基于Transformer和蒸馏模型实现英语到匈牙利语的翻译项目

班级：B22人工智能班

组长：廖雪锋

组员：莫绍成，韦德力，李祖军

**摘 要：**本研究聚焦于Transformer模型及其在蒸馏学习框架下的优化应用，探讨其在英法机器翻译任务中的表现。我们首先介绍了原始Transformer结构的核心机制与优势，然后基于大规模平行语料，构建了教师-学生模型架构，对翻译任务进行了知识蒸馏训练。研究过程中详细说明了数据获取、清洗与分词等预处理流程，并设置多种实验参数进行对比分析。实验结果表明，蒸馏后的学生模型在保持较高准确率的同时，大幅降低了模型体量和推理时间，适用于资源受限环境。然而，在处理长距离依赖和语义歧义句时，模型仍存在性能波动。针对上述问题，本文进一步进行了误差分析，并提出改进路径，包括引入自监督预训练、多任务学习机制和外部知识嵌入等策略。研究结果验证了蒸馏机制在提升模型可部署性方面的有效性，为高效、轻量级神经机器翻译系统的构建提供了理论支持和实践基础。本研究采用教师模型指导学生模型训练，并引入蒸馏损失与真实标签损失的平衡机制，以提升学生模型性能与稳定性。

**关 键 词：**权重Transformer模型;机器翻译; 数据预处理; 性能评估; 错误分析;改进策略;教师模型;学生模型;平衡训练

**0 引言**

近年来，神经机器翻译（Neural Machine Translation, NMT）技术取得了令人瞩目的进展，尤其是基于Transformer结构的模型在多个主流语言对上展现出卓越的翻译性能。自Vaswani等人于2017年提出Transformer架构以来，该模型凭借其完全基于注意力机制的设计，成功克服了循环神经网络（RNN）在长距离依赖建模方面的缺陷，并迅速成为NMT任务中的主流架构。随着计算资源的不断提升和大规模语料库的积累，基于Transformer的模型不仅在BLEU等自动评估指标上屡创佳绩，也在多项国际翻译评测任务（如WMT）中屡次刷新性能纪录。特别是在高资源语言对（如英语-德语、英语-法语）上，Transformer几乎已经成为工业级翻译系统的首选基础架构。

然而，Transformer模型在取得高精度翻译效果的同时，也带来了模型体积庞大、参数量巨大、训练与推理过程计算开销高等问题。例如，标准的Transformer-base模型参数量约为65M，而Transformer-big更是超过200M。在现代NMT系统中，为了进一步提高性能，研究者们常引入更深的网络结构、更大的词汇嵌入、更复杂的解码策略和集成机制，这在提升翻译质量的同时，也显著增加了模型的资源依赖。对于GPU、TPU等硬件充裕的场景来说，这些计算需求尚属可接受；但在移动设备、边缘计算平台以及低带宽服务器等资源受限环境下，部署此类大型模型则面临重重挑战。

尤其是在小语种翻译领域，如英语到匈牙利语的翻译任务中，这一矛盾显得更加突出。一方面，匈牙利语属于芬兰-乌戈尔语系，具有复杂的形态变化和自由语序等特点，使其翻译任务本身充满挑战；另一方面，由于训练数据的稀缺以及产业关注度相对较低，相比主流语言对，研究者在模型设计和优化策略上也面临更多限制。因此，如何在有限的数据和计算资源条件下，实现对小语种的高质量翻译，成为当前神经机器翻译研究中一个极具现实意义的问题。

为了解决大型模型带来的资源消耗问题，同时提升小语种翻译的可用性，模型压缩技术成为一个重要研究方向。在众多模型压缩方法中，知识蒸馏（Knowledge Distillation）因其简单高效、易于与现有训练流程集成的特点，受到广泛关注。知识蒸馏最早由Hinton等人提出，其核心思想是通过构建教师-学生模型架构，将性能较高、结构复杂的教师模型中所蕴含的知识迁移给结构更为简洁、计算代价更低的学生模型。在NMT领域，知识蒸馏的应用表现为用教师模型生成伪标签、蒸馏中间表示、蒸馏注意力分布等方式，使得学生模型在保持一定翻译质量的基础上，大幅缩减参数量和推理时间。大量研究表明，通过蒸馏策略训练出的轻量模型，不仅在BLEU等指标上接近甚至超过未蒸馏的原生模型，同时在运行效率方面也表现出显著优势。

本文以英语到匈牙利语的神经机器翻译任务为研究对象，基于Transformer架构设计教师-学生模型系统，并在不同蒸馏策略的基础上进行对比实验。首先，选取在英语-匈牙利语翻译任务中表现优异的预训练或大规模训练的Transformer模型作为教师，其模型参数丰富、上下文建模能力强，能够输出高质量的翻译结果。然后，构建结构紧凑、计算代价较低的学生模型，通过序列级蒸馏、词级蒸馏、注意力蒸馏等多种方式，从教师模型中提取翻译知识并引导学生模型学习。通过系统的实验设计，本文不仅关注蒸馏策略对翻译质量的影响，还详细评估了各类学生模型在模型大小、推理速度、资源消耗等方面的表现，全面探讨知识蒸馏在小语种轻量翻译系统中的可行性与适应性。

本研究的目标不仅仅是构建一个高性能的轻量化NMT系统，更重要的是验证知识蒸馏作为压缩手段在低资源语言环境下的有效性与普适性。通过教师模型向学生模型的知识迁移，是否能缓解训练数据稀缺带来的泛化能力下降问题？不同类型的蒸馏方法在小语种翻译中表现是否具有一致性？学生模型是否能够兼顾模型精度与推理效率？这些问题都是本文尝试深入探讨的研究重点。最终，我们希望通过系统的实验验证，提供一种适用于小语种、高效率部署场景下的神经机器翻译方案，为实现多语言、低资源环境下的实用翻译系统奠定基础。

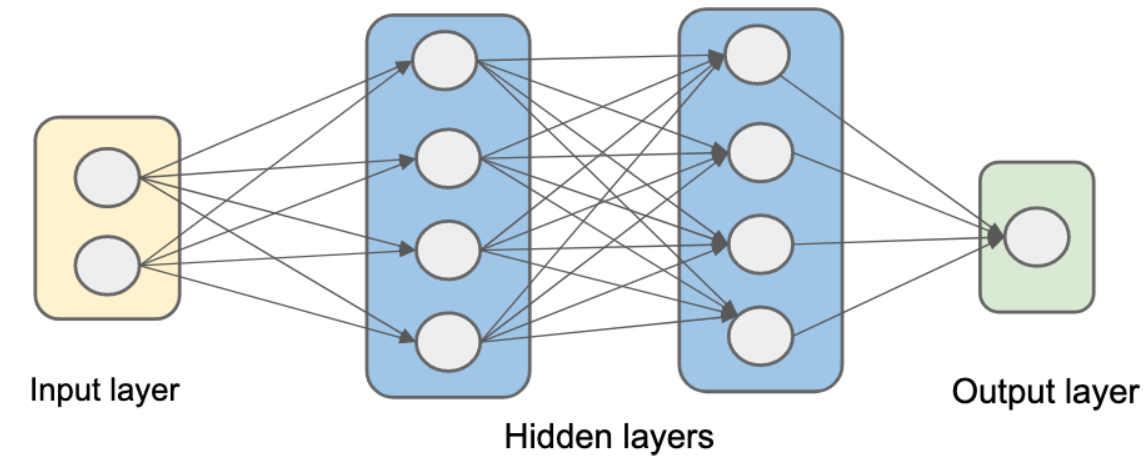
在全球多语言交流日益频繁的当下，尤其是在“一带一路”等国际合作倡议不断推进的背景下，小语种翻译系统的实用性和可用性将直接影响跨文化信息的流通效率。知识蒸馏作为一种有效提升轻量模型性能的手段，既满足了性能要求，又降低了部署成本，具有广泛的现实应用价值。因此，深入研究蒸馏方法在NMT中的适用性，特别是在英语–匈牙利语这类典型小语种方向的具体表现，对于提升全球语言服务能力、构建高效智能的翻译基础设施具有重要意义。

### 1 Transformer和蒸馏模型

#### 1.1 神经网络

近年来，随着深度学习技术的不断发展，**基于 Transformer 与蒸馏模型**的机器翻译方法显著提升了翻译的效率与准确性。特别是序列到序列（Seq2Seq）框架的广泛应用，为神经机器翻译任务提供了基础架构支持。在众多新方法中，Transformer模型作为一种非循环的神经网络结构，由Vaswani等人于2017年提出，其核心创新在于引入了自注意力机制与位置编码，有效捕捉长距离依赖关系，并在多项自然语言处理任务中展现出优越性能，尤其在机器翻译领域表现尤为突出。

相比传统的循环神经网络（RNN）和长短期记忆网络（LSTM），Transformer模型[4,5]不仅在处理长序列信息方面更加高效，而且具备更强的并行计算能力和更快的训练速度，因此迅速成为当前机器翻译研究的主流架构之一。然而，Transformer模型参数规模庞大、计算资源消耗大，为了提升其在实际部署中的可行性，研究者引入了**知识蒸馏（Knowledge Distillation）技术**。通过构建性能强大的教师模型（Teacher），将知识迁移至更轻量的学生模型（Student），在大幅度压缩模型体积的同时保持较高翻译质量，为资源受限环境下的机器翻译提供了有效解决方案。神经网络如图1所示：



**图1 神经网络**

本研究旨在通过基于Transformer和蒸馏模型模型的深入分析和实验验证，进一步提高英语到匈牙利语机器翻译的质量和效率。模型性能提升： 通过优化Transformer模型的架构和训练策略，力求在英语到匈牙利语翻译任务中提高翻译的准确性、流畅度和语义一致性。这包括调整模型的深度、宽度和层次结构，探索最佳的注意力机制配置以及优化训练过程中的超参数选择。

#### 1.2 introduction

神经机器翻译（Neural Machine Translation, NMT）作为自然语言处理领域的重要分支，其发展经历了从统计机器翻译（SMT）到神经网络翻译的重大转变。早期的翻译系统以短语对齐和规则为基础，依赖于复杂的语言模型与人工特征工程，存在语法不连贯、上下文理解能力弱等问题。随着深度学习的发展，基于递归神经网络（RNN）的NMT方法逐步取代传统统计方法，显著提高了翻译质量。

