摘要

最近基于Transformer的上下文词表示，包括BERT和XLNet，在自然语言处理的多个领域中展现出最先进的性能。在特定任务的数据集上微调这些经过训练的上下文模型一直是实现卓越性能的关键。虽然对于词汇应用（仅涉及语言模态的应用），微调这些预训练模型是直接的，但对于多模态语言（一门关注面对面交流建模的不断发展的领域）却并非易事。预训练模型不具备接受视觉和声学两种额外模态所需的组件。在本文中，我们提出了对BERT和XLNet的扩展，称为Multimodal Adaptation Gate（MAG）。MAG允许BERT和XLNet在微调过程中接受多模态非语言数据。它通过生成对BERT和XLNet的内部表示进行转换来实现，该转换是基于视觉和声学模态进行条件约束的。在我们的实验中，我们研究了常用的CMU-MOSI和CMU-MOSEI数据集，用于多模态情感分析。通过对MAG-BERT和MAG-XLNet进行微调，显著提高了情感分析性能，超过了以前的基准线以及仅针对语言的BERT和XLNet的微调。在CMU-MOSI数据集上，MAG-XLNet首次在自然语言处理社区实现了人类级别的多模态情感分析性能。

1 引言

人类面对面的交流是语言、声学和视觉模态的无缝融合。在日常互动中，我们共同利用所有这些模态来传达我们的意图和情感。理解这种面对面交流属于一个日益增长的自然语言处理研究领域，称为多模态语言分析（Zadeh等人，2018b）。这个领域面临的最大挑战是有效地将交流的三个支柱进行建模。这使得人工智能系统能够理解多感官信息，而不忽视非语言因素。在对话系统和虚拟现实等许多应用中，这种能力对于保持用户互动的高质量至关重要。

近期在自然语言处理领域中，上下文词表示的成功主要归功于新的基于Transformer的模型，如BERT（Devlin等，2018）和XLNet（Yang等，2019）。这些基于Transformer的模型已经显示出在下游任务上的性能提升（Devlin等，2018）。然而，它们真正的下游潜力来自于对其预训练模型进行特定任务的微调（Devlin等，2018）。对于仅涉及语言模态的词汇数据集，这通常很容易实现。然而，对于多模态语言的微调既不是简单的，也没有被研究过，因为BERT和XLNet只接受语言输入。因此，在将BERT和XLNet应用于多模态语言时，人们必须要么（a）放弃非语言信息，对语言进行微调，要么（b）仅提取词表示，并继续使用先进的多模态模型进行研究。

在本文中，我们提出了一种成功的框架，用于对BERT和XLNet进行多模态输入的微调。我们的框架使得BERT和XLNet的核心结构保持完整，并且仅将一个精心设计的多模态适应门（MAG）附加到模型上。使用基于非语言行为的注意力机制，MAG将信息丰富的视觉和声学因素映射到具有轨迹和幅度的向量上。在微调过程中，这个适应向量修改BERT和XLNet的内部状态，使得模型能够无缝地适应多模态输入。在我们的实验中，我们使用了CMU-MOSI（Zadeh等，2016）和CMU-MOSEI（Zadeh等，2018d）这两个多模态语言数据集，重点研究了多模态情感分析这一核心自然语言处理任务。我们比较了MAG-BERT和MAG-XLNet在分类和回归情感分析中与上述（a）和（b）情景的性能。我们的研究结果表明，使用MAG对这些先进的预训练Transformer进行微调可以持续改善性能，即使BERT和XLNet从未在多模态数据上进行过训练。

因此，本文的贡献总结如下:

- 我们提出了一种高效的框架，用于对多模态语言数据进行BERT和XLNet的微调。该框架使用了一个名为Multimodal Adaptation Gate（MAG）的组件，对模型的开销很小。

- 与情景（a）和（b）相比，MAG-BERT和MAG-XLNet在CMU-MOSI和CMU-MOSEI数据集上取得了最新的最先进性能。对于CMU-MOSI数据集，MAG-XLNet的性能与已报告的人类表现相当。

