3000 万的模型（Se-shadri 等人，2021年）。对于更大的模型，参数缓存甚至会适得其反。编译器优化带来的收益正在递减，而服务延迟几乎与模型参数数量成正比。我们的案例中，最小的翻译模型约有 5000 万个参数。因此，要降低延迟，归根结底只能提高存储器带宽。

量化可以显著降低推理成本。二值化是一种极端情况，此时权重和

[1]康奈尔大学 [2]Google Research。通信作者：Yichi Zhang <yz2499@cornell.edu>、Ankush Garg<ankugarg@google.com>、Yuan Cao <yuancao@google.com>、Łukasz Lew <lew@google.com>、Behrooz Ghorbani <ghor- bani@google.com>、Zhiru Zhang <zhiruz@cornell.edu>、Orhan Firat <orhanf@google.com>。

矩阵乘法（matmul）中的激活值被量化为单个比特。与脑浮点格式（bfloat16）（Abadi 等人，2016年）[1]相比，二值化可将权重大小减少 16×，从而显著降低内存和通信开销。此外，二值化矩阵乘法可通过同或运算和随后的人口计数来完成，据估计，这种方法比 bfloat16 对应操作 256× 更节能（Zhang 等人，2022年）。在准确率与效率的权衡方面，二值化在 ImageNet 上取得了成功（Zhang 等人，2022年）。

已有研究表明，BERT 可以用于预训练二值化（Bai 等人，2020；Qin 等人，2022；Liu 等人，2022）。然而，需要注意的是，BERT 和机器翻译模型虽然都以 Transformer 为核心（Vaswani 等人，2017年），但二者存在很大差异。一个关键区别在于架构：机器翻译模型同时包含编码器和解码器，而 BERT 只有编码器。这一差异可能会影响编码器量化的质量，因为解码器中的每个交叉注意力层都需要编码器的输出。另一个区别是，机器翻译模型的推理会生成一段文本序列，而 BERT 执行的是单次文本分类。这一点至关重要，因为输出翻译序列中的每个词都会影响下一个词的生成。因此，一个词的采样分布非常关键，二值化后应当予以保留；而对于 BERT，只需保留逻辑值的峰值即可。由于这些差异，将 BERT 的二值化技术直接应用于机器翻译，很容易导致模型质量下降。

在这项工作中，我们研究了用于神经机器翻译的二值化 Transformer。据我们所知，这是对这一主题的首次研究。每个 Transformer 模块都包含一个注意力层和一个前馈网络（FFN）。我们分别对权重和激活值进行二值化，以研究它们各自对模型质量的影响。我们发现，二值化权重对准确率没有显著影响，但传统的激活值二值化方法会因激活值幅度爆炸而导致性能不佳。随后，我们提出了一种新的激活值二值化方法，该方法使用简单的缩放因子和额外的残差连接。

为了解该 1 位 Transformer 在实践中的扩展特性，我们进一步在内部生产规模翻译数据集上对其进行了评估。该数据集包含三十亿个句对。我们首次证明，即使在域外数据上，1 位权重 Transformer 的扩展性和泛化能力也与浮点模型同样出色。我们还分析了从两个模型输出中采样的句子，发现 1 位 Transformer 生成的翻译质量与其浮点对应模型相近。因此，二值化可能成为未来机器翻译模型服务的一种潜在方案。

[1]在本文余下部分中，“float”指 bfloat16。

2. 相关工作

Transformer 的成功推动了大量将其量化为更低精度的研究。在本节中，我们回顾其中一部分启发了我们方法的工作。

Transformer 量化。先前的大部分工作都聚焦于 8 位 Transformer。Bhandare 等人(2019) 报告称，在 WMT14 英德翻译任务中使用 8 位时，BLEU 分数下降不到 0.5。Prato 等人(2019) 表明，8 位 Transformer 保持了翻译质量。对于非生成式任务，Zafrir 等人(2019) 将 BERT 量化为 8 位，质量损失微乎其微。然而，当位数降至 4 位时，Prato 等人(2019) 报告称，机器翻译的 BLEU 分数下降了 8 分；Aji 和 Heafield (2019) 则报告称，采用对数量化方案几乎不会造成 BLEU 损失。

围绕 1 位 Transformer 的探索主要集中于 BERT。通常会直接进行二值化，重点则放在改进训练方案上。Bai 等人 (2020年) 首先尝试将三值 BERT 拆分为二值 BERT，然后进行微调。在 GLUE 基准上，其平均准确率达到 41%。Qin 等人(2022年) 提议从浮点模型中蒸馏每个中间层的输出。最近，Liu 等人 (2022年) 提议对模型进行渐进式量化，例如从 32 位量化到 4 位，再到 2 位，最终量化到 1 位，并将 GLUE 准确率提升至 73.5%。

二值化视觉模型。Courbariaux 等人(2016年) 开创了对二值化深度神经网络的研究。最近，PokeBNN（Zhang 等人，2022年）在 ImageNet 识别任务上建立了帕累托 SOTA。我们沿用了 PokeBNN 的二值化函数和训练方案。

泛化能力。Hooker 等人（2019）表明，压缩模型在域外（OOD）数据上的泛化能力较差。我们尤其关注在域外设置下评估 BMT，并分析其泛化能力。

3. 算法与模型架构

本节将介绍对基于 Transformer 的机器翻译模型进行二值化的方法。我们首先定义二值化方程，然后说明将这些方程直接应用于 Transformer 会导致模型质量下降，原因是点积方差膨胀。接着，我们提出一种缩放因子来解决这一问题，并讨论使用层归一化（Ba 等人，2016年）替代固定缩放因子。最后，我们结合并介绍为提升二值化模型质量而进行的必要架构调整。

3.1. 二值化方程

我们遵循 PokeBNN（Zhang 等人，2022年）中定义的方法，其中包含一个重要的超参数“B”。将浮点数值转换为二进制

值的函数总结如下。

$$\text{clip}\left(x, x_{min}, x_{max}\right) := \min\left(x_{max}, \max\left(x_{min}, x\right)\right)$$
$$x_b := \left(\text{floor}\left(\text{clip}\left(\frac{x}{B}, -1+\epsilon, 1-\epsilon\right)\right) + 0.5\right) \times B$$

其中， x 是输入张量， ϵ 是一个较小的浮点数，用于在取向下取整时防止溢出，而 B 是二值化界限。在反向传播中，忽略向下取整函数，即 $\frac{\partial \text{floor}(x)}{\partial x} := 1$，称为直通估计器（Courbariaux 等人，2016年）。整个二值化函数的梯度为 $\frac{\partial x_b}{x} = 1_{x \in [-B,B]}$，否则为零。因此，界限 B 充当超参数，用于控制输入值的范围，只有处于该范围内的输入值才具有非零梯度。注意，B 也充当输出的缩放因子，因为二值化函数将 $x \to \{-\frac{B}{2}, +\frac{B}{2}\}$ 映射为 B 也可以推广为向量，具体取决于二值化的粒度。不过，最细的粒度是每个点积对应一个界限值，即每个缩并维度对应一个界限值，这样便可加速二值化矩阵乘法。

对于形式为 $A \cdot W$ 的 Transformer 全连接层，其中 $A^{N \times d_{\text{model}}}$ 是输入激活值， $W^{d_{\text{model}} \times d_k}$ 是模型权重，我们改为计算二值化矩阵乘法 $A_b \cdot W_b$。在所有实验中，我们分别对权重和激活值采用二值化界限 B_W 和 B_A。

$$B_W = \max(\text{abs}(W), \text{axis} = d_{\text{model}})$$
$$B_A = \max(\text{abs}(A), \text{axis} = d_{\text{model}})$$

其中 轴 是沿其取 最大值 的维度。只使用一个轴意味着界限是按通道、按样本设定的（Lew 等人，2022年）。和 $B_W^{1 \times d_k}$ 都是向量，包含沿缩并维度取得的最大绝对值。请注意，权重二值化界限 B_W 在推理时是静态的，但每次训练迭代都会更新。激活值界限 B_A 则是动态的。