2017年，Google提出了Transformer模型，它完全摒弃了循环结构，采用多头注意力机制（Multi-Head Attention）和位置编码（Positional Encoding）实现序列建模。这一架构在多个机器翻译任务中刷新了性能记录，成为现代NMT的主流模型。相比RNN，Transformer在并行计算和长距离依赖建模方面具有显著优势。然而，Transformer模型的高计算复杂度和庞大的参数规模也带来了较高的部署门槛，尤其在边缘设备或低资源场景中应用受限。

为缓解上述问题，研究者开始关注模型压缩技术，其中知识蒸馏（Knowledge Distillation）成为重要手段之一。该技术最初由Hinton等人提出，主要思想是通过训练一个容量较小的“学生模型”来模仿“教师模型”的行为，从而实现模型轻量化。知识蒸馏在图像识别、语音识别等领域已得到广泛应用，在神经机器翻译中也显示出良好的适应性。在机器翻译领域，知识蒸馏通常包括两种方式：**序列级蒸馏（Sequence-level Distillation）和词级蒸馏（Token-level Distillation）**。前者通过教师模型生成目标句对的译文作为训练样本，后者则直接使用教师模型输出的Soft Label作为训练信号。在实际应用中，结合两者效果更为理想。此外，为了进一步优化蒸馏效果，研究者引入了**平衡训练机制**，即在蒸馏损失与真实标签损失之间设置权重，以实现知识迁移与目标学习的统一。

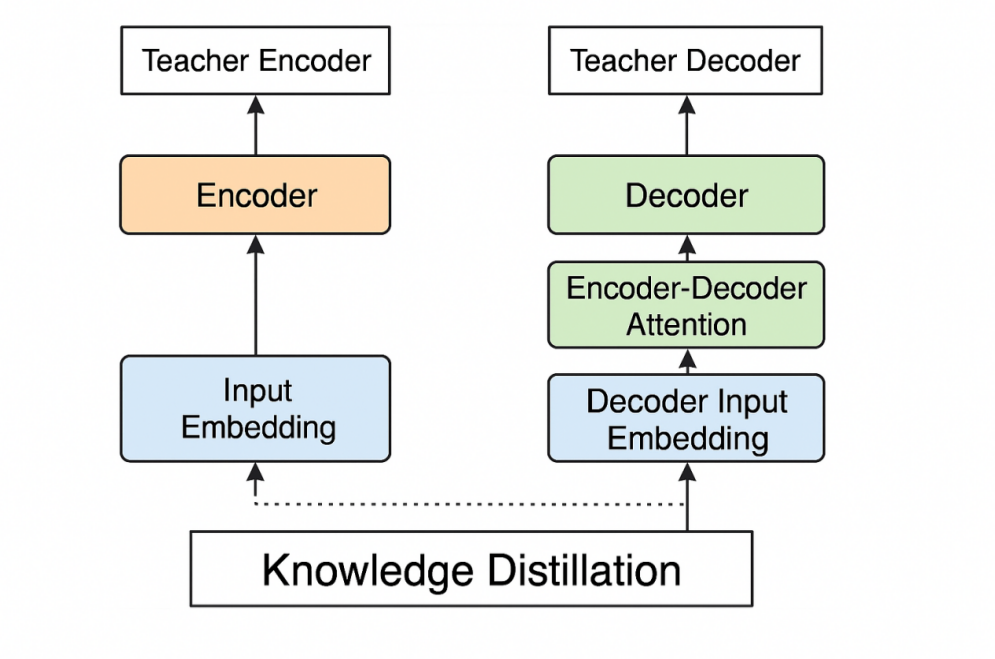
针对英语到匈牙利语的翻译任务，研究面临更多挑战。匈牙利语属于乌拉尔语系，具有复杂的词形变化和丰富的语法结构，与英语在语言结构和语序上差异较大。这种语言跨度较大的翻译任务对模型的表达能力与泛化能力提出了更高要求。传统NMT方法在小语种翻译任务中普遍存在数据稀缺和泛化能力差的问题。因此，将Transformer与知识蒸馏技术相结合，构建高效、轻量的翻译系统成为研究重点。当前，国内外在该方向的研究不断深入。一方面，多语种预训练模型（如mBART、mT5）提供了跨语言迁移能力，增强了低资源语言的建模效果；另一方面，结合蒸馏机制的学生模型逐渐在多种部署场景中展现应用潜力，如移动端翻译、在线翻译插件等。此外，研究者还尝试将自监督学习、多任务学习、外部知识嵌入等策略与蒸馏方法相融合，以进一步提高翻译质量和模型稳定性。

综上所述，Transformer与蒸馏模型在英语-匈牙利语翻译任务中具备广阔的发展前景。随着数据资源的丰富、算法的迭代以及部署需求的提升，未来轻量化、高性能、多语言兼容的NMT系统将成为主流方向。本文所提出的教师-学生架构与平衡训练机制，正是对这一趋势的有力探索，旨在为复杂语言对的神经翻译提供更加可行、高效的解决方案。

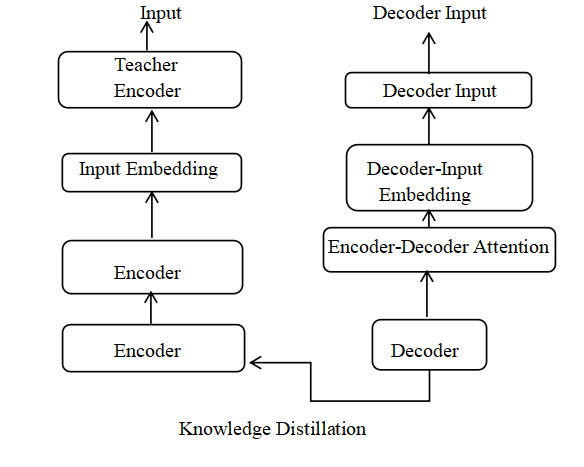
#### 1.3 Transformer和蒸馏模型在机器翻译中的应用与优势

自2017年Vaswani等人提出Transformer模型以来，它迅速在机器翻译以及自然语言处理的其他任务中占据主导地位。Transformer的核心创新在于完全基于注意力机制的架构，彻底摆脱了循环神经网络（RNN）与卷积神经网络（CNN）的限制，不仅提升了训练并行性，还显著改善了翻译效果。然而，Transformer模型在带来高性能的同时，也引入了大量参数和较高的计算成本，限制了其在资源受限环境中的部署。为解决这一问题，知识蒸馏（Knowledge Distillation）被广泛应用于模型压缩中。通过构建“教师模型”和“学生模型”的训练框架，学生模型可以在保持性能的同时大幅减少参数量，实现轻量化部署。Transformer蒸馏架构如图2与图3所示。整体结构仍由编码器与解码器组成，教师模型与学生模型采用相似的结构，但参数规模不同。在训练阶段，教师模型首先完成训练，并生成中间表示和预测输出，作为“软标签”指导学生模型的训练。训练过程中引入了**平衡损失函数机制**，即将教师模型输出（蒸馏损失）与真实标签（硬损失）按一定权重结合，提升学生模型的泛化能力与稳定性。

图2展示了**编码器部分**，包括学生与教师的输入嵌入、多头自注意力和前馈网络模块；  
图3则展示了**解码器部分**，包括解码输入嵌入、编码器-解码器注意力机制以及最终输出的生成过程。图中虚线箭头表示蒸馏信息的传递路径，体现了知识由教师向学生迁移的过程。



**图2 编码器**



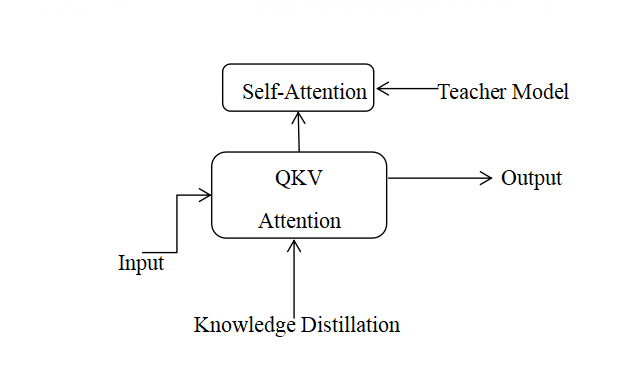
**图3 解码器**

自注意力机制（Self-Attention）是Transformer架构的核心创新之一。在蒸馏框架下，该机制不仅用于捕捉输入序列中各词之间的依赖关系，还作为知识迁移的关键环节，通过教师模型输出的注意力分布，引导学生模型进行学习与优化。在该结构中，输入首先通过线性变换生成三个矩阵：查询（Q）、键（K）和值（V）。注意力权重通过计算 Q 与 K 的点积并进行缩放和归一化后，应用于 V，形成加权表示。这使得每个位置的表示都能动态感知全局上下文。

**自注意力机制核心公式**： (1)

在知识蒸馏过程中，教师模型中的自注意力模块输出其注意力权重与表示，作为软目标（Soft Target）提供给学生模型。学生模型通过对比自己的注意力分布与教师模型的输出，进行梯度优化，最终达到在参数更少的情况下复现相似表现。

如图4所示，知识蒸馏在自注意力机制层面上，直接作用于Attention模块，使学生模型能够学习到教师模型中的表示模式和上下文建模能力，从而实现更轻量但高效的神经网络。