2 相关工作

本文的研究与以下研究领域相关：

2.1 多模态语言分析

多模态语言分析是自然语言处理中的一个新兴研究趋势（Zadeh等，2018b），它帮助我们从文本、视觉和声学等多个模态理解语言。这些分析特别关注情感分析（Poria等，2018）、情感识别（Zadeh等，2018d）和人格特征识别（Park等，2014）等任务。这一领域的研究常常关注于新颖的多模态神经架构（Pham等，2019；Hazarika等，2018）和多模态融合方法（Liang等，2018；Tsai等，2018）。与本文内容相关的是，我们讨论了该领域中的一些模型，包括TFN、MARN、MFN、RMFN和MulT。
张量融合网络（TFN）（Zadeh等，2017）创建一个多维张量，明确捕捉三种模态之间的所有可能的交互：单模态、双模态和三模态。
多注意力循环网络（MARN）（Zadeh等，2018c）使用三个独立的混合LSTM记忆体，具有传播跨模态交互的能力。
记忆融合网络（MFN）（Zadeh等，2018a）通过多视图门控内存将来自三个独立LSTM的信息进行同步。
循环记忆融合网络（RMFN）（Liang等，2018）以多阶段方式捕捉模态之间的微妙交互，使每个阶段能够关注一部分信号。
用于不对齐多模态语言序列的多模态Transformer（MulT）（Tsai等，2019）采用三个Transformer，每个Transformer用于一个模态，以自注意的方式捕捉与其他两个模态的交互。三个Transformer的信息通过后期融合进行聚合。

2.2 预训练语言表示

从大型语料库中学习词表示是自然语言处理领域的一个活跃研究领域（Mikolov等，2013；Pennington等，2014）。Glove（Pennington等，2014）和Word2Vec（Mikolov等，2013）对许多NLP任务的最新技术水平做出了贡献。这些词表示的一个主要缺点是它们是非上下文的。最近，基于大型文本语料库训练的上下文语言表示模型在问答、情感分类、词性标注和相似性建模等多个NLP任务上取得了最新的技术成果（Peters等，2018；Devlin等，2018）。最早的两个值得注意的基于上下文表示的模型是ELMO（Peters等，2018）和GPT（Radford等，2018）。然而，它们只捕捉到单向上下文，因此错过了句子中词之间更微妙的交互。BERT（双向编码器表示来自Transformer）（Devlin等，2018）通过使用Transformer捕捉双向上下文，提供了更好的表示效果，优于ELMO和GPT。XLNet（Dai等，2019）通过构建一个能够捕捉输入的所有可能因子分解的自回归模型，提供了新的上下文表示。对于BERT和XLNet的预训练模型进行微调是实现最新技术水平的关键因素。尽管先前的工作已经探索了使用BERT对多模态数据进行建模（Sun等，2019），但据我们所知，直接对多模态数据进行BERT或XLNet的微调在先前的工作中尚未得到探索。

3 BERT和XLNet

为了更好地理解本文中提出的多模态框架，我们首先概述BERT和XLNet模型。我们首先简要介绍Transformer和Transformer-XL模型中的操作，然后概述BERT和XLNet。

3.1 Transformer

Transformer是一种非递归神经架构，用于建模序列数据（Vaswani等，2017）。Transformer模型的优越性能主要归功于多头自注意力模块。使用该模块，序列的每个元素都通过与其他序列元素的关联进行关注。图2总结了Transformer层（共M个层）的内部操作。通常，Transformer使用编码器-解码器范式。一堆编码器后面跟着一堆解码器，将输入序列映射到输出序列。在输入经过编码器和解码器堆栈之前，还会应用一个带有位置输入嵌入的附加嵌入步骤。

3.2 Transformer-XL

Transformer-XL（Dai等，2019）是Transformer的扩展，它提供了两个改进：a）增强了Transformer捕捉长距离依赖关系的能力（特别是对于上下文碎片化的情况），b）提高了更好地预测前几个符号的能力（这些符号通常对于序列的其余部分至关重要）。它通过一种循环机制将上下文信息从一个段传递到下一个段，并通过相对位置编码机制实现状态复用而不引起时间上的混淆。

3.3 BERT

BERT是一种成功的语言模型，提供了丰富的上下文词表示（Devlin等，2018）。它采用自编码方法，即将输入标记的一部分屏蔽掉，然后根据所有其他非屏蔽标记预测这些标记，从而学习屏蔽标记的向量表示。我们使用用于单句分类任务的BERT变体。首先，通过添加标记嵌入、段嵌入和位置嵌入，从词片段标记序列生成输入嵌入。然后，在这些输入嵌入之上应用多个编码器层。每个编码器都有一个多头注意力层和一个前馈层，每个层后面都有一个残差连接和层归一化。一个特殊的[CLS]标记被添加到输入标记序列的前面。因此，对于长度为N的输入序列，我们从最后一个编码器层得到N+1个向量，其中第一个向量用于在经过仿射变换后预测输入的标签。