3.2. 二值化中的方差膨胀

我们首先将二值化函数应用于前馈网络（前馈神经网络），其他模块则保持浮点形式。我们发现，直接对权重进行二值化能够保持模型质量，但对输入激活值进行二值化会导致机器翻译任务中的训练无法收敛。为了理解这一现象的原因，我们分析了有无二值化时点积幅值的方差。分析表明，同时对权重和激活值进行二值化会在统计上增大幅值，导致神经网络内部的信号传播异常（Brock 等人，2021年）。下面将详细介绍这一分析。

设全连接层的每个权重都经过随机初始化，并从零均值正态分布中采样， $w \sim \mathcal{N}(0, \sigma_w^2)$。假设每个输入激活值都与权重相互独立，且均服从 $a \sim \mathcal{N}(0, \sigma_a^2)$。应用二值化函数后， w_b 和 a_b 仍以零为中心，且取 $-\frac{B}{2}$ 或 $+\frac{B}{2}$ 的概率相等，即它们服从如下定义的概率质量函数：

$$\Pr\left(x_b\right) = \begin{cases} \frac{1}{2} & x_b = -\frac{B}{2} \\ \frac{1}{2} & x_b = +\frac{B}{2} \end{cases}$$

因此，二值化乘法的方差为

$$\text{Var}\left(a_b \cdot w_b\right) = \mathbb{E}\left[a_b^2\right] \cdot \mathbb{E}\left[w_b^2\right] - \mathbb{E}^2\left[a_b\right] \cdot \mathbb{E}^2\left[w_b\right]$$
$$= \sum_{a_b} a_b^2 \cdot \Pr\left(a_b\right) \cdot \sum_{w_b} w_b^2 \cdot \Pr\left(w_b\right) - 0 = \frac{B^4}{16}$$

因此，二值化点积的方差为

$$\text{Var}\left(A_b \cdot W_b\right) = \sum_{n=0}^{D-1} \text{Var}_n\left(a_b \cdot w_b\right) = \frac{B^4}{16} \cdot D$$

其中， D 是点积的维度，即前馈神经网络中的隐藏投影维度，而 n 是向量中每个元素的索引。

按照相同的分析，浮点点积的方差为

$$\text{Var}\left(A \cdot B\right) = \sigma_a^2 \cdot \sigma_w^2 \cdot D$$

请注意，常用的 Xavier 初始化器（Glorot&Bengio，2010年）会使各层激活值的方差相等。 σ_w^2 因此将初始化为 $\frac{1}{D}$，所以 $\text{Var}\left(A \cdot W\right) = \sigma_a^2$，其通常在 1 的量级。

与此同时，常用的二值化界限为 $B \in [1, 3]$（Courbariaux 等人，2016年；Zhang 等人，2022年；Bethge 等人，2021年）。我们的 Transformer 前馈神经网络在所有实验中均采用隐藏投影维度 $D = 4096$。因此，$\text{Var}\left(A_b \cdot W_b\right) \gg \text{Var}\left(A \cdot W\right)$。二值化会使点积方差至少膨胀 256×倍，这会反映在全连接层输出的幅值上。另请注意，$\text{Var}\left(A_b \cdot W_b\right) \propto D$，这表明宽度更大的 Transformer 可能更容易出现收敛问题。

3.3. 缩放因子作为解决方案

受缩放因子的启发 $\sqrt{d_k}$ 在缩放点积注意力 注意力（Q， K， V） $= \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$ 在原始 Transformer（Vaswani 等人，2017年）中，我们为每个二值化全连接层提出一个缩放因子，即：

$$\text{Dense}\left(A_b\right) = \frac{A_b \cdot W_b}{s}$$

缩放因子 s 是一个超参数，用于抑制点积方差膨胀；在注意力层中

$\sqrt{d_k}$ 可防止点积进入 softmax 函数的梯度较小区域。根据第 3.2 节的分析，其值估计为 $s \propto \sqrt{D}$，以抵消 D 对方差的乘法效应。

为验证缩放因子的幅值如何影响训练损失，我们在 s 第5节中进行了扫描。实际上，$s \geq 64$可以使训练收敛。

3.4. 用层归一化替换缩放因子

虽然缩放因子 s 能够实现前馈神经网络的二值化，但它需要进行超参数调优，这对于拥有数十亿参数的翻译模型而言可能颇具挑战。为解决这一不足，我们提出使用层归一化（LayerNorm）（Ba 等人，2016年）直接替代缩放因子，其形式如下：

$$\mathrm{LN}(x) = \frac{x - \mathbb{E}[x]}{\sqrt{\mathrm{Var}(x) + \epsilon}} \cdot \gamma + \beta$$

其中，γ 和 β 是可学习参数。除了 γ 能够纳入缩放因子 s这一点之外，层归一化还有以下优点。

缩放因子在训练过程中变得动态且自适应。二值化函数采用动态界 B，因此 Var($A_b \cdot W_b$)会发生变化。层归一化中的可学习参数γ 能够更好地捕捉点积方差的变化，从而对其进行适当归一化。

层归一化还会重新分布输入激活值。它使得可以对包含全为正值的张量进行二值化。直接二值化的前馈神经网络结构如下：

$$\mathrm{FFN}(A) = \max(0, A_b W_{1b} + b_1)_b W_{2b} + b_2$$

其中 W_1，b_1 和 W_2，b_2 分别是第一个和第二个全连接层的权重和偏置。可以注意到，激活值$\max(0, A_b W_{1b} + b_1)$ 全为正值。此时，二值化函数会将整个张量映射为一个常数 $+\frac{B}{2}$，这会妨碍模型训练。然而，借助层归一化，激活值会重新分布，其正值和负值的数量也会更加均衡。这使得第二个全连接层能够进行常规的$\{-1, +1\}$（双极）二值化。Qin 等人（2022年）；Liu 等人（2022年）在二值化 BERT 中改用 $\{0, 1\}$ 二值化，以解决正值恒定的问题。由于 $A \in \{0, 1\}^{N \times D}$ 且 $W \in \{-1, +1\}^{D \times K}$，这会产生三值矩阵乘法；若在二值硬件加速器上计算，则会带来相当大的额外开销。

完整的所提出的 1 位前馈神经网络结构如下：

$$\mathrm{FFN}(A) = \mathrm{LN}(\mathrm{LN}(\max(0, A_b W_{1b} + b_1))_b \cdot W_{2b} + b_2)$$

在进行注意力二值化时，出于相同的原因，我们在每个线性投影层的输出上添加层归一化，以验证这一点。我们在第5 节中证实，层归一化中的动态自适应缩放因子确实优于固定缩放因子。

3.5. 注意力层中的残差连接

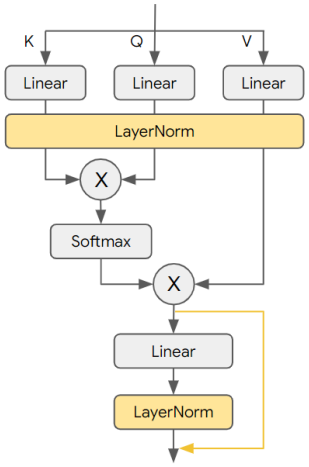


图 *1*。BMT 多头注意力层——与原始 Transformer 的差异以黄色标出。所有线性投影和爱因斯坦求和都可以二值化。

在注意力层中，我们还在线性投影层的输出处添加了捷径连接。在二值神经网络中，由于采用直通估计器，二值化层的梯度是近似得到的。随着二值化层不断堆叠，这最终会使优化朝着不同的方向发展。Liu 等人(2018年)提出在二值神经网络中添加额外的残差连接，这成为部分解决这一问题的一种有效方法。因此，我们在模型中采用了这种方法。需要注意的是，对于 QKV（查询、键、值）线性投影，这种修改并非必要。原始 Transformer 中环绕整个注意力层的快捷连接具有相同作用。我们还将在第 5节的消融实验中展示快捷连接的有效性。