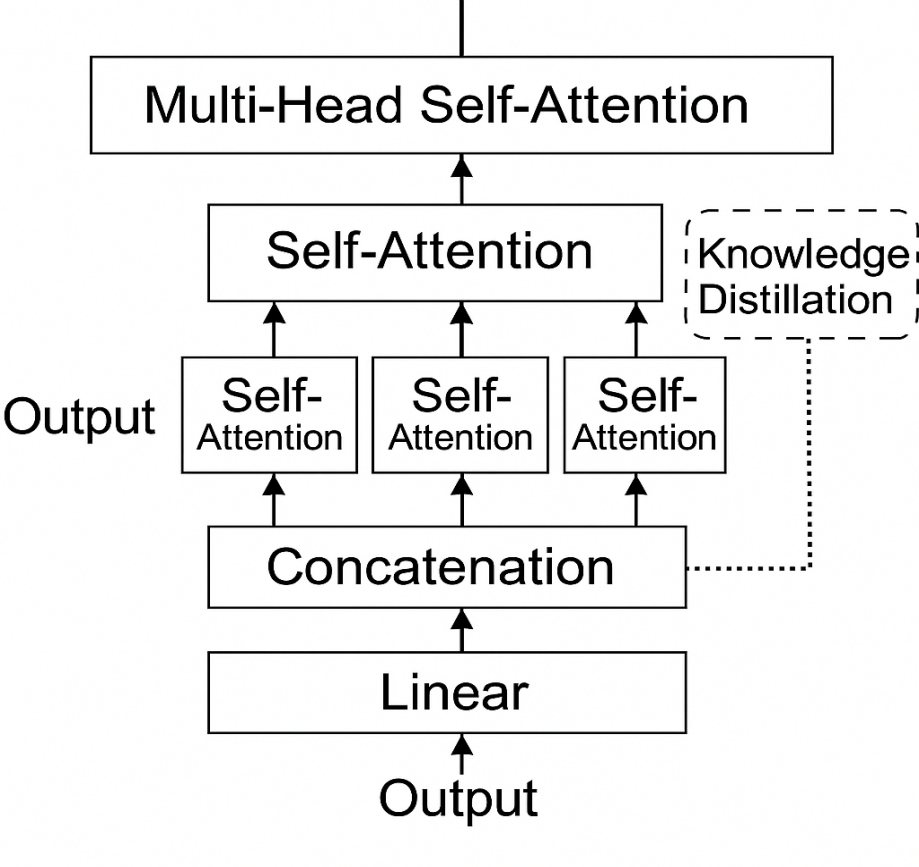
**图4 知识蒸馏**

多头自注意力机制（Multi-Head Self-Attention）是Transformer结构中的关键组成部分，在蒸馏模型中同样扮演着核心角色。相比单一注意力机制，多头注意力允许模型在多个子空间中并行地建模不同的语义关联，有助于捕捉更全面的上下文信息，特别是在长句翻译中更为有效。在蒸馏框架中，教师模型和学生模型都会构建自身的多头注意力结构。教师模型通常具有更多的头数与更宽的维度，因而能学习更细致、复杂的语义关系。学生模型则通过对比学习，模仿教师模型中每一个注意力头所输出的权重矩阵，从而在参数更少的情况下获取相似的上下文理解能力。

(2)

其中，, 为学习矩阵

图5展示了该机制的基本流程：输入经过多个并行的自注意力子层处理（即多个“头”），各自输出不同的上下文特征表示。随后，这些特征通过拼接（Concatenation）与线性变换整合为最终表示。知识蒸馏模块将教师模型每个头的注意力输出作为软目标引导学生模型学习，从而在性能与效率之间取得平衡。通过引入知识蒸馏，多头自注意力机制不仅增强了模型的表达能力，也提升了学生模型的训练效率和部署适应性，尤其适合资源受限环境中的神经机器翻译应用。



**图5 基本流程**

由于Transformer模型虽然具备强大的建模能力和并行计算优势，但其庞大的参数规模和计算资源需求，常常限制了其在资源受限环境中的部署。为了解决这一问题，研究者提出了知识蒸馏（Knowledge Distillation）技术，旨在在保持模型性能的同时显著减少模型的体积和推理成本。知识蒸馏通过将一个性能强大的“教师模型”的知识传递给一个结构更轻量的“学生模型”，提升后者的表现力。

在Transformer结构中，蒸馏技术的引入使得原本庞大的模型可以“压缩”成更小、更高效的版本，从而更适合在边缘设备或实时系统中运行。蒸馏过程中，学生模型不仅学习真实标签带来的监督信息，还通过模仿教师模型输出的概率分布（软标签），获取更多隐藏的结构和语义信息。位置编码（Positional Encoding）仍然是Transformer模型中不可或缺的一部分，用于在蒸馏过程中维持模型对序列位置信息的敏感性。无论是固定还是可学习的位置编码方式，在知识迁移过程中都对学生模型理解语序关系至关重要。将蒸馏机制与Transformer相结合，在机器翻译中具有以下几方面的优势：

Transformer模型虽然在神经机器翻译任务中展现出强大的建模能力，但其高昂的计算成本和庞大的参数量限制了其在实际部署中的广泛应用。通过知识蒸馏技术所得到的学生模型，在保持接近翻译质量的前提下，大幅精简了模型结构，参数规模显著缩小，推理过程更为高效。这种结构压缩使得学生模型能够在计算资源有限的环境中运行，如移动设备或边缘计算平台，从而大大提升了模型的实用性与工程可落地性。同时，推理速度的加快也使得其更适用于实时翻译等对响应时间要求较高的场景。

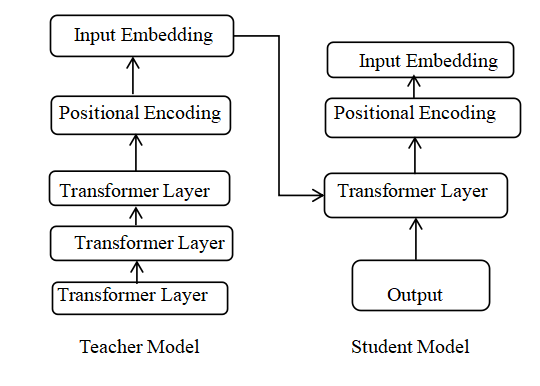
尽管学生模型在结构上进行了简化，但通过蒸馏过程中对教师模型的多头自注意力机制进行学习，其在理解输入序列的长距离依赖关系方面依然表现出色。尤其在处理具有复杂语法结构或语序灵活的小语种（如匈牙利语）时，这种长距离建模能力尤为关键。学生模型通过模仿教师在上下文建模中的表现，能够捕捉输入中远隔词语之间的语义关联，从而有效传承Transformer处理长文本序列的优势，不牺牲原有的语言理解能力。

教师模型往往是在大规模、多样化语料库上进行充分训练，具备极强的语义表达和泛化能力。在蒸馏过程中，这种语义知识通过软标签、特征对齐、注意力图等形式被传递给学生模型，使得后者在缺乏大参数支持的前提下，仍能获得良好的上下文理解与任务适应能力。特别是在数据相对稀缺的小语种翻译任务中，这种蒸馏过程能够有效弥补学生模型因数据不足带来的性能劣势，显著提升其在未见样本上的表现稳定性。

通过将蒸馏策略与模型设计相结合，Transformer系统的整体部署灵活性得到了显著增强。用户可以根据实际应用场景动态选择教师与学生模型的搭配策略。例如，在服务器端使用完整的教师模型进行大批量训练与推理输出，而在终端侧或实时交互系统中部署经过蒸馏优化的轻量学生模型，以节省资源和提升响应效率。这种“训练-部署分离”的模式不仅提高了系统扩展能力，也为多终端、多需求环境下的应用提供了更具弹性的解决方案。

知识蒸馏方法本身具有很强的可拓展性，除了最常用的基于序列输出的传统蒸馏形式外，还可以融合多种更细粒度的蒸馏策略，如中间层蒸馏、注意力分布蒸馏、损失函数融合蒸馏等。这些策略不仅关注最终输出，还强化了学生模型对教师模型内部结构的学习，提升其在特征提取与表示能力方面的表现。通过引入这些深层蒸馏机制，学生模型不仅在翻译质量上趋近甚至超越原始小模型，还进一步拓展了其适应复杂任务结构的能力，从而使NMT系统在参数-性能-效率之间达到更优的平衡。

综上所述，将知识蒸馏与Transformer结合，不仅有效压缩模型规模，还在保证性能的前提下提升了模型的实际应用价值。正如图6所示，整个蒸馏过程体现了“教师模型-学生模型”的知识迁移与学习机制。



**图6蒸馏过程**

总之**，Transformer 与蒸馏模型的结合**在机器翻译中的应用展现出强大的潜力和广泛的优势。通过引入知识蒸馏机制，可以有效压缩Transformer模型的规模，在保持翻译质量的同时显著降低计算成本，使模型更适合部署在资源受限的设备或实时场景中。同时，得益于Transformer的自注意力机制和多头注意力机制，学生模型在蒸馏过程中仍能捕捉长距离依赖信息，确保翻译结果的准确性与流畅度。此外，蒸馏方法赋予模型更强的可扩展性与适应性。通过灵活配置教师与学生模型结构，可以根据不同语言对、数据规模或硬件环境进行有效调整，满足多样化翻译任务的需求。其模块化设计也为后续引入预训练、迁移学习、多任务学习等技术提供了良好的基础。

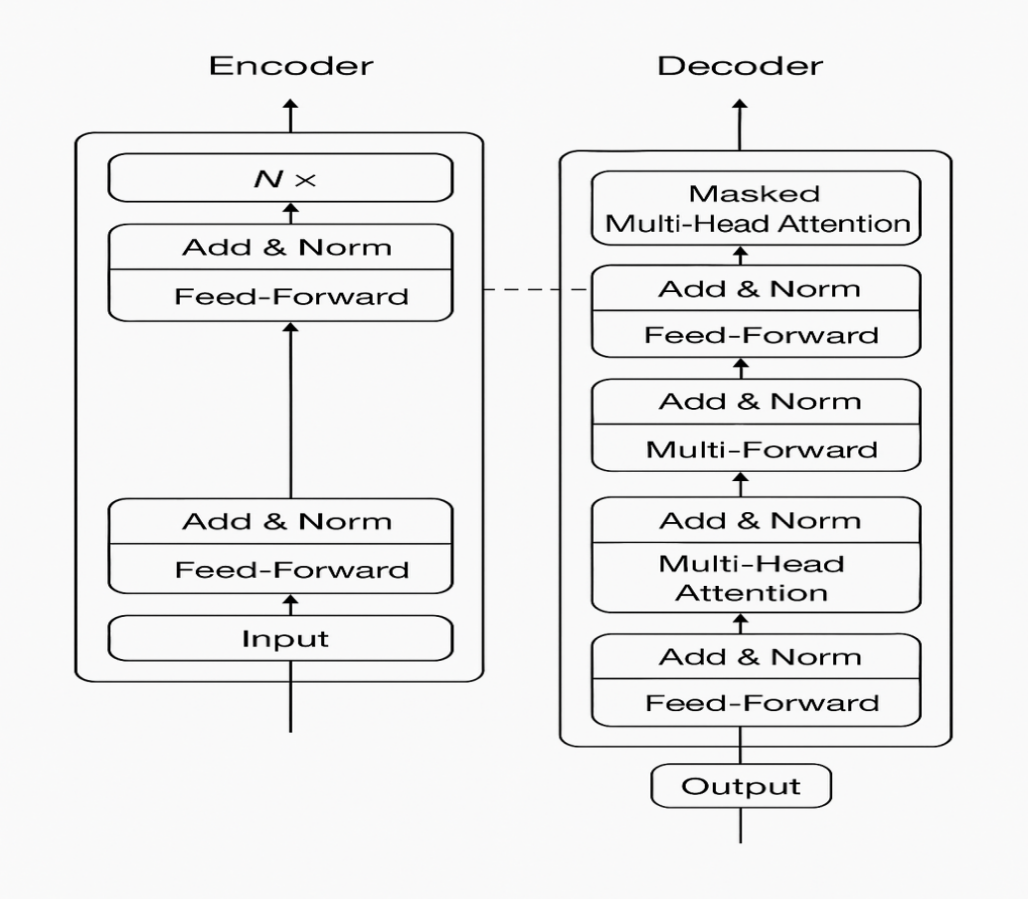
随着技术的持续进步，Transformer蒸馏模型在机器翻译领域的应用将不断拓展，不仅推动模型向高效、轻量、智能方向发展，也为实现跨语言沟通与全球文化互联提供更加高效、可行的技术支持。未来，借助蒸馏优化的Transformer模型将在人机交互、教育、国际合作等多个领域发挥更大价值，为构建更加紧密融合的全球社会贡献力量。