3.4 XLNet

XLNet（Yang等，2019）旨在改进BERT模型的两个关键方面：a）屏蔽标记之间的独立性，b）预训练和微调之间在训练和推断中的差异，因为推断输入没有屏蔽标记。XLNet是一种自回归模型，因此不需要屏蔽特定的标记。然而，自回归模型通常捕捉单向上下文（向前或向后）。XLNet可以通过最大化所有可能的因子分解顺序的似然性来学习双向上下文。实质上，它随机采样多个因子分解顺序，并在每个顺序上训练模型。因此，它可以通过考虑所有可能的排列方式（期望中）来建模输入。

XLNet利用了TransformerXL（Dai等，2019）的两个关键思想：相对位置和段重复机制。与BERT类似，它也有一个输入嵌入器，然后是多个编码器。嵌入器将输入标记转换为向量，其中包括标记嵌入、段嵌入和相对位置嵌入信息。每个编码器由一个多头注意力层和一个前馈层组成，每个层后面都有一个残差加法和归一化层。嵌入器的输出被馈送到编码器中，以获得输入的上下文表示。

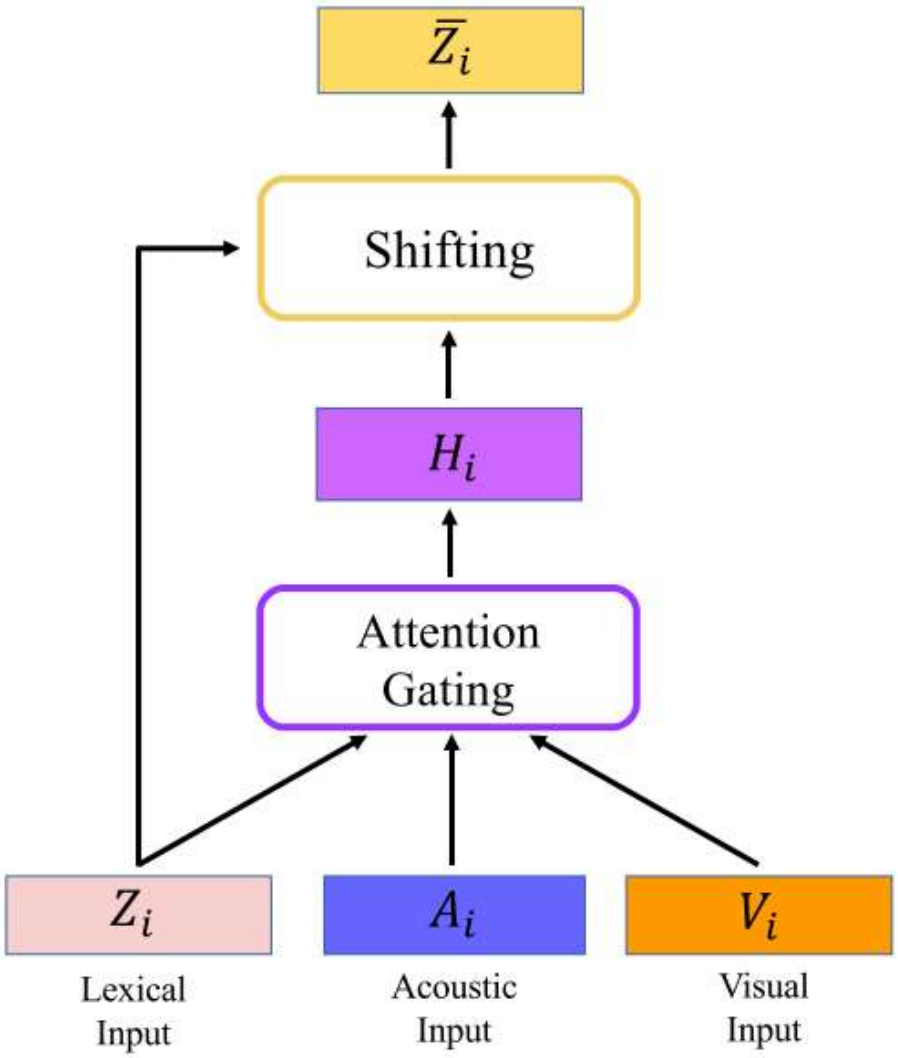

图1：多模态适应门（MAG）以词汇输入向量及其视觉和声音伴随作为输入。随后，对词汇和非语言维度进行的注意力操作用于将多模态数据融合到另一个向量中，然后将其与输入的词汇向量相加（位移）。