修改后的完整注意力架构如图 1所示，其中突出显示了它与原始架构的差异。额外的层归一化和快捷连接均按元素进行操作，开销很小，尤其是与二值化带来的收益相比。

4. 实验

本节将通过实证评估我们提出的二值化 Transformer 在不同规模机器翻译任务上的表现。为研究二值化不同层的影响，我们首先在 WMT2017 英德翻译数据集（Bojar 等人，2017年）上评估标准的 6 层编码器-解码器（6L6L）Transformer。随后，我们选择 1 比特权重模型变体，研究其在内部翻译

表 *1*。WMT17 英德数据集上的 BMT 结果。二值化的激活值或权重以勾号标记。未标记张量的数据类型保持为 bfloat16。BLEU 评估采用束宽 4。

	注意力 1-BIT						前馈神经网络 1-BIT		指标	
	A_{QKV}	W_{QKV}	A_{OUT}	W_{OUT}	QK 爱因斯坦求和	分数-值爱因斯坦求和	A_{FFN}	W_{FFN}	验证损失	BLEU
1									1.39	26.35
2		✓		✓				✓	1.38	25.93
3							✓	✓	1.40	25.44
4		✓		✓			✓	✓	1.51	24.11
5	✓	✓		✓			✓	✓	1.72	21.55
6		✓	✓	✓			✓	✓	1.60	21.06
7	✓	✓	✓	✓			✓	✓	1.89	17.87
8		✓		✓	✓			✓	1.76	18.27
9		✓		✓	✓	✓		✓	2.81	9.42

数据集上的实际缩放定律。我们还分析了 1 比特模型和浮点模型生成的翻译样本，以比较它们的质量。

4.1. WMT 结果

我们对 Transformer 中五种不同的矩阵乘法进行二值化。在注意力层中，包括：(1) QKV 线性投影；(2) 查询与键之间的激活值-激活值矩阵乘法（QK 爱因斯坦求和）；(3) 注意力分数与值之间的激活值-激活值矩阵乘法（Score-V 爱因斯坦求和）；(4) 输出线性投影。在前馈神经网络中，有两个类型相同的全连接层。为了研究它们各自的影响，我们分别对其权重和激活值进行二值化。在实验中，我们采用以下训练细节。

模型。 我们使用 6L6L Transformer 作为基础模型。嵌入维度为 1024。每个多头注意力层有 16 个注意力头，合并所有注意力头后，QKV 的维度为 1024。前馈网络中的隐藏投影维度为 4096。Dropout 层的 Dropout 率为 0.1。

优化器。 Adam 优化器（Kingma 和 Ba，2014年）与 $\beta_1 = 0.9$ 和 $\beta_2 = 0.98$配合使用。不应用权重衰减。

调度器。 我们采用三阶段训练方案，每个阶段的学习率（LR）都遵循余弦衰减，从基础学习率降至零。每个阶段开始时都会启动一次量化事件。我们首先使用浮点数训练模型。在第二阶段，所有权重都将进行二值化。在最后阶段，权重和激活值都将进行二值化。

损失。 我们在训练过程中应用知识蒸馏（KD）。知识蒸馏可以通过在交叉熵损失函数中使用教师模型经过 softmax 的 logits 替换真实标签来实现，因此用户可以选择是否使用。

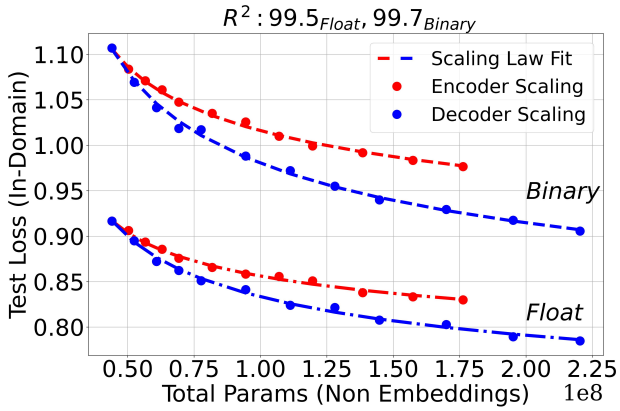
训练。 我们使用 1024 的批量大小。基础学习率为 0。001。第一个学习率周期包含 50000 个步骤，其余周期包含 88339 个步骤。我们使用 4×8TPU 拓扑训练模型。

观察结果。 WMT2017 英德翻译数据集上的评估结果如表 1所示。我们主要依据验证损失来比较模型质量，因为 BLEU 分数的波动较大（Ghorbani 等人，2021年）。从表中我们可以得出以下关键观察结果。

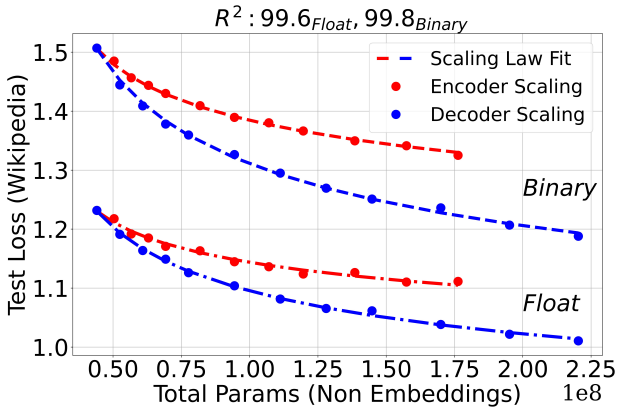
仅权重二值化可保持模型损失不变。 浮点 6L6L Transformer 基线（第 1 行）的验证损失为 1。39。相比之下，将所有全连接层权重（包括注意力层和前馈网络中的权重）进行二值化后，损失甚至更低（1。38，第 2 行），但 BLEU 分数略微下降了约 0。4。这两项指标均表明，1 位权重模型的翻译质量与浮点基线相近。因此，二值化有望在保持质量的同时，将模型大小压缩 16× 。

前馈网络二值化取得了令人鼓舞的结果。 将整个前馈网络进行二值化，即同时对激活值和权重进行二值化，而保留其他层为浮点形式，同样得到了相近的验证损失（1。4，第 3 行），与浮点基线相比没有明显差异。借助我们提出的 BMT，这是机器翻译任务中首次实现前馈网络激活二值化同时保持损失不变。这一引人关注的 1 位前馈网络变体可能对混合专家（MOE）模型十分有用，因为前馈网络占模型总参数量的 50% 至 90%（Lepikhin 等人，2021年）。进一步结合 1 位全连接层权重二值化后，损失降至 1。51 （第 4 行），但 BLEU 分数相比浮点模型低了 2。2 。总体而言，前馈网络二值化展现出了很大的潜力。

注意力激活是影响二值化模型质量的关键瓶颈。在 1-bit 权重和 1-bit 前馈神经网络激活值模型变体的基础上，进一步将注意力层中所有全连接层的输入激活值二值化（第 7 行；包括键、查询、值以及输出投影全连接层的输入激活值），会导致 1.89 的损失。这是模型质量下降幅度最大的情况。将其中每个单独的激活张量二值化，都会使损失至少增加 0.3（第 5 行和第 6 行）。此外，将两个激活间矩阵乘法（查询-键 einsum 运算和注意力分数-值 einsum 运算）二值化尤其具有挑战性。将这两种激活间矩阵乘法也二值化后，1-bit 权重模型的



(a) 域内 $\{p_e, p_d\}$: $\{0.18, 0.31\}_{\text{Float}}$, $\{0.16, 0.28\}_{\text{Binary}}$