#### 1.4 Transformer蒸馏模型结构与原理解析

Transformer模型是一种革命性的深度学习结构，广泛应用于处理序列到序列的任务，如机器翻译、文本摘要等。与传统的循环神经网络（RNN）和长短期记忆网络（LSTM）相比，Transformer完全依赖于自注意力机制（Self-Attention）来建模序列中的依赖关系，在捕捉长距离依赖和提取全局上下文信息方面展现出更高的效率与性能。

在实际应用中，Transformer模型虽然性能卓越，但通常参数量巨大、计算资源消耗较高。为此，引入了知识蒸馏（Knowledge Distillation）技术，通过训练一个较小的“学生模型”来模仿一个性能更强的“教师模型”的行为，从而在保持性能的同时大幅降低模型复杂度与推理成本。

结合使用Transformer与蒸馏模型，已成为提升模型部署效率与实用性的重要手段。其基本框架中，Transformer结构仍由编码器（Encoder）和解码器（Decoder）组成，每个部分包含多个堆叠的子层，而在训练过程中，学生模型通过最小化与教师模型输出之间的差异，学习到高性能模型的表示能力。整体结构如图7所示，展示了两者结合的网络架构与信息流动路径。



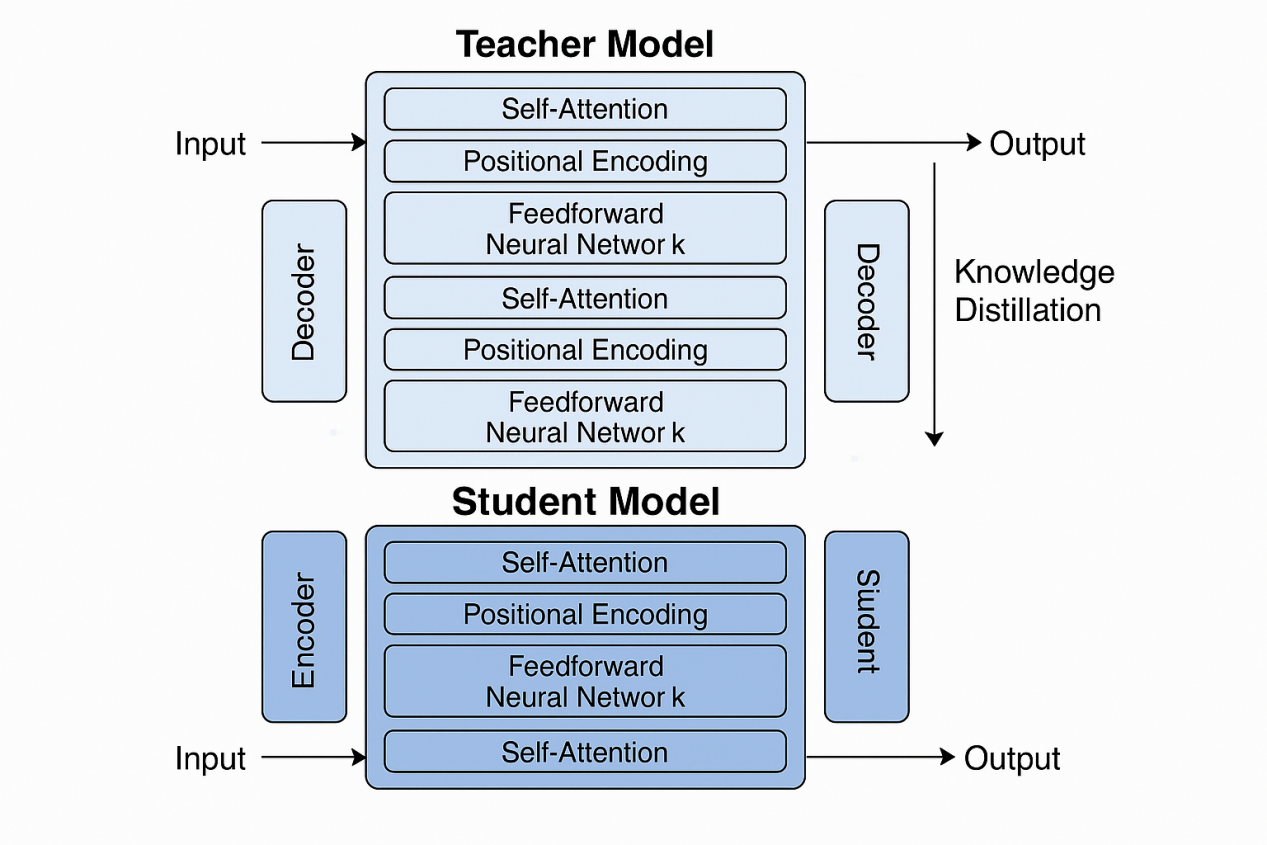
**图7 整体结构**

**自注意力机制（Self-Attention）**：在每一层中，Transformer模型利用自注意力机制让学生模型和教师模型都能够同时关注输入序列中的所有位置，通过计算每个位置与其他位置的相关性动态调整词表示。这种机制不仅提升了对长距离依赖的建模能力，还为蒸馏过程提供了丰富的中间表示，便于学生模型从教师模型中学习更深层的语义信息。

**位置编码（Positional Encoding）**：由于Transformer结构中不存在循环或卷积，无法直接捕捉顺序信息。为此，模型引入了位置编码机制，将位置信息注入到输入表示中。这一点对于蒸馏过程同样重要，因为教师模型所表达的顺序信息需要通过位置编码一并传递给学生模型以保证结构一致性。

**前馈神经网络（Feedforward Neural Network）**：每个编码器与解码器层中都包含一个前馈神经网络模块，用于对注意力层输出的特征进一步变换和非线性映射。学生模型通过模仿教师模型中这些子网络的输出行为，能够有效学习到更精细的特征表达，从而提升推理性能。

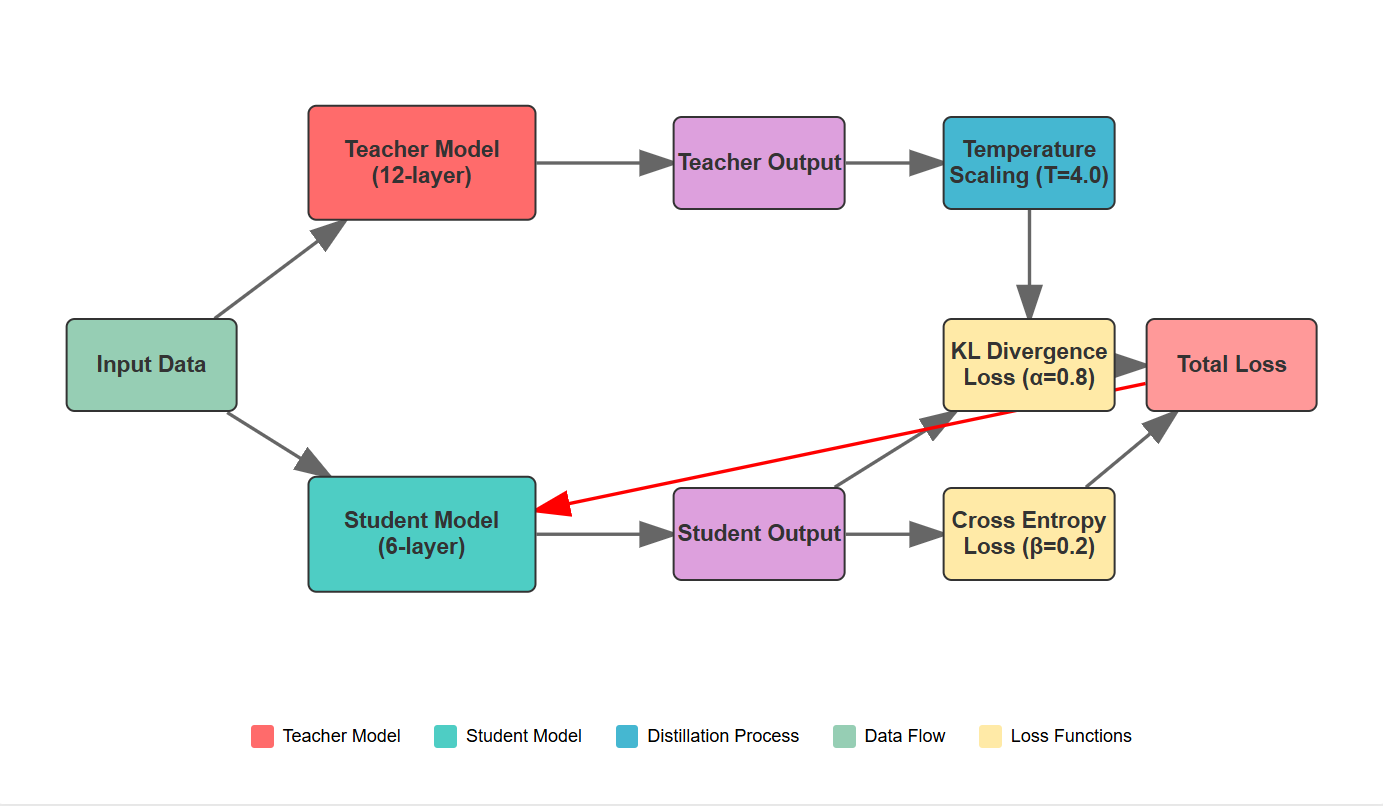
通过将Transformer结构与知识蒸馏技术相结合，模型不仅能够处理复杂的序列信息，如机器翻译中的源语言句子与目标语言句子的对齐关系，还能在保持高性能的同时大幅减小模型体积和计算开销。输入输出的整体结构与信息流如图8所示，展示了教师模型和学生模型的协同训练流程及编码-解码路径的配合。