4 多模态适应门（MAG）

在多模态语言中，词汇输入伴随着视觉和声音信息，即手势和韵律与语言同时出现。考虑一个语义空间，该空间捕捉了单词的潜在概念（潜在空间中的位置）。在没有多模态伴随的情况下，语义空间直接依赖于语言流形。简单来说，每个单词根据其在语言结构（即句子）中的意义，在此语义空间的某个部分。非语言行为可以影响单词的意义，从而影响单词在语义空间中的位置。语言和非语言伴随行为共同决定了单词在语义空间中的新位置。在本文中，我们将这个新位置视为将仅语言位置与位移向量相加的结果；位移向量具有轨迹和大小，将单词的仅语言位置在非语言行为的影响下转移到新的位置。这就是多模态适应门（MAG）的核心理念。在RAVEN（Wang等，2018）中，研究了这种位移的特别吸引人

的实现方式，其中使用跨模态自注意力计算位移，以突出相关的非语言信息。图1显示了本文研究的MAG。基本上，一个MAG单元接收三个输入，一个纯粹是词汇的，一个是视觉的，最后一个是声音的。用(Z_i, A_i, V_i)表示序列中第i个词的这些输入。我们通过将词汇向量与声音和视觉信息分别连接起来，将这个位移分解为双模态因子$[Z_i; A_i]$和$[Z_i; V_i]$，并使用它们生成两个门控向量g_i^v和g_i^a：

$$g_i^v = R(W_{gv}[Z_i; V_i] + b_v) \quad (1)$$

$$g_i^a = R(W_{ga}[Z_i; A_i] + b_a) \quad (2)$$

其中，W_{gv}和W_{ga}是用于视觉和声音模态的权重矩阵，b_v和b_a是标量偏置。$R(x)$是非线性激活函数。这些门控向量在词汇向量的条件下突出显示视觉和声音模态中的相关信息。
然后，我们通过将A_i和V_i分别乘以它们各自的门控向量，将它们融合在一起，创建一个非语言位移向量H_i：

$$H_i = g_i^a \times (W_a A_i) + g_i^v \times (W_v V_i) + b_H \quad (3)$$

其中，W_a和W_v分别是声音和视觉信息的权重矩阵，b_H是偏置向量。
随后，我们使用Z_i和其非语言位移向量H_i之间的加权求和来创建一个多模态向量$\overline{Z_i} = Z_i + \alpha H_i$：

$$\overline{Z_i} = Z_i + \alpha H_i \quad (4)$$

$$\alpha = min(\frac{||Z_i||_2}{||H_i||_2}\beta, 1) \quad (5)$$

其中，β是通过交叉验证过程选择的超参数。$||Z_i||_2$和$||H_i||_2$分别表示Z_i向量和H_i向量的L_2范数。我们使用缩放因子α，使得非语言位移H_i的影响保持在一个理想的范围内。
最后，我们对$\overline{Z_i}$应用了层归一化和dropout层。