(b) 域外 $\{p_e, p_d\}$: $\{0.13, 0.25\}_{\text{Float}}$, $\{0.13, 0.25\}_{\text{Binary}}$

图 2. 对域内和域外数据进行缩放定律研究——对于域内数据，缩放定律拟合的 R^2 值在浮点模型和二值模型上分别达到 99.5 和 99.7。同样，对于域外数据（维基百科），R^2 值分别为 99.6 和 99.8。所有评估数据集上的缩放定律拟合，以及斜率（p_e 和 p_d）见图8 和图9 （附录A）。

激活间矩阵乘法二值化模型仅取得 9.4 的 BLEU 分数（最后一行）。注意力层激活值是实现全二值化翻译模型目前面临的瓶颈。

4.2. 缩放定律研究

尽管第4.1 节展示了令人鼓舞的结果，但一个尚未解答的问题是，二值化 Transformer 扩大规模时，性能是否会下降。众所周知，随着模型大小扩大，神经语言模型的损失遵循幂律（Kaplan 等人，2020年），这被称为“缩放定律”。它被广泛用于预测模型在大规模情况下的性能。因此，我们在内部翻译数据集上对浮点模型和二值化模型进行了缩放定律研究，并比较二者的差异。我们训练了一组翻译模型，并使用以下公式拟合损失，这与 Ghorbani 等人(2021年)的做法类似：

$$L(N_e, N_d) = \alpha \left(\frac{\bar{N}_e}{N_e}\right)^{p_e} \left(\frac{\bar{N}_d}{N_d}\right)^{p_d} + L_\infty,$$

其中 L 是每个 token 的损失，N_e，N_d 分别是编码器和解码器的参数数量。L_∞ 是模型在容量无限时达到的不可约损失。N_e (N_d) 是基线 6L6L Transformer 的参数数量，在曲线拟合过程中用作数值稳定性的归一化常数。为便于分析，我们仅研究权重二值化模型的缩放定律。仅压缩权重的模型也可带来延迟方面的线性提升（Seshadri 等人，2021年），并带来 16× 内存消耗方面的提升（与 blfloat16 相比）。

数据集。 为研究二值模型在容量受限情形下的扩展行为（即性能不受训练数据限制），我们使用内部的大型英语到德语（En → De）平行语料。该训练集包含 30 亿个网络爬取的句对。我们也特别关注在域外（OOD）设置下评估 BMT，并评估其泛化能力，因为先前在图像领域的研究表明，与未压缩模型相比，压缩模型（经过权重剪枝或量化）的质量在 OOD 数据上下降幅度要大得多，也就是说，模型压缩会加剧脆弱性（Hooker 等人，2019）。因此，为了对 BMT 进行稳健评估，我们使用了 11 个评估集，其中一个是域内（ID）数据集，其分布与训练集相似，其余均为分布外（OOD）数据集。对于 ID 数据集，我们抽取 2000 个训练样本，并将其从训练数据中移除。10 个 OOD 评估集分为四类：（i）网络领域；（ii）新闻领域；（iii）维基百科；（iv）专利。此外，这些数据集要么是“原始来源”，要么是“原始目标文本”。原始来源数据集的源端（英语）为自然文本，而目标端（德语）则由人工或机器翻译而成。原始目标文本数据集的目标端（德语）为自然文本，再回译为英语源句。我们作此区分，是为了研究二值化对句子“风格”的影响，因为自然语言具有丰富的多样性，不同于简单直白的翻译腔句子（翻译腔）（Freitag 等人，2020）（更多详情见附录 A.1）。

模型与训练。 我们训练了两组 Transformer，即编码器缩放模型和解码器缩放模型。编码器缩放模型的解码器深度固定为 6 层，同时将编码器深度按 $\{6, 8, 10, 12, 14, 18, 22, 26, 30, 36, 42, 48\}$ 层的规模递增，共得到 12 个模型。解码器缩放模型也采用相同方式，即以类似的方式逐步增加解码器深度。由于训练数据充足，训练过程中我们没有使用标签平滑。二值模型在训练时未使用知识蒸馏。（关于超参数和训练的更多细节，请参见附录 A.2 。）

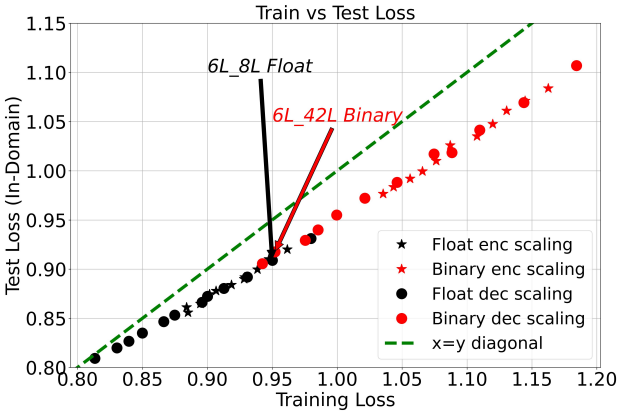


图 *3.* 训练损失与同分布测试损失。我们观察到，所有评估数据集的训练损失与测试损失之间都存在相似的线性关系。

观察结果。 图 2比较了二值模型和浮点模型在同分布和分布外数据集上的缩放曲线，更多内容见附录 A.3。图3 比较了它们的训练损失与域内测试损失。我们得到以下观察结果：

无论是编码器缩放还是解码器缩放，二值模型都表现出与对应浮点模型相似的缩放行为。 图 2a($p_e = 0.16$, $p_d = 0.28$) 中二值模型拟合幂律的指数仅略低于浮点模型（$p_e = 0.18$, $p_d = 0.31$），这表明随着参数数量增加，二值模型的损失会快速下降。这一趋势在图 2b所示的分布外维基百科数据集上同样成立。二值模型在分布外数据上的泛化能力与浮点模型一样好（所有分布外评估数据集的缩放定律拟合见附录 A.3）。我们还注意到二值模型与浮点模型的损失之间存在差距，而这一现象在 WMT 实验中并未观察到。我们推测，这是因为内部生产规模数据集更具挑战性，需要更高的模型容量才能学好。

对于相同的训练损失，二值模型和浮点模型能够达到相同的泛化性能。如图3所示，二值模型和浮点模型的损失在一条直线上良好地对齐，并且几乎重叠于 $0.95 \sim 1.0$ 区域中。未检测到这两类模型的归纳偏置存在可测量的差异。此外，二值模型只需更少的参数位数即可达到特定的性能水平。例如，一个具有195M个参数（195M Bit）的6层42层二值 Transformer，其3× 规模比具有52M个参数（832M Bit）且损失相同的6层8层浮点模型小4倍。这种内存节省在模型部署于资源受限环境中时尤其具有优势（Seshadri 等人，2021年）。

4.3. 生成质量

我们在图中考察机器翻译模型的生成质量4，采用两种解码策略：a）束搜索解码；b）最小贝叶斯风险（MBR）解码（Kumar 和 Byrne，2004年）。

束搜索。 对束搜索解码生成的样本质量[2] 使用标准的去分词 BLEU 分数进行评估（Pa- pineni 等人，2002），使用 sacreBLEU 库（Post，2018年）。[3]。

MBR。Freitag 等人（2022年）表明，束搜索解码选择的是概率较高而非质量较高的样本，尤其是对于大型模型而言，这一点通过人工评估得以衡量。他们提出了如下定义的基于 MBR 的解码策略：

$$h^{\text{MBR}} = \arg\max_{h \in \mathcal{H}} \frac{1}{|\mathcal{H}_{\text{model}}|} \sum_{y \in \mathcal{H}_{\text{model}}} u(h, y)$$