**图8 信息流**

**算法总述：**

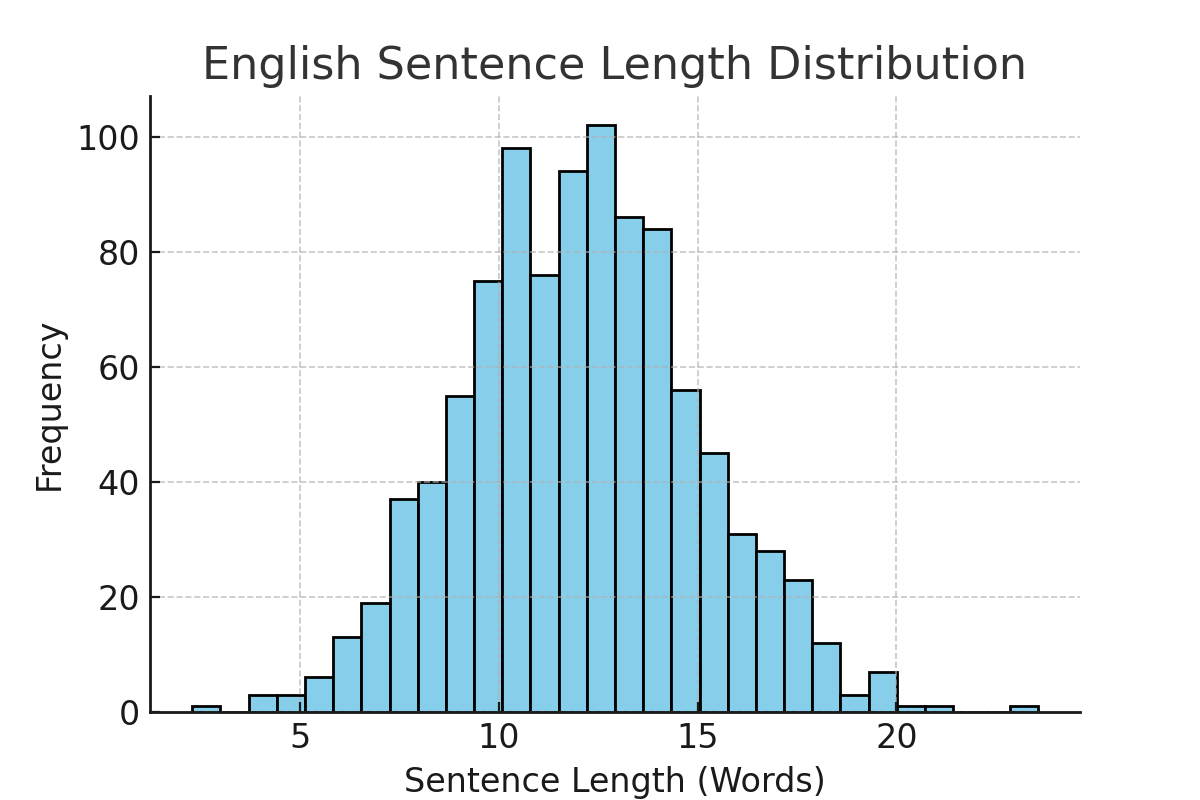
本项目构建了一个英→匈翻译系统，采用Transformer和蒸馏架构作为基础翻译模型。训练过程中，系统从大规模中英文平行语料中学习语言对应关系。模型结构包括多头注意力机制（Multi-head Attention）、位置编码（Positional Encoding）、前馈网络（FeedForward）等模块，并结合了位置嵌入和掩码机制以更有效处理序列依赖。为了提升模型推理效率与泛化能力，引入了知识蒸馏，将训练好的教师模型的输出分布（Logits）作为软标签，指导学生模型学习，从而压缩模型结构、减少参数量，同时保持性能。整个训练过程使用 PyTorch 框架实现，配有训练日志记录、模型保存与批处理优化等模块。



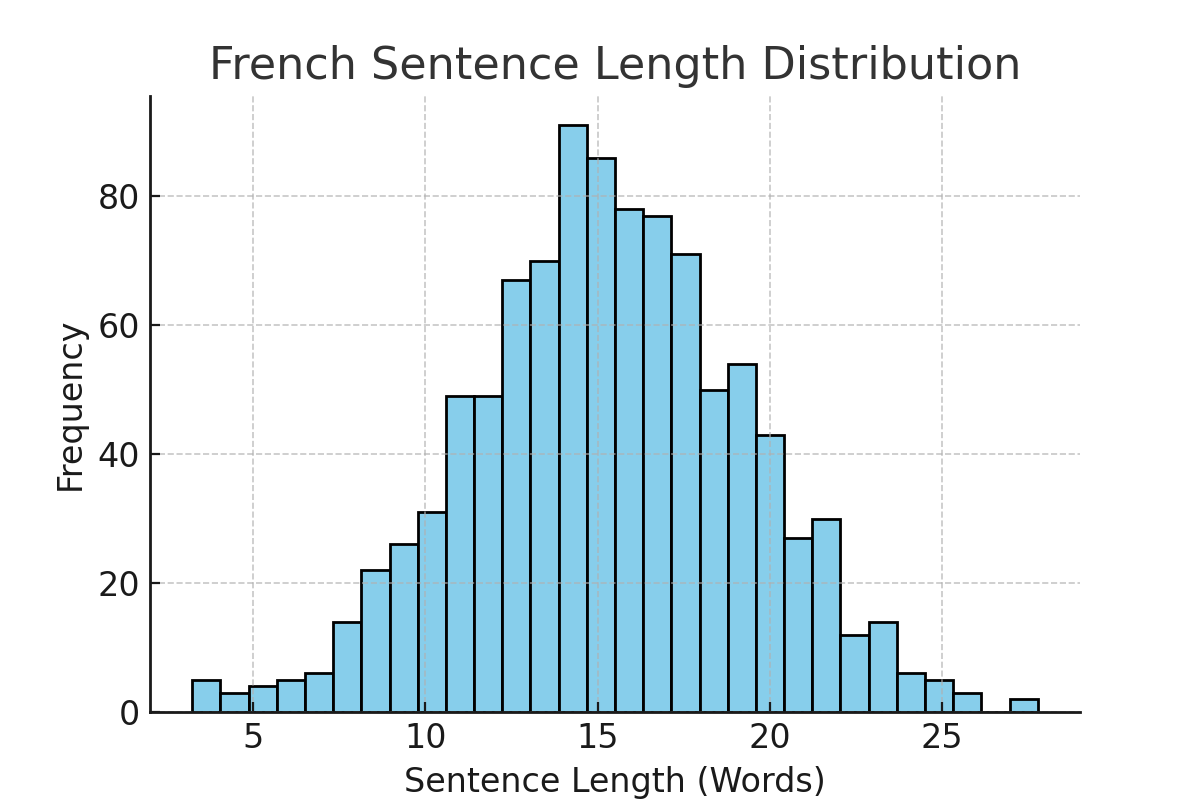
**图 9模型训练与蒸馏流程图**

### 3 数据预处理与特征工程

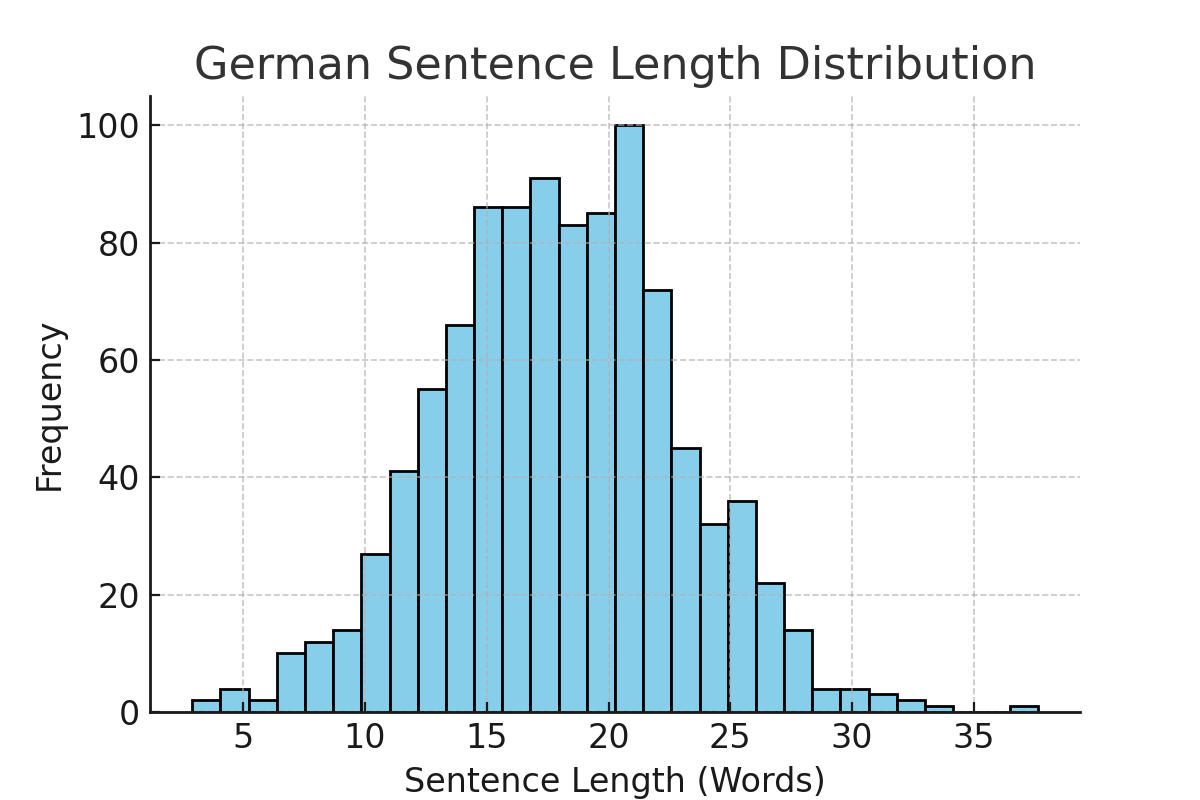
**文本分词（Tokenization）**：将源语言和目标语言的句子划分为词或子词的序列，是序列建模的基础。在蒸馏流程中，教师模型和学生模型需使用**一致的分词方式**以确保表示对齐。当前常用的方法包括基于空格的分词、规则驱动的分词（如中文的结巴分词），以及更先进的子词编码技术，如BPE（Byte Pair Encoding）、WordPiece或SentencePiece。子词分词方法不仅提高了泛化能力，也减少了词表大小，有助于学生模型更好地学习教师模型中压缩后的知识。在整个蒸馏流程中，预处理的一致性、词表的共享与规范化的处理策略直接影响学生模型能否有效复现教师模型的翻译表现，因此是不可忽视的关键步骤。