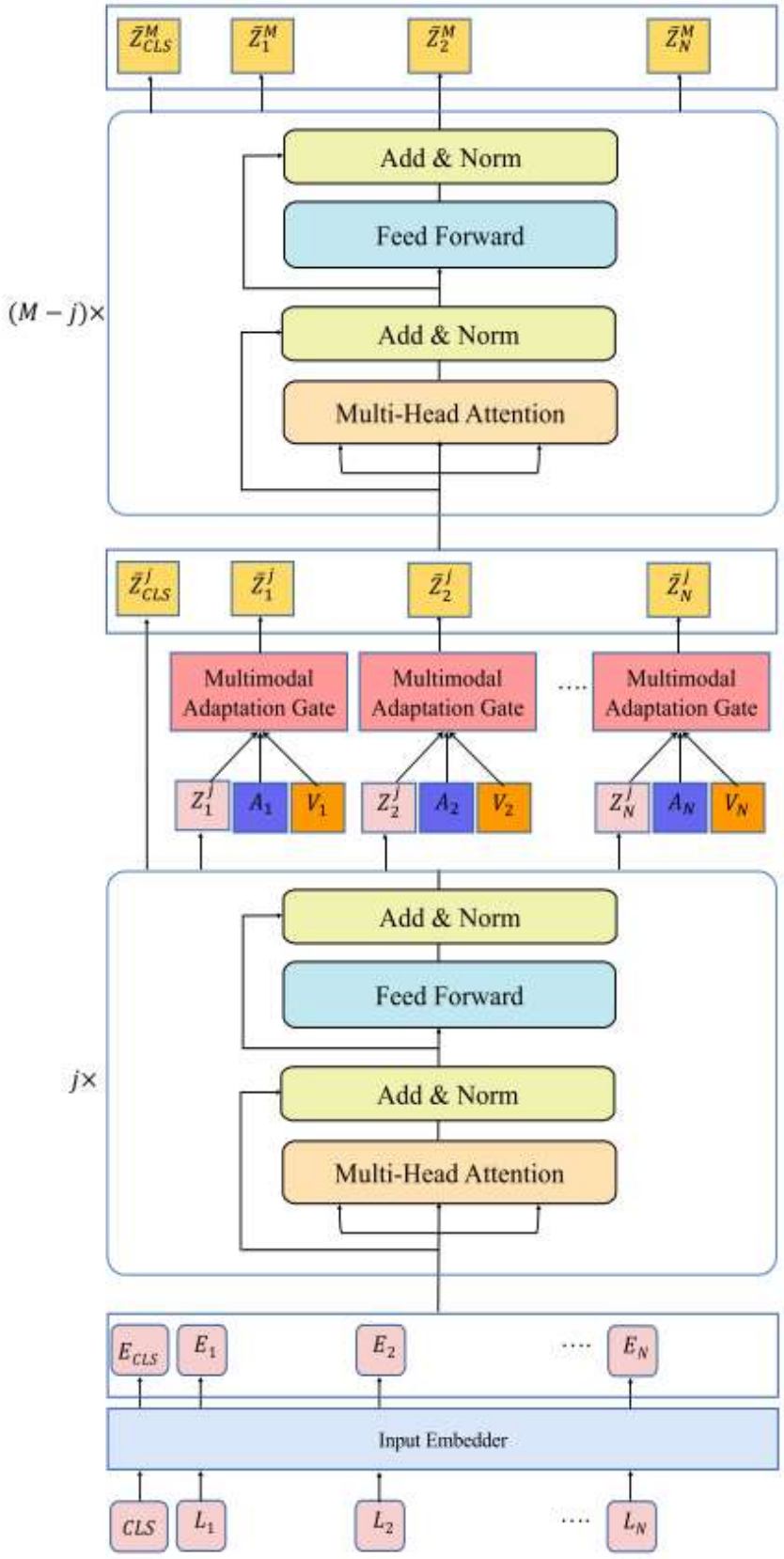

图2：最佳查看方式为放大并以彩色显示。BERT/XLNet的Transformer架构，并在第j层应用了MAG。我们在预训练的Transformer中考虑了总共M层。MAG可以应用于预训练Transformer的不同层。

4.1 MAG-BERT

MAG-BERT是将MAG应用于BERT网络的特定层的组合（图2展示了MAG-BERT和MAG-XLNet的结构）。在每一层中，BERT包含了序列中第i个单词的词汇向量。对于相同的单词，在多模态语言环境中也可以获得非语言伴随信息。MAG本质上是与BERT中的目标层建立联系；这种联系允许多模态信息渗入BERT模型并位移词汇向量。MAG内部的操作使得BERT中的词汇向量能够通过在语义空间中改变其位置来适应多模态信息。除了MAG的连接，对BERT结构没有进行任何改变。

给定一个长度为N的语言序列$L = [L_1, L_2, \dots L_N]$，其中包含单词片段标记，我们在L末尾添加一个$[CLS]$标记，以便之后用于类别标签预测。然后，我们将$L$输入到输入嵌入器，该嵌入器在添加标记、段落和位置嵌入后输出$E = [E_{CLS}, E_1, E_2, \dots E_N]$。然后，我们将$E$输入到第一个编码层，然后连续应用$j$个编码器。经过编码过程后，我们得到输出$Z^j = [Z^j_{CLS}, Z^j_1, Z^j_2, \dots Z^j_N]$，表示经过j层编码后的词汇嵌入。

为了将音频-视觉信息注入这些嵌入中，我们准备了一个三元组序列$[(Z^j_i, A_i, V_i) \colon \forall i \in \{CLS, [1, N]\}]$，将$Z^j_i$与相应的$(A_i, V_i)$配对。每个三元组都经过多模态适应门，将第i个三元组转换为$\overline{Z}^j_i$ - 对应词汇嵌入的统一多模态表示。

由于我们的BERT模型中存在$M = 12$个编码器层，我们将$\overline{Z}^j = [\overline{Z}^j_1, \overline{Z}^j_2, \dots \overline{Z}^j_N]$输入到下一个编码器，并依次应用$M - j$个编码器层。最后，我们从第$M$个编码器层得到$\overline{Z}^M$。作为第一个元素的$\overline{Z}^M_{CLS}$代表$[CLS]$标记，它包含了进行类别标签预测所需的信息。因此，$\overline{Z}^M_{CLS}$经过一个仿射变换，产生一个实数，可以用于预测类别标签。