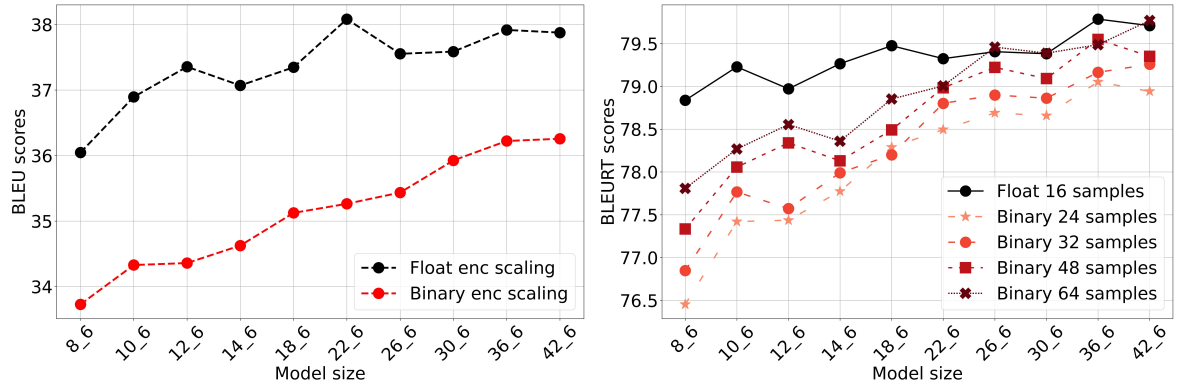
其中， h^{MBR} 是模型根据源句 x生成的解码结果， $\mathcal{H}_{\text{model}}$ 是从模型 $p(.|x)$中采样得到的假设集合， u 是用于评估假设 h 相对于参考译文 y质量的效用函数。Freitag 等人（2022年）展示了 BLEURT 模型（Sellam 等人，2020）作为效用函数的有效性。BLEURT 是一种回归模型，它将假设 h 与参考译文 y 拼接起来，并生成一个介于 [0,1], 之间的标量分数，用于衡量假设的质量，而不受其句子结构、长度或与参考译文的词语重合程度影响。同样，我们使用 BLEURT 作为效用函数进行 MBR 解码，以根据源句解码序列。为了衡量样本质量，我们在评估集中针对给定源句（x），计算解码假设（h_{MBR}）与参考译文（r）之间的 BLEURT(h, r)。随后，对整个评估集上的 BLEURT分数取平均值。

观察结果。图4a 展示了编码器缩放模型的 BLEU 分数（即解码器深度=6，编码器深度各异）。图4b 绘制了编码器缩放模型的 BLEURT 分数，其中基线是采用 MBR 解码、使用 16 个样本的浮点模型。我们观察到以下结果：

二值模型可以用更小的模型大小达到与浮点模型相同的 BLEU 分数。图4a 显示，随着模型大小增加，二值模型的 BLEU 分数会持续提升。尽管在相同模型深度下，二值模型的表现比浮点模型差 2–3 个 BLEU 分数点，但 42L6L 二值模型达到了与 10L6L 浮点模型相同的 BLEU 分数，同时 5×模型大小更小。

增加样本量可以使二值模型的生成质量达到浮点模型的水平。 在图 4b中，较大的样本量始终能提升二值模型的生成质量。当样本量为 4× ，即 64 个样本时，二值模型质量大致与浮点模型相当。此外，随着模型大小增加，二值模型的 BLEURT 得分也会提高。

[2]束宽 =4 =0 6，长度惩罚。[3]大小写混合 +引用数。指数平滑 + 13a分词



(a) 束搜索 - BLEU　　(b) MBR-BLEURT

图 *4.* 编码器缩放时二值化模型与浮点模型的翻译质量对比。(a) 束搜索解码：域内测试集上的 BLEU 分数 (b) MBR解码：域内测试集上的 BLEURT分数

5. 消融实验

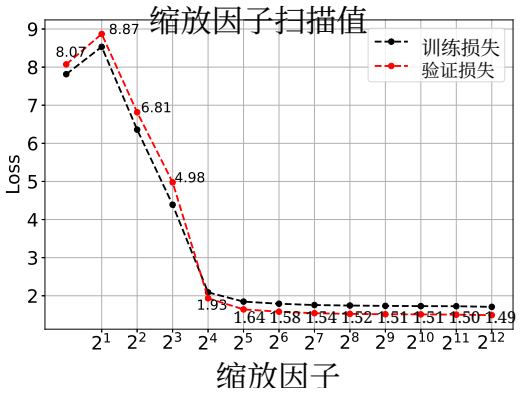


图 *5.* 不同缩放因子下模型的损失 s.

缩放因子消融实验。 我们仅对前馈神经网络进行二值化，并将缩放因子 s 按 2 的幂次从 1 （相当于不应用缩放因子）扫描至4096。我们在图5中绘制最终训练损失和验证损失。

当将 s 增加到 64 时，模型损失会急剧下降。 $s \leq 8$ 的模型产生的译文质量几乎是随机的。较大的缩放因子确实解决了收敛问题。损失在 $s = 64$ 时就已饱和，且仅略差于浮点基线（1.39）。这与我们的预期完全一致，即 $s \propto \sqrt{D}$。当 $s > 64$时，模型损失仍在略微改善。我们推测这是因为界 B 是动态的。由于 Var $(A_b \cdot W_b) \propto B^4$，即使 B 发生微小变化，也会使理论上的最优 s 产生很大偏差。

BMT 注意力层消融实验。我们仅对注意力输出投影线性层进行二值化。模型训练了88339 步，二值化事件从第 50000步开始。我们绘制了从第40000 步开始的损失曲线，如图 6所示。

采用固定的缩放因子可使损失改善近 0.2。这与先前的观察结果一致

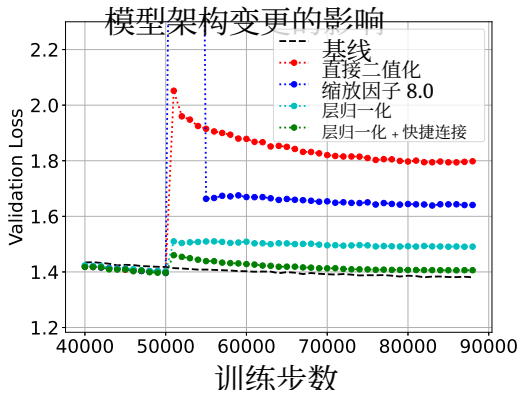


图 *6.* 注意力输出线性层中不同结构的损失 层。

即缩放因子有助于收敛。层归一化可直接替代缩放因子，不仅能使模型收敛到更低的损失，还能在二值化后更快地恢复损失。这是预期之中的，因为 γ 层归一化中的参数是可学习的，因此能够更好地适应动态界 B，如第3.4节所分析。二值化后，损失几乎达到饱和。在输出投影层周围添加快捷连接可以消除信息瓶颈，有助于模型收敛到与浮点基线近似相同的质量。

6. 结论

所提出的方法使机器翻译得以实现二值化。简单而有效的缩放因子是关键。二值化Transformer在缩放行为和翻译质量方面与浮点模型相似。因此，二值化有望成为未来模型服务的一种选择。

尚未解答的问题：如何更好地对注意力的 ein- sums 进行二值化？对于扩展二值 Transformer，增加深度还是宽度更好？如果与 4 位和 8 位量化相结合，哪种混合精度方案更好？

参考译文

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M. 等。TensorFlow：异构分布式系统上的大规模机器学习。*arXiv* 预印本 *arXiv:1603.04467*，2016年。

Aji, A. F. 和 Heafield, K. 使用 4 位精度及更高精度进行神经机器翻译。*arXiv* 预印本 *arXiv:1909.06091*，2019。