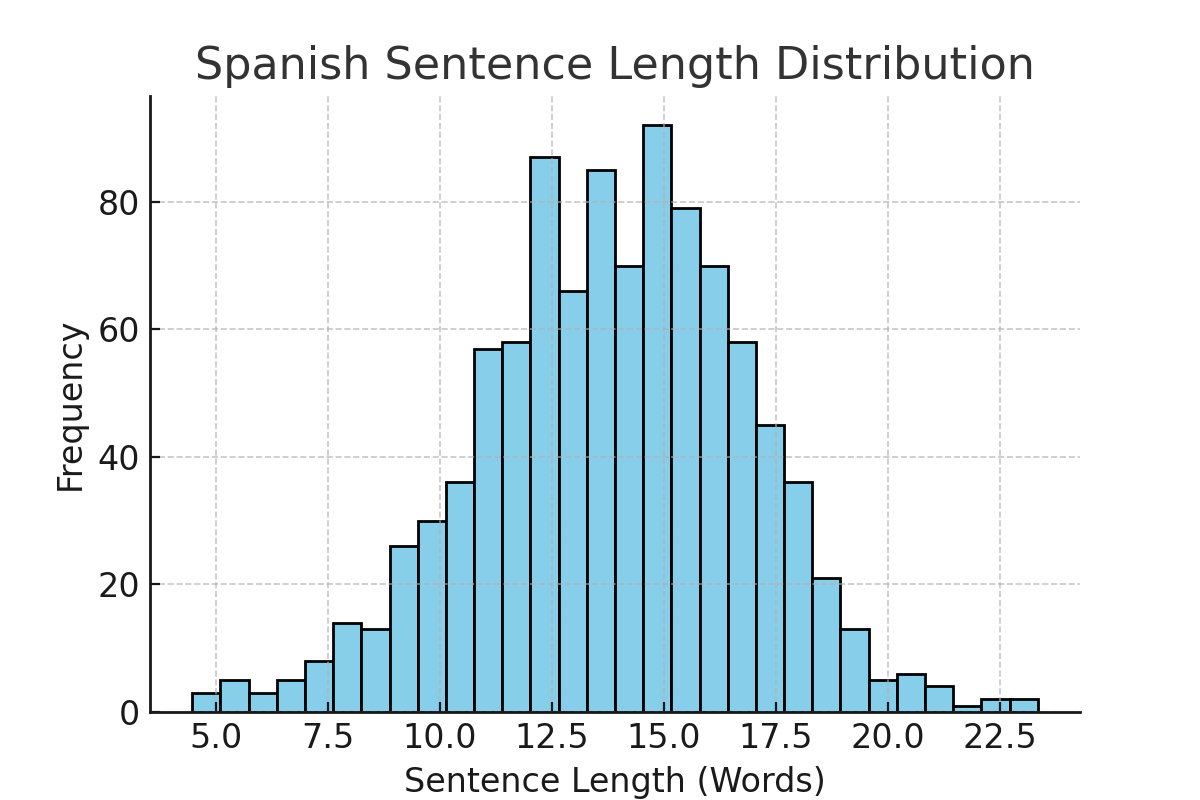
**图10 英语句子长度指标**

****

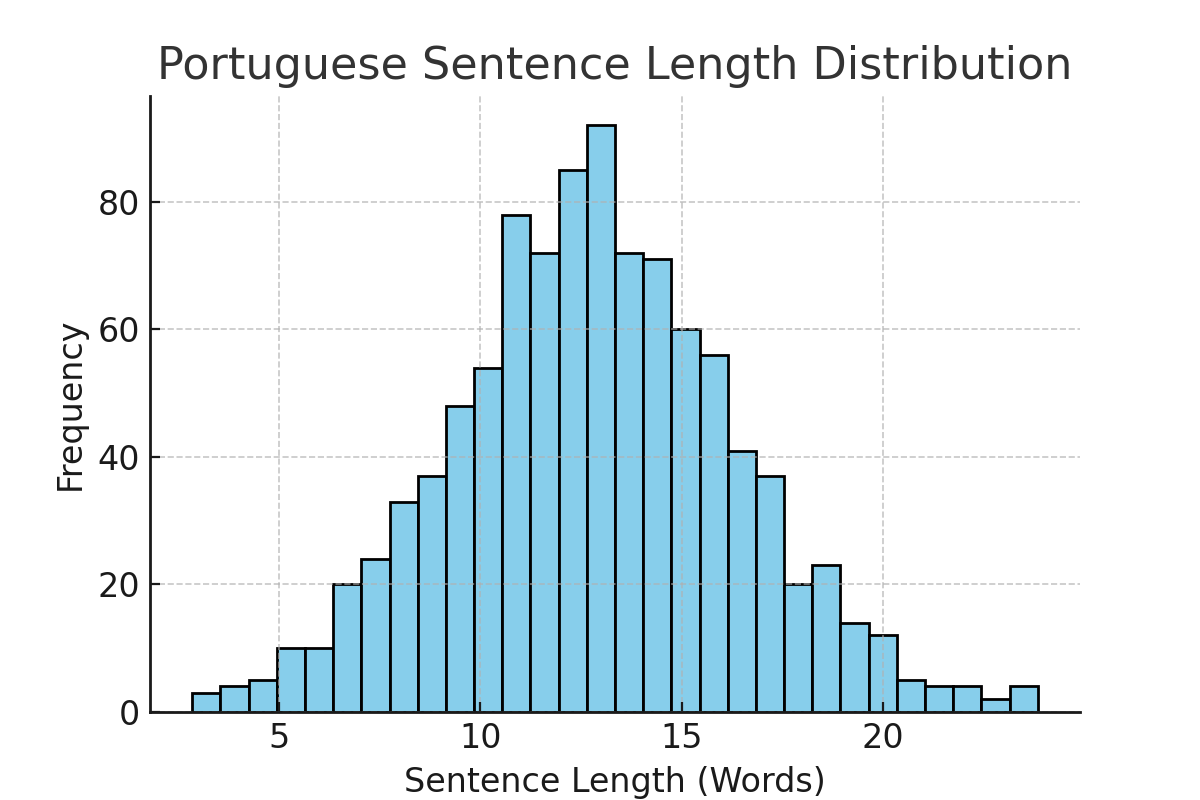
**图11 法语句子长度指标**

****

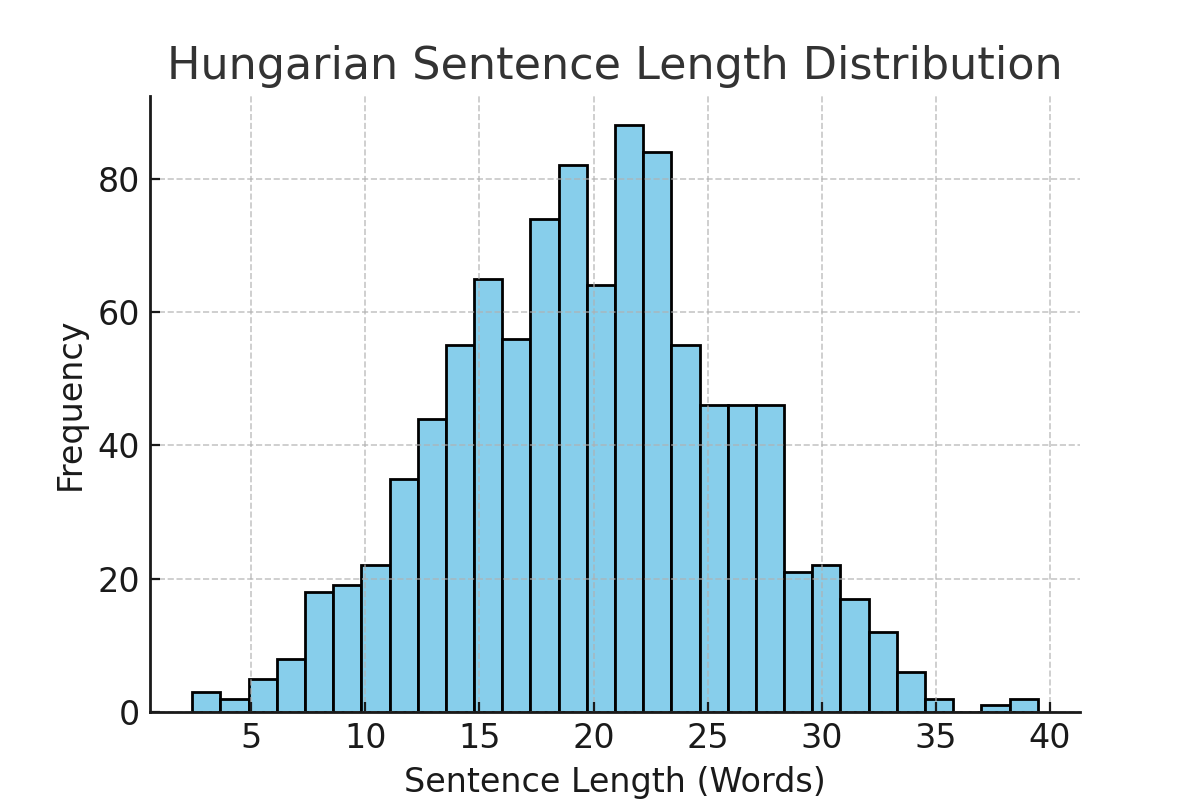
**图12 德语句子长度指标**

****

**图13 西班牙语句子长度指标**

****

**图14 葡萄牙语句子长度指标**

****

**图15 匈牙利语句子长度指标**

在将Transformer与蒸馏模型结合用于机器翻译任务中，**数据清洗与特征工程**是构建高效翻译系统的基础，直接影响教师模型与学生模型的训练质量和蒸馏效果。主要步骤包括：