4.2 MAG-XLNet

与MAG-BERT类似，MAG-XLNet 也具有在任意层使用MAG注入音频-视觉信息的能力。在其任意层的每个位置j处，它保存对应于该位置的词汇向量。利用该位置可用的音频-视觉信息，它可以调用MAG来获得在多模态空间中适当位移的词汇向量。尽管 MAG-XLNet 在很大程度上遵循图2中呈现的一般范例，但它使用了 XLNet 特定的嵌入器和编码器。另一个关键区别是$[CLS]$标记的位置。与BERT不同，$[CLS]$标记附加在输入标记序列的右侧，因此在所有中间表示中，与$[CLS]$对应的向量将是最右侧的一个。按照相同的逻辑，最终编码层的输出将是$\overline{Z}^M = [\overline{Z}^M_1, \overline{Z}^M_2, \dots \overline{Z}^M_N, \overline{Z}^M_{CLS}]$。最后一项$\overline{Z}^M_{CLS}$经过仿射变换后可用于类别标签预测。

5 实验

本节概述了本文中的实验。我们首先描述数据集，然后描述提取的特征、基准线和实验设置。

5.1 CMU-MOSI数据集

CMU-MOSI（CMU多模态意见情感强度）是一个专注于多模态情感分析的多模态语言数据集（Zadeh等，2016）。CMU-MOSI包含来自93个Youtube电影评论视频的2199个视频片段。该数据集中的情感强度注释为实值，范围为$[-3, +3]$。

5.2 计算描述符

对于每个模态，可用以下计算描述符：

语言：我们使用Youtube API进行视频转录，然后进行手动纠正。

声学：使用COVAREP（Degottex等，2014）提取以下相关特征：基频、准开率、归一化幅度比、声门源参数（H1H2、Rd、Rd conf）、VUV、MDQ、前3个共振峰、PSP、HMPDM 0-24和HMPDD 0-12、小波响应的光谱倾斜/斜率（峰值/斜率）、MCEP 0-24。

视觉：对于视觉模态，使用Facet库（iMotions，2017）提取一组视觉特征，包括面部动作单元、面部标记、头部姿势、凝视跟踪和HOG特征。

对于每个单词，我们按照（Chen等，2017）中建立的约定对齐所有三个模态。首先，使用强制对齐（Yuan和Liberman，2008）获取语言和音频之间的单词对齐。然后，每个单词的边界表示同时出现的视觉和声学特征（FACET和COVAREP）。随后，对于每个单词，将同时出现的声学和视觉特征在每个特征上进行平均，从而得到与单词i对应的A_i和V_i向量。

5.3 基准模型

我们将MAG-BERT和MAG-XLNet的性能与各种用于多模态语言分析的最先进模型进行比较。这些模型使用提取的BERT和XLNet词嵌入作为它们的语言输入：

- **TFN（Tensor Fusion Network）**通过创建一个多维张量来显式建模模态内部和模态间的动态（Zadeh等，2017），该张量捕捉了三个模态之间的单模态、双模态和三模态交互。
- **MARN（Multi-attention Recurrent Network）**使用混合LSTM记忆来建模视图间的交互，并使用多注意力块（MAB）来建模跨模态的交互（Zadeh等，2018c）。
- **MFN（Memory Fusion Network）**具有三个独立的LSTM来分别建模每个模态，并使用多视图门控内存来实现它们之间的同步（Zadeh等，2018a）。
- **RMFN（Recurrent Memory Fusion Network）**通过递归多阶段方式捕捉模态内部和模态间的信息（Liang等，2018）。
- **MulT（用于非对齐多模态语言序列的多模态Transformer）**使用三组Transformer，并以后期融合的方式组合它们的输出来建模多模态序列（Tsai等，2019）。我们使用原始模型的对齐变体，该变体在性能上优于非对齐变体。

我们还将我们的模型与仅使用语言模态进行微调的BERT和XLNet进行比较，以衡量MAG框架的成功。

5.4 实验设计

本文中的所有模型都是使用Adam（Kingma和Ba，2014）优化器进行训练，学习率在$\{0.001、0.0001、0.00001\}$之间。我们对每个模型使用了$\{0.1、0.2、0.3、0.4、0.5\}$的dropout进行训练。TFN、MARN、MFN、RMFN、LFN中的LSTM使用了$\{16、32、64、128\}$的潜在大小。对于MulT，我们在网络中使用了$\{3、5、7\}$个层和$\{1、3、5\}$个注意力头。所有模型都使用了CMU-MOSI的指定验证集来寻找最佳超参数。