Akhbardeh,F.,Arkhangorodsky,A.,Biesialska,M., Bojar,O., Chatterjee, R., Chaudhary, V., Costa-jussa, M.R.,Espa˜ na-Bonet, C., Fan, A., Federmann, C., Freitag, M., Graham, Y., Grundkiewicz, R., Haddow, B., Harter, L., Heafield, K., Homan, C., Huck, M., Amponsah-Kaakyire, K., Kasai, J., Khashabi, D., Knight, K., Kocmi, T., Koehn, P., Lourie, N., Monz, C., Morishita, M., Na- gata, M., Nagesh, A., Nakazawa, T., Negri, M., Pal, S., Tapo, A. A., Turchi, M., Vydrin, V., and Zampieri, M. 《2021 年机器翻译会议（WMT21）论文集》。载于第六届机器翻译会议论文集，第 1–88 页，线上，2021 年 11 月。计算语言学协会。网址：https://aclanthology.org/2021.wmt-1.1。

Ba, J. L., Kiros, J. R., andHinton, G. E. 层归一化。*arXiv* 预印本 *arXiv:1607.06450*，2016年。

Bai, H., Zhang, W., Hou, L., Shang, L., Jin, J., Jiang, X., Liu, Q., Lyu, M., and King, I. BinaryBERT：突破 BERT 量化的极限。*arXiv* 预印本 *arXiv:2012.15701*，2020。

Barrault, L., Bojar, O., Costa-juss`a, M.R., Federmann,C., Fishel, M., Graham, Y.,Haddow, B., Huck, M.,Koehn,P., Malmasi, S., Monz, C., M¨uller, M., Pal, S., Post, M., and Zampieri, M. 2019 年机器翻译会议（WMT19）论文集。在 第四届机器翻译会议论文集（第 *2* 卷：共享任务论文，第 *1* 天）。计算语言学协会，2019。网址：https://aclanthology.org/W19-5301。

Bethge, J., Bartz, C., Yang, H., Chen, Y., 和 Meinel, C. Meliusnet：一种用于二值神经网络的改进网络架构。载于*IEEE/CVF* 冬季计算机视觉应用大会论文集，第 1439–1448 页，2021年。

Bhandare, A.、Sripathi, V.、Karkada, D.、Menon, V.、Choi, S.、Datta, K. 和 Saletore, V. Transformer神经机器翻译模型的高效8位量化。*arXiv* 预印本 *arXiv:1906.00532*，2019年。

Bojar, O.、Graham, Y. 和 Kamran, A. WMT17 指标共享任务的结果。载于第二届机器翻译会议论文集，2017年。

Brock, A., De, S., 和 Smith, S. L. 描述信号传播特性，以缩小未归一化 ResNet 中的性能差距。*arXiv* 预印本 *arXiv:2101.08692*，2021年。

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S. 等。PaLM：利用路径扩展语言建模。*arXiv* 预印本 *arXiv:2204.02311*，2022年。

Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., 和 Bengio, Y. 二值化神经网络：使用权重和激活值约束为+ 1 或- 1的深度神经网络训练。*arXiv* 预印本 *arXiv:1602.02830*，2016年。

Freitag, M., Grangier, D., 和 Caswell, I. BLEU 也许有罪，但参考译文并非无辜。收录于*2020*年自然语言处理经验方法会议（*EMNLP*）论文集。计算语言学协会，2020。网址：https://aclanthology.org/2020.emnlp-main.5。

Freitag, M., Grangier, D., Tan, Q., 和 Liang, B. 重视质量而非模型概率：使用神经指标进行最小贝叶斯风险解码。计算语言学协会会刊，10：811–825，2022年。doi：10.1162/tacl_a_00491。

Ghorbani, B., Firat, O., Freitag, M., Bapna, A., Krikun, M., Garcia, X., Chelba, C. 和 Cherry, C. 神经机器翻译的缩放定律。2021年。

Glorot, X. 和 Bengio, Y. 理解训练深度前馈神经网络的困难。载于第十三届人工智能与统计国际会议论文集，第 249–256 页。JMLR研讨会与会议论文集，2010年。

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A. 等。训练计算最优的大型语言模型。*arXiv*预印本 *arXiv:2203.15556*，2022年。

Hooker, S., Courville, A. C., Dauphin, Y. N., and Frome, A. 选择性脑损伤：衡量模型剪枝的差异化影响。*CoRR*,abs/1911.05248, 2019. URLhttp://arxiv.org/abs/1911.05248。

Jouppi, N. P., Yoon, D. H., Ashcraft, M., Gottscho, M., Jablin, T. B., Kurian, G., Laudon, J., Li, S., Ma, P., Ma, X., 等。借鉴三代产品的经验教训，塑造 Google 的 TPUv4i：工业产品。载于*2021 ACM/IEEE* 第*48*届年度国际计算机体系结构研讨会（*ISCA*），第 1–14 页。IEEE，2021年。

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., 和 Amodei, D. 神经语言模型的缩放定律。*arXiv* 预印本 *arXiv:2001.08361*，2020。Kingma, D. P. 和 Ba, A. Adam：一种随机优化方法。*arXiv* 预印本 *arXiv:1412.6980*，2014年。Kumar, S. 和 Byrne, W. 统计机器翻译的最小贝叶斯风险解码。载于计算语言学协会北美分会人类语言技术会议论文集：*HLT-NAACL 2004*，第 169–176 页，美国马萨诸塞州波士顿，2004年5月2日至5月7日。计算语言学协会。网址：https://aclanthology.org/N04-1022。Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., 和 Chen, Z. {GS}hard：借助条件计算和自动分片扩展巨型模型。载于国际学习表征会议，2021年。网址：https://openreview.net/forum?id=qrwe7XHTmYb。Lew, L., Feinberg, V., Agrawal, S., Lee, J., Malmaud, J., Wang, L., Dormiani, P., 和 Pope, R. Aqt：精确量化训练），2022年。网址：http://github.com/google/aqt。Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., 等。用语言模型解决定量推理问题。*arXiv* 预印本 *arXiv:2206.14858*，2022年。Liu, Z., Wu, B., Luo, W., Yang, X., Liu, W., 和 Cheng, K.-T. Bi-Real Net：通过增强表征能力和改进训练算法提升 1比特卷积神经网络的性能。载于欧洲计算机视觉会议论文集（*ECCV*），第 722–737 页，2018年。Liu, Z., Shen, Z., Li, S., Helwegen, K., Huang, D., 和 Cheng, K.-T. Adam 和训练策略如何助力二值神经网络优化。载于国际机器学习会议，第 6936–6946 页。机器学习研究论文集，2021年。Liu, Z., Oguz, B., Pappu, A., Xiao, L., Yih, S., Li, M., Krishnamoorthi, R., 和 Mehdad, Y. Bit：稳健二值化的多蒸馏Transformer。*arXiv* 预印本 *arXiv:2205.13016*，2022年。Narayanan, D., Shoeybi, M., Casper, J., LeGresley, P., Patwary, M., Korthikanti, V., Vainbrand, D., 和 Catanzaro, B. 使用 Megatron 将语言模型训练扩展至万亿参数，2021年。Papineni, K., Roukos, S., Ward, T. 和 Zhu, W.-J. Bleu：一种用于机器翻译自动评估的方法。载于 计算语言学协会第 *40* 届年会论文集，第 311–318 页，美国宾夕法尼亚州费城，2002年7月。计算语言学协会。doi：10.3115/ 1073083.1073135。网址：https://aclanthology.org/P02-1040。Post, M. 关于清晰报告 BLEU 分数的呼吁。载于 第三届机器翻译会议：研究论文集，第 186–191 页，比利时布鲁塞尔，2018年10月。计算语言学协会。网址：https://www.aclweb.org/anthology/W18-6319。Prato, G., Charlaix, E. 和 Rezagholizadeh, M. 用于机器翻译的全量化 Transformer。*arXiv* 预印本 *arXiv:1910.10485*，2019。Qin, H., Ding, Y., Zhang, M., Yan, Q., Liu, A., Dang, Q., Liu, Z. 和 Liu, X. Bibert：精确的全二值化 BERT。*arXiv* 预印本 *arXiv:2203.06390*，2022。Sellam, T., Das, D. 和 Parikh, A. BLEURT：学习用于文本生成的稳健指标。载于计算语言学协会第 *58* 届年会论文集，第 7881–7892 页，线上，2020年7月。计算语言学协会。doi：10.18653/v1/2020.acl-main.704。网址：https://aclanthology.org/2020.acl-main.704。Seshadri, K., Akin, B., Laudon, J., Narayanaswami, R. 和 Yazdanbakhsh, A. 对用于卷积神经网络的边缘 TPU 加速器的评估。*arXiv* 预印本 *arXiv:2102.10423*，2021。Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. 和 Polosukhin, I. 注意力就是你所需要的一切。神经信息处理系统进展，第 30 卷，2017。Zafrir, O., Boudoukh, G., Izsak, P. 和 Wasserblat, M. Q8bert：量化的 8 位 BERT。载于 *2019* 年第五届能源高效机器学习与认知计算——*NeurIPS* 专刊研讨会（*EMC2-NIPS*），第 36–39 页。IEEE，2019。Zhang, Y., Zhang, Z. 和 Lew, L. Pokebnn：以二值追求轻量级精度。载于 *IEEE/CVF* 计算机视觉与模式识别会议论文集，第 12475–12485 页，2022年。