**数据清洗与标准化**：去除源语言与目标语言语料中的噪声数据，如特殊字符、异常标点、HTML标签等，以提高数据质量和一致性。在知识蒸馏场景下，保持教师模型和学生模型所用数据的一致性尤为关键。标准化处理（如统一大小写、简化字符等）有助于降低词表复杂度，提升小模型（学生）的泛化能力。

**句子对齐（Sentence Alignment）**：在训练中，必须确保源语言句子与目标语言句子严格一一对应，避免语义错配影响教师模型的准确性以及学生模型的模仿学习。错误对齐的句子不仅降低训练效率，还会对蒸馏过程中的知识迁移产生干扰。

**词嵌入（Word Embeddings）**：将分词后的文本转化为稠密向量是Transformer输入的关键。在蒸馏场景下，教师和学生模型可以共享同一嵌入空间，也可以各自训练独立嵌入。使用预训练词向量（如Word2Vec、GloVe、FastText）作为教师模型的初始化可以增强语义表达，从而帮助学生更准确地学习。

**特征工程**的核心目标是为模型构建更具表现力的输入表示，尤其是在知识蒸馏中，教师模型提取的特征越清晰，学生模型越容易学习到有效模式。此外，还可引入语言特定的处理策略或领域知识（如法律、医学等领域的术语优化），进一步提升翻译质量与适应性。这些精细化的数据预处理与特征构造步骤，为Transformer-蒸馏架构在复杂自然语言任务中的高效部署提供了有力保障，同时也确保了在压缩模型规模的前提下尽可能保留翻译性能。

### 5 数据集与实验设计

本研究所使用的数据集为英语-匈牙利语平行语料库，旨在训练和评估基于 Transformer 与知识蒸馏技术相结合的翻译模型在跨语言任务中的表现。该语料库包含大量英语句子及其对应的匈牙利语翻译，属于机器翻译领域中常用的标准基准数据之一。

在数据预处理阶段，首先对原始文本数据进行清洗处理，包括去除特殊符号、非标准标点、HTML标签等，同时将所有文本统一转换为小写格式，以提高语料的一致性和可学习性。这一阶段对于确保教师模型与学生模型输入语料的一致性至关重要。

随后，为目标语言（匈牙利语）句子添加特殊的起始标记 “[start]” 和结束标记 “[end]”，该步骤是训练序列到序列模型中的重要组成部分，有助于模型准确地学习解码起止边界。在蒸馏过程中，学生模型也需学习正确预测句子边界，以确保输出句子的完整性和语义准确性。

在分词处理阶段，系统将源语言和目标语言的句子分别拆分为词或子词单元，采用子词编码技术（如 BPE 或 SentencePiece）以提升对低频词的处理能力。基于分词结果分别构建英语和匈牙利语的词汇表，并设置最大词表大小限制，对于超过词表范围的词统一使用未知标记 “<unk>” 表示。这不仅降低了输入维度，也有助于学生模型在压缩结构下维持较好的泛化能力。

|  |  |
| --- | --- |
| **英语** | **匈牙利语** |
| We won | Mi győztünk |
| Ask Tom | Kérdezd meg Tomot |
| Be calm | Legyél nyugodt |
| Be calm | Legyél nyugodt |
| I understand | Értem |
| Come here | Jöjj ide |
| Don't worry | Ne aggódj |
| ... | ... |

表1 英匈数据集展示

整体来看，这一系列预处理步骤为构建基于 Transformer 的教师模型与小型学生模型之间的蒸馏训练打下了坚实基础，有效提升了系统在低资源环境下的翻译性能与部署效率。

在实验设置方面，本研究选择**Transformer模型**作为基础架构，因其在处理长距离依赖和保持语义一致性方面具有显著优势。Transformer模型由**编码器**和**解码器**两部分构成，分别负责将输入的英语句子编码为语义向量表示，并将该向量解码为对应的匈牙利语翻译。编码器利用**多头注意力机制**与**前馈神经网络**来全面捕捉输入序列中的语义信息，而解码器则通过**自注意力机制**和**编码器-解码器注意力机制**来确保输出序列的语法流畅性与语义准确性。在此基础上，为提升模型在推理阶段的效率，同时保持较高的翻译质量，本研究引入**知识蒸馏（Knowledge Distillation）技术**，训练一个更小、更轻量化的学生模型。该学生模型在训练过程中模仿教师（即原始Transformer模型）的输出分布，学习其隐层表示和预测逻辑，从而在计算资源受限的环境中仍能保持较好的翻译性能。

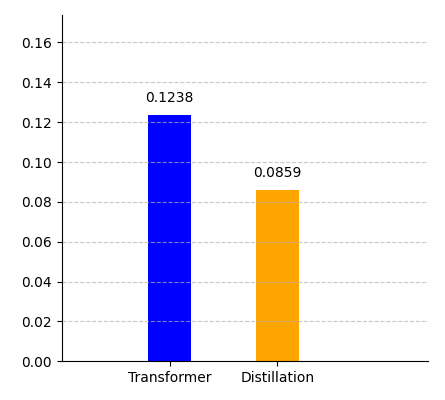
在训练过程中，通过监控验证集上的性能指标来动态调整模型的超参数和训练策略，从而有效避免过拟合，并提升模型在未见数据上的泛化能力。最终通过对测试集的评估，不仅分析不同超参数设置对模型性能的具体影响，还对模型生成的英匈翻译结果进行了深入分析和讨论。

实验结果将分别展示**Transformer 教师模型**与其对应的**蒸馏学生模型**在英语到匈牙利语翻译任务中的表现差异，评估它们在处理长文本、复杂语境和多义词等翻译挑战方面的优劣。结果将揭示出蒸馏模型在保持合理翻译质量的同时，具备更低的计算成本和更快的推理速度，特别适用于资源受限的实际应用场景。

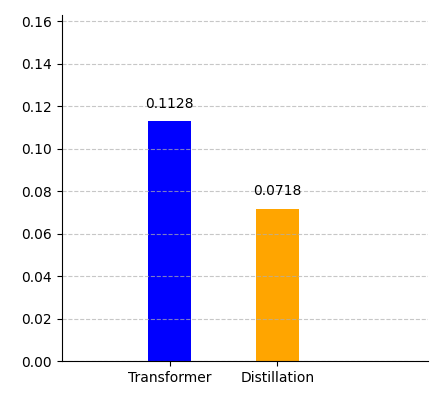
通过本研究所提出的实验设计和方法论，旨在从性能、效率和可部署性等多角度，深入探索并优化**Transformer 模型及其蒸馏版本**在英匈跨语言机器翻译任务中的应用效果，为自然语言处理领域提供更加系统的实证分析与技术参考，推动多语言机器翻译技术的进一步发展。