我们在CMU-MOSI数据集上进行了两个不同的评估任务：i）二元分类和ii）回归。我们将其制定为回归问题，并报告了平均绝对误差（MAE）和模型预测与真实标签的相关性。此外，我们将回归输出转换为分类值，以获得二元分类准确率（BA）和F1分数。除了MAE之外，所有指标的较高值表示更好的性能。对于BA和F1，我们使用了两个评估指标，一个是(Zadeh等人，2018d)中使用的，另一个是(Tsai等人，2019)中使用的。

6 结果和讨论

表格1显示了本文实验的结果。我们从这个表格中总结了以下观察结果：

Task Metric	BA↑	F1↑	MAE↓	Corr↑
Original (glove)				
TFN	73.9/–	73.4/–	0.970/–	0.633/–
MARN	77.1/–	77.0/–	0.968/–	0.625/–
MFN	77.4/–	77.3/–	0.965/–	0.632/–
RMFN	78.4/–	78.0/–	0.922/–	0.681/–
LFN	76.4/–	75.7/–	0.912/–	0.668/–
MulT	–/83.0	–/82.8	–/0.871	–/0.698
BERT				
TFN	74.8/76.0	74.1/75.2	0.955	0.649
MARN	77.7/78.9	77.9/78.2	0.938	0.691
MFN	78.2/79.3	78.1/78.4	0.911	0.699
RMFN	79.6/80.7	78.9/79.1	0.878	0.712
LFN	79.1/80.2	77.3/78.1	0.899	0.701
MulT	81.5/84.1	80.6/83.9	0.861	0.711
BERT	83.5/85.2	83.4/85.2	0.739	0.782
MAG-BERT	**84.2/86.1**	**84.1/86.0**	**0.712**	**0.796**
XLNet				
TFN	78.2/80.1	78.2/78.8	0.914	0.713
MARN	78.3/79.5	78.8/79.6	0.921	0.707
MFN	78.3/79.9	78.4/79.1	0.898	0.713
RMFN	79.1/81.0	78.6/80.0	0.901	0.703
LFN	80.2/82.9	79.1/81.6	0.862	0.701
MulT	81.7/84.4	80.4/83.1	0.849	0.738
XLNet	84.7/86.7	84.6/86.7	0.676	0.812
MAG-XLNet	**85.7/87.9**	**85.6/87.9**	**0.675**	**0.821**
Human	85.7/-	87.5/-	0.710	0.820

表1：在CMU-MOSI数据集上的情感预测结果。最佳结果以粗体显示。MAG-BERT和MAG-XLNet的性能优于基线模型及其仅语言微调的对应模型。BA表示二元准确率（越高越好，F1同理），MAE表示平均绝对误差（越低越好），Corr表示皮尔逊相关系数（越高越好）。对于BA和F1，我们报告了两个数字："/"左侧的数字是基于（Zadeh等人，2018c）计算的，右侧的数字是基于（Tsai等人，2019）计算的。CMU-MOSI的人类表现报告为（Zadeh等人，2018a）。

6.1 MAG-BERT的性能

在CMU-MOSI数据集的所有指标中，我们观察到MAG-BERT的性能优于使用BERT词嵌入的最先进多模态模型。此外，MAG-BERT的表现也优于微调的BERT。这实质上表明，MAG组件使BERT模型能够在微调期间适应多模态信息，从而实现了更优越的性能。

6.2 MAG-XLNet的性能

MAG-XLNet也观察到与MAG-BERT类似的性能趋势。除了优于基线和微调的XLNet之外，MAG-XLNet在CMU-MOSI数据集上实现了接近人类水平的性能。此外，我们使用微调的XLNet嵌入来训练MulT，并获得以下性能：$83.6/85.3$，$82.6/84.2$，0.810，0.759，低于MAG-XLNet和XLNet。值得注意的是，表1中MAG-XLNet和XLNet之间的学生t检验的p值在所有指标上都低于$10e-5$。

表1中报告的实验的动机如下：我们从预训练的BERT和XLNet模型中提取了词嵌入，并使用这些嵌入训练了基线模型。由于BERT和XLNet通常被认为提供比Glove更好的词嵌入，因此将MAG-BERT/MAG-XLNet与使用Glove嵌入训练的先前模型进行比较是不公平的。因此，我们重新使用BERT/XLNet嵌入对先前的工作进行了重新训练，以在本文提出的方法和先前的工作之间建立更公平的比较。根据表1中的信息，我们观察到MAG-BERT/MAG-XLNet模型在使用BERT/XLNet/Glove模型的各种基线模型方面表现出色。