A. 缩放定律研究详情

A.1. 数据集

用于缩放定律的评估数据集概览（第 4.2 节）见表格。十个分布外评估数据集涵盖四类：(i) 网络领域，(ii) 新闻领域，(iii) 维基百科，(iv) 专利。它们属于“原始来源”或“原始目标”数据集。网络领域包含两个原始来源数据集和一个原始目标数据集；维基百科和专利领域则各包含一个原始来源数据集。新闻领域采用公开可用的 WMT newstest2019（Barrault 等，2019）和 WMT newstest2021（Akhbardeh 等，2021）。该领域包含五个数据集：来自 WMT newstest2019（Freitag 等人，2020）的原始来源、原始目标、源端原文释义（Freitag 等人，2020）和高质量源端原文（Barrault 等，2019），以及来自 WMT newstest2021（Akhbardeh 等，2021）的 wmt-reference-C。

数据集名称	领域	Type	来源
训练子集	Web	混合	内部
专利	专利	混合	内部
网络领域1	Web	原始来源	内部
网络领域2	Web	原始来源	内部
网络领域3	Web	原始目标文本	内部
维基百科	维基百科	原始来源	内部
WMT 高质量	News	原始来源	WMT newstest2019（Freitag 等人，2020）
WMT 参考语料 C	News	原始来源	WMT newstest2021 Ref-C（Akhbardeh 等，2021年）
WMT释义版	News	原始来源	WMT newstest2019（Freitag 等人，2020）
WMT原始源文本	News	原始来源	WMT newstest2019（Barrault 等，2019）
WMT原始目标文本	News	原始目标文本	WMT newstest2019（Barrault 等，2019）

表2. 第4.2节使用的评估数据集。

A.2. 模型与训练详情

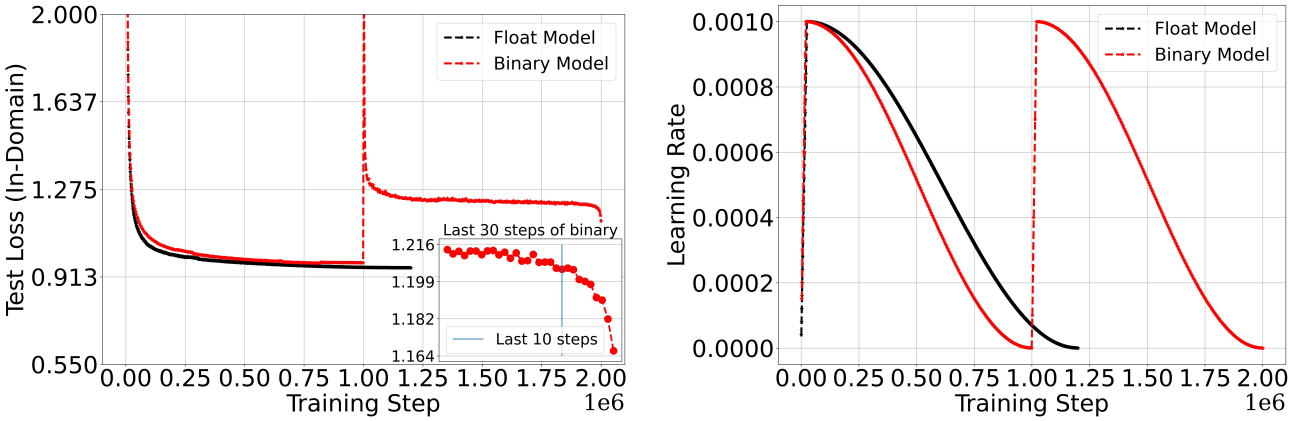
第 4.2 节中的所有模型，其嵌入维度均为 512，隐藏投影维度为 2048，注意力头数为 8。来源端和目标端共享嵌入参数。同一个嵌入矩阵（转置后）也用于解码器端线性读出（softmax）参数。所有模型均使用 Adam 优化器（Kingma 和 Ba，2014年），并采用余弦学习率调度。由于训练数据充足，我们在训练过程中未使用标签平滑。在实验中，启用标签平滑会导致所有模型在开发集上的表现都不佳。一个模型（6 个编码器层、8 个解码器层）的训练和学习率曲线如图 7所示。浮点模型训练 5 个 epoch，二值模型则分两个阶段训练 9 个 epoch：浮点阶段和二值化阶段。在二值模型训练的两个阶段中，均使用独立但相同的学习率调度（含预热）。值得注意的是，二值模型的大量训练（即损失下降）发生在最后 10 步，此时学习率极低。最后 15 步的学习率原始值为[5.0e-7, 3.1e-7, 1.6e-7, 6.4e-7, 1.0e-8, {2.5e-15}x10]。我们还尝试在最后一个 epoch 中为二值模型设置恒定学习率，取值为{1e-8, 1e-11, 1e-15}，以替代原有调度，但观察到模型质量下降（损失进入平台期）。Liu 等人(2021年)；Zhang 等人(2022年)也观察到了这一现象：二值模型训练的最后阶段，即使学习率极低，仍会发生显著的学习。我们将对这一现象的进一步研究留待未来工作。

A.3. 缩放定律拟合

所有十个 OOD评估数据集上的缩放定律拟合结果如图 8所示。斜率 p_e 和 p_d 如图 9 和表 3所示。

B. 生成质量

解码器缩放模型的生成质量如图 10所示。我们观察到的行为与第 4.3节中编码器缩放模型的表现类似。在相同模型深度下，二值模型的 BLEU 分数比相应的浮点模型低 2–3 个 BLEU 分数。基于 MBR-BLEURT 的解码质量会随着样本量的增加而持续提升。



（a）损失　　（b）学习率

图7。训练过程中，6层8层浮点模型和二值模型的测试损失与学习率曲线。

数据集	浮点模型		二值模型	
	p_e	p_d	p_e	p_d
训练子集	0.18	0.31	0.16	0.28
专利	0.20	0.30	0.19	0.32
网络领域1	0.14	0.25	0.14	0.27
网络领域2	0.19	0.37	0.16	0.30
网络领域3	0.12	0.18	0.14	0.23
维基百科	0.13	0.25	0.12	0.25
WMT 高质量	0.20	0.31	0.18	0.30
WMT 参考语料 C	0.24	0.34	0.17	0.27
WMT释义版	0.14	0.36	0.12	0.31
WMT原始源文本	0.22	0.37	0.23	0.36
WMT原始目标文本	0.15	0.22	0.12	0.20