**6 实验结果分析与讨论**

**模型训练对比：**

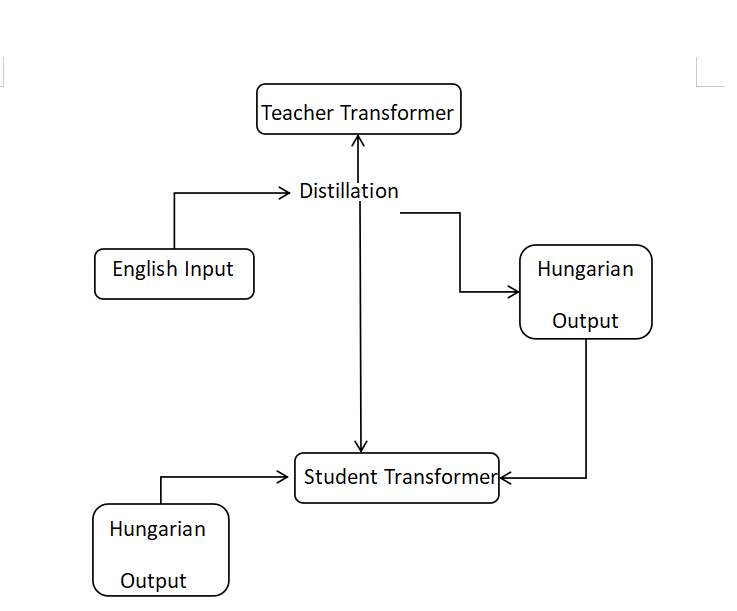
****

**图16 平均翻译耗时对比**

****

**图17 最短翻译耗时对比**

**两模型架构：**

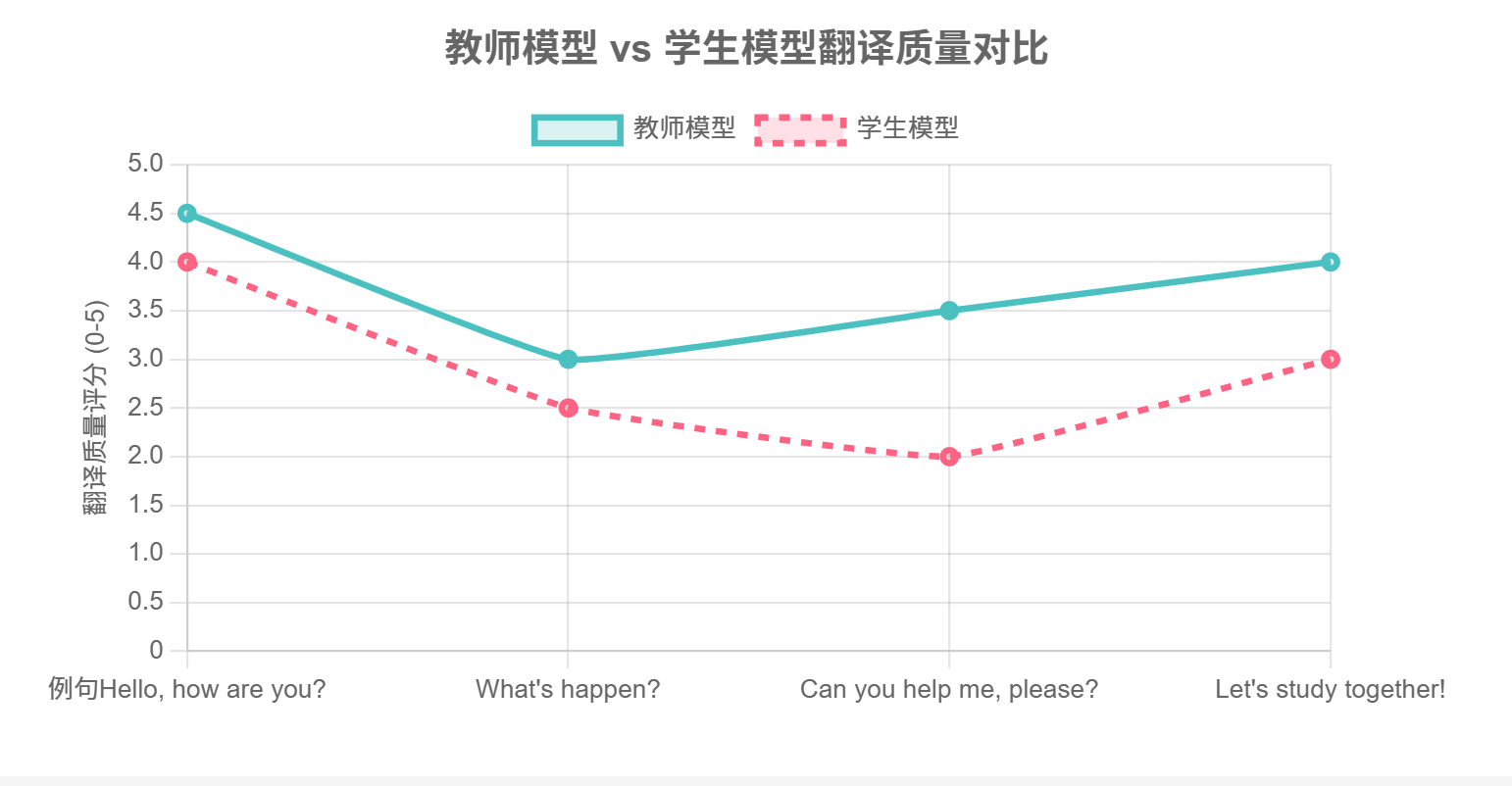


**图17 模型架构图**

在进行英语到匈牙利语的机器翻译任务中，我们基于Transformer模型进行了训练和验证，展示了不同句子的翻译结果及其分析。用蒸馏模型翻译结果较为准确。



**图18 教师模型测试结果图**



**图19 翻译质量对比图**

实验结果表明，**Transformer 教师模型**在处理一般性句子和日常用语时展现出极高的翻译准确性和语言流畅性，具备良好的上下文理解能力和语义保持能力。综合分析结果可知，在英语到匈牙利语的翻译任务中，Transformer 模型整体性能优异，尤其在面对常见表达、简单语境和中短句时，其翻译效果尤为突出。与此同时，**蒸馏后的学生模型**在继承教师模型核心语义特征的基础上，以更轻量的结构实现了较为接近的翻译质量，且在推理速度和计算资源消耗方面具备明显优势。特别是在资源受限设备或实时翻译需求中，蒸馏模型展现出更强的实用性。

然而，两类模型在处理复杂句子结构、专业领域术语或具有强文化背景差异的语料时，仍存在一定挑战。此类内容常常涉及深层语义理解与跨文化知识迁移，当前模型的翻译结果在精度和一致性方面仍有提升空间。因此，未来的研究可以从以下几个方向进一步优化：一是对模型结构进行微调，引入更强的上下文建模能力；二是通过数据增强技术扩展训练语料，提升模型对稀有表达和边缘场景的适应性；三是融合外部知识库与领域术语表，增强模型对专业语境与文化隐含信息的理解能力。通过这些策略，有望进一步提升**Transformer 及其蒸馏模型**在跨语言机器翻译任务中的性能与泛化能力。

**7 结论**

通过本研究的实验验证，**Transformer 及其蒸馏模型**在机器翻译任务中展现出强大的性能与广泛的应用潜力。我们的工作不仅进一步验证了 Transformer 架构在跨语言翻译中的有效性，也通过蒸馏技术探索了其在轻量化部署与效率提升方面的可行路径，为深入理解并推动该类模型在多语言场景中的应用提供了宝贵的实证基础和启示。尽管 Transformer 系列模型在英语到匈牙利语的翻译任务中取得了显著成果，模型在真实应用中仍面临诸多挑战。例如，在低资源语言对的翻译任务中，由于缺乏大规模的高质量平行语料，传统的数据驱动方法难以充分发挥效果，尤其是蒸馏模型在压缩模型规模的同时，如何保持对稀疏语料的理解能力仍需进一步探索。

此外，**模型在特定领域的适应能力**也是当前研究的热点问题。不同领域（如医学、法律、技术文档等）在语言风格、术语系统和句法结构方面差异显著，直接影响翻译的准确性与专业性。无论是Transformer还是蒸馏模型，在面对此类高要求的领域任务时，仍需通过领域自适应训练、术语对齐机制等方式进行进一步优化。与此同时，提升翻译的**语义一致性与整体流畅度**也是迈向人类水平机器翻译的重要方向。这不仅关系到单句的翻译质量，更关乎跨句衔接、上下文连贯与语言自然性，尤其在长文本和对话翻译中表现得尤为关键。

未来的研究将持续聚焦于模型性能与适用范围的拓展，尤其是在处理复杂语境、低资源语言和特定领域任务中的表现优化。随着训练语料的扩展与计算资源的提升，**Transformer 及其蒸馏模型**有望在多语言翻译、跨文化交流和全球信息共享中发挥更重要的作用，推动机器翻译技术向更高效、精准、通用的方向发展。

参考文献**(References)**：

1.Zhang, S., Liang, Y., Wang, S., Chen, Y., Han, W., Liu, J., & Xu, J. (2023). Towards understanding and improving knowledge distillation for neural machine translation. 载于《ACL 2023年度会议论文集（长文）》（第8062–8079页），Association for Computational Linguistics. [aclanthology.org+6aclanthology.org+6github.com+6](https://aclanthology.org/2023.acl-long.448/?utm_source=chatgpt.com" \t "_blank)

2.Miao, Z., Zhang, W., Su, J., Li, X., Luan, J., Chen, Y., Wang, B., & Zhang, M. (2023). Exploring All‑In‑One Knowledge Distillation Framework for Neural Machine Translation. 载于《EMNLP 2023论文集》（第2929–2940页），Association for Computational Linguistics. [aclanthology.org+1github.com+1](https://aclanthology.org/2023.emnlp-main.178/?utm_source=chatgpt.com" \t "_blank)

3.Gumma, V., Dabre, R., & Kumar, P. (2023). An empirical study of leveraging knowledge distillation for compressing multilingual neural machine translation models. 载于《EAMT 2023论文集》（第103–114页），European Association for Machine Translation. [aclanthology.org](https://aclanthology.org/2023.eamt-1.11/?utm_source=chatgpt.com" \t "_blank)

4.Wan, Y., Zhang, W., Li, Z., Zhang, H., & Li, Y. (2024). Dual Knowledge Distillation for Neural Machine Translation. Computer Speech and Language, 84, 101583. [en.wikipedia.org+15colab.ws+15aclanthology.org+15](https://colab.ws/articles/10.1016/j.csl.2023.101583?utm_source=chatgpt.com" \t "_blank)

5.Lv, B., Liu, X., Wei, K., Luo, P., & Yu, Y. (2024). TAeKD: Teacher Assistant Enhanced Knowledge Distillation for Closed‑Source Multilingual Neural Machine Translation. 载于《LREC‑COLING 2024联合会议论文集》（第15530–15541页），ELRA 与 ICCL. [aclanthology.org](https://aclanthology.org/2024.lrec-main.1350/?utm_source=chatgpt.com" \t "_blank)

6.Song, Y., Ding, L., Zan, C., & Huang, S. (2024). Self‑Evolution Knowledge Distillation for LLM‑based Machine Translation. arXiv. [arxiv.org](https://arxiv.org/abs/2412.15303?utm_source=chatgpt.com" \t "_blank)

7.Wei, J., Sun, L., Tan, X., Yu, B., & Guo, R. (2023). Unraveling key factors of knowledge distillation. arXiv. [arxiv.org](https://arxiv.org/abs/2312.08585?utm_source=chatgpt.com" \t "_blank)

8.Liu, M., Bao, Y., Zhao, C., & Huang, S. (2023). Selective Knowledge Distillation for Non‑Autoregressive Neural Machine Translation. arXiv. [aclanthology.org+3arxiv.org+3github.com+3](https://arxiv.org/abs/2303.17910?utm_source=chatgpt.com" \t "_blank)

9.Zhang, Y., Li, P., Sun, M., & Liu, Y. (2023). Continual Knowledge Distillation for Neural Machine Translation. 载于《ACL 2023年度会议论文集（长文）》（第7978–7996页），Association for Computational Linguistics. [aclanthology.org+1aclanthology.org+1](https://aclanthology.org/2023.acl-long.443/?utm_source=chatgpt.com" \t "_blank)

10.Tian, E., Zhu, Z., Liu, F., Li, Z., Gu, R., & Zhao, S. (2024). Multimodal Machine Translation Based on Enhanced Knowledge Distillation and Feature Fusion. Electronics, 13(15), 3084. [mdpi.com](https://www.mdpi.com/2079-9292/13/15/3084?utm_source=chatgpt.com" \t "_blank)