表3。 相同数据的表格表示（p_e & p_d），如图 9所示。

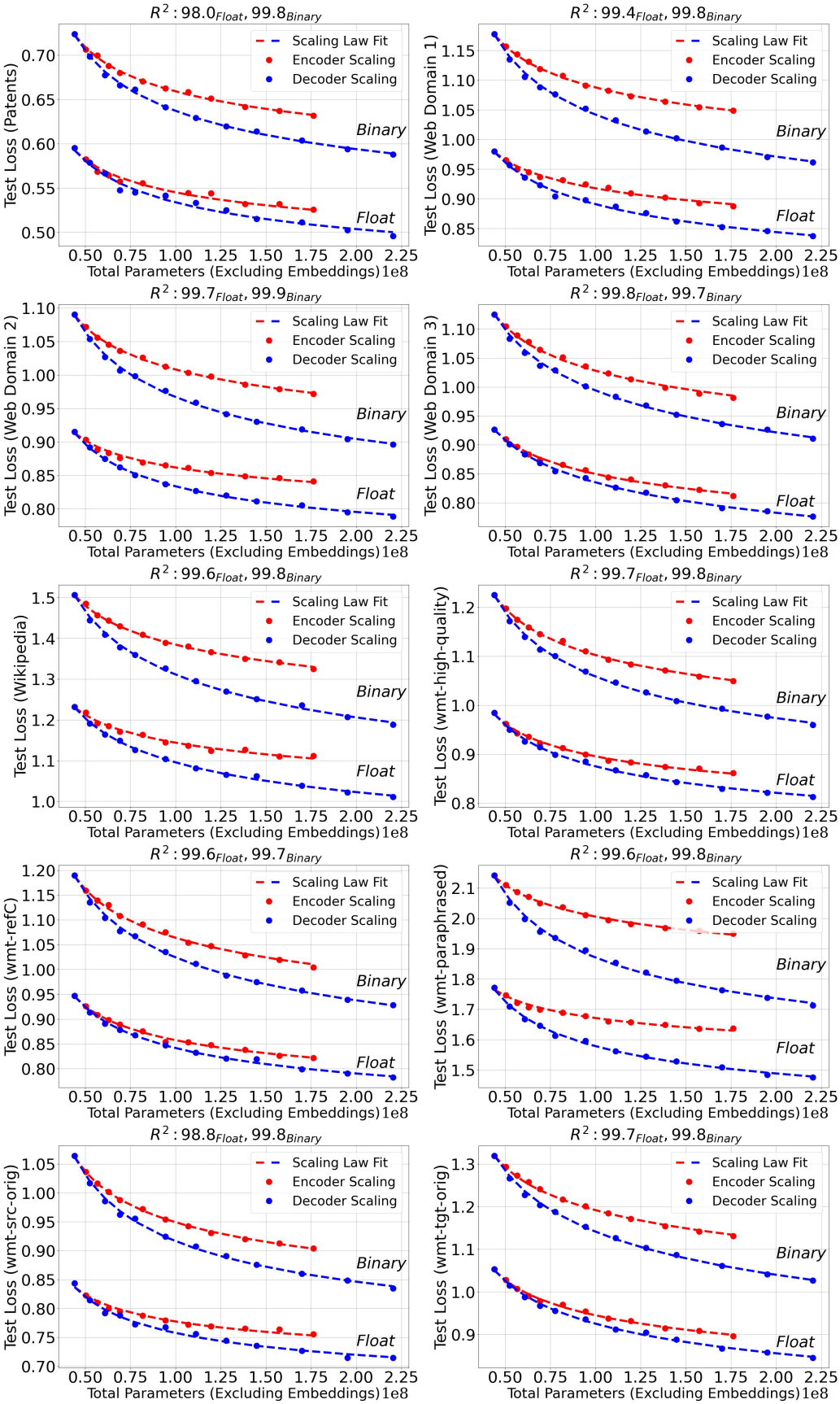


图8。 第 4.2 节定义的评估数据集上的 BLEU 分数4.2

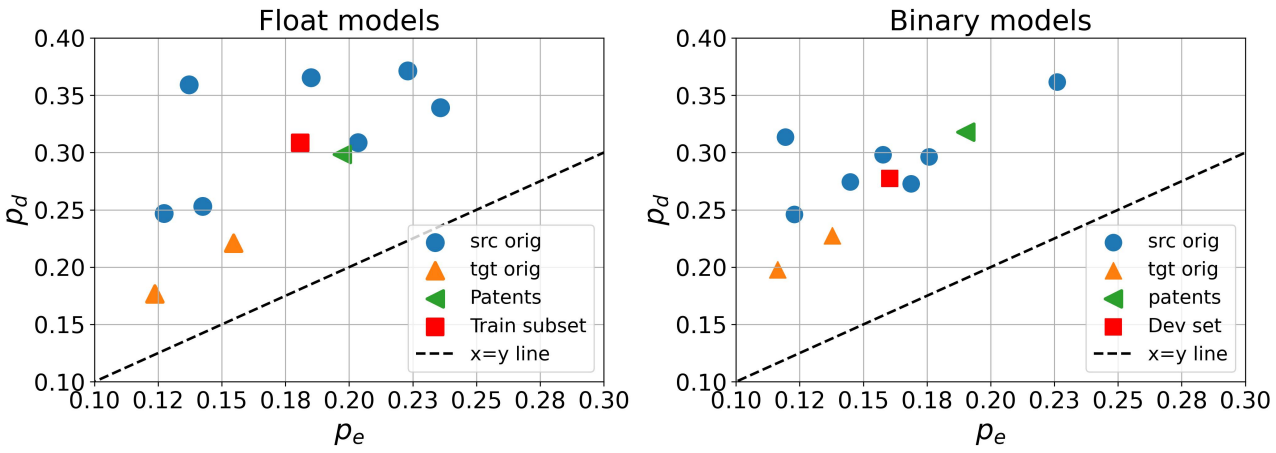
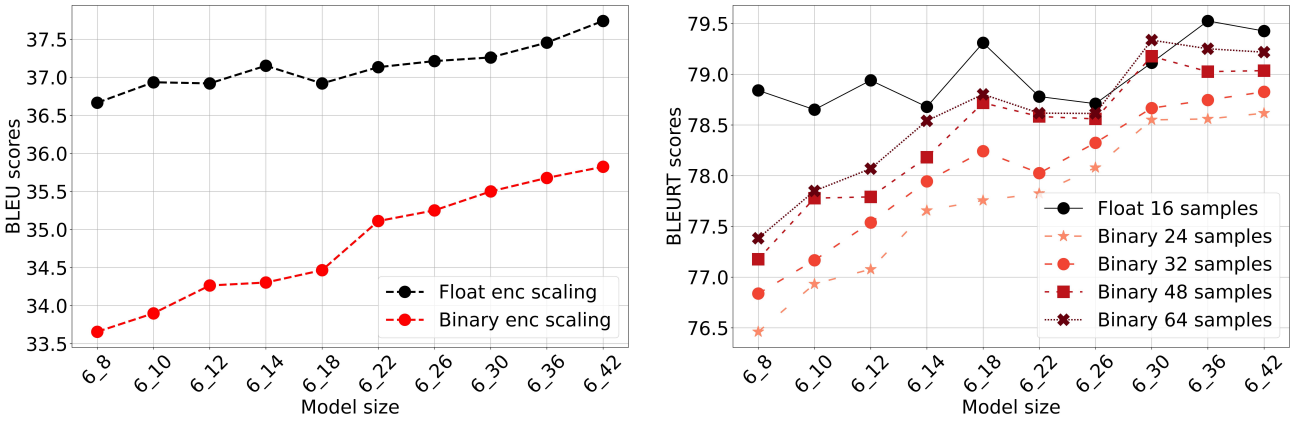


图*9*。根据第 p_e 和 p_d节定义的缩放定律，编码器和解码器的缩放斜率（即 4.2）。原始值见表 3。



（a）束搜索 BLEU　　（b）MBR-BLEURT

图*10*。 解码器缩放中，二值化模型与 bfloat16模型的翻译质量对比。