6.3 不同层级上的适应性

我们还研究了在XLNet的不同编码器层级上应用MAG的效果。具体而言，我们首先将MAG应用于嵌入层的输出。随后，我们将MAG应用于XLNet的第j层（$j \in \{1, 4, 6, 8, 12\}$）。然后，我们在所有XLNet层上应用MAG。从表2中，我们观察到较早的层级更适合应用MAG。

我们认为较早的层级允许更好地整合多模态信息，因为它们允许单词的转换从网络的开始发生。如果单词的语义应该根据非语言伴随信息发生变化，那么初始层应该反映出语义的转变，否则这些层只是单模态工作。此外，BERT的较高层级学习了关于语言特征的句法和语义结构的更抽象和更高级的信息（Coenen等人，2019）。由于我们模型中的声学和视觉信息与话语中的每个单词相对应，因此MAG更难将从较晚的层级提取的向量进行转移，因为该向量的信息性质将非常抽象。

Model	E	1	4	6	8	12	A	$\oplus$	$\odot$
MAG-XLNet	80.1	85.6	84.1	84.1	83.8	83.6	64.0	60.0	55.8

表2：XLNet模型的变体结果：在XLNet模型的不同层次上应用MAG，输入级别的拼接和所有模态的加法。“E”表示在XLNet的嵌入层之后立即应用MAG，“A”表示在嵌入层和所有后续编码层之后应用MAG。$\oplus$和$\odot$分别表示输入级别的加法和拼接所有模态。在初始层应用MAG的效果总体上更好。

6.4 输入级别的串联和加法

从表2中，我们可以看到输入级别的串联和模态的加法都表现不佳。对于串联，我们只是简单地串联所有模态。对于加法，我们将音频和视觉信息添加到语言嵌入中，将它们都映射到语言维度后相加。这些结果显示了我们使用MAG这样的先进融合机制的合理性。

6.5 类似数据集上的结果

我们还在CMU-MOSEI数据集上进行了实验，以研究我们的方法在其他多模态语言数据集上的泛化能力（Zadeh等人，2018d）。与CMU-MOSI在话语级别上有情感标注不同，CMU-MOSEI在句子级别上有情感标注。CMU-MOSEI的实验方法与原始论文类似。为了比较，我们仅比较表1中排名前三的模型的二元准确率和F1分数。在BERT类别中，我们比较了MulT（使用BERT嵌入）、BERT和MAG-BERT的性能，分别为MulT的[83.5, 82.9]，BERT的[83.9, 83.9]，MAG-BERT的[84.7, 84.5]。类似地，在XLNet类别中，MulT（使用XLNet嵌入）、XLNet和MAG-XLNet的结果分别为MulT的[84.1, 83.7]，XLNet的[85.4, 85.2]，MAG-XLNet的[85.6, 85.7]。因此，MAG-BERT和MAG-XLNet的优越性能也适用于CMU-MOSEI数据集。

6.6 微调效果

我们研究了MAG-BERT和MAG-XLNet的优越性能是否与模型的成功微调有关，或者与其他因素（例如具有与BERT或XLNet类似架构的任何Transformer）相关，无论是否进行了预训练，都能够实现优越性能。通过在MAG-BERT和MAG-XLNet中随机初始化BERT和XLNet的权重，我们得到了在CMU-MOSI的二元准确率为70.1和70.7的结果。这表明MAG-BERT和MAG-XLNet的成功性能是由于成功的微调。即使在更大的CMU-MOSEI数据集上，我们得到了MAG-BERT和MAG-XLNet的准确率分别为76.8和78.4，这进一步证实了使用MAG框架进行微调的成功性。

#	Spoken words + acoustic and visual behaviors	Ground Truth	MAG-XLNet	XLNet
1	"And it really just lacked what made the other movies more enjoyable." + Frustrated and disappointed tone	-1.4	-1.41	-0.9
2	"But umm I liked it." + Emphasis on tone + positive shock through sudden eyebrow raise	1.8	1.9	1.2
3	"Except their eyes are kind of like this welcome to the polar express." + tense voice + frown expression	-0.6	-0.6	0.8
4	"Straight away miley cyrus acting miley cyrus, or lack of, she had this same expression throughout the entire film" + sarcastic voice + frustrated facial expression	-1.0	-1.2	0.2

表3：CMU-MOSI数据集的示例。ground truth 情感标签介于强烈负面（-3）和强烈正面（+3）之间。对于每个示例，我们展示了MAG-XLNet和XLNet的 ground truth 情感和预测输出。XLNet似乎主要复制语言模态，而MAG-XLNet成功地整合了非语言信息。

6.7 定性分析

在表3中，我们列举了一些示例，其中MAG-XLNet通过考虑非语言信息正确调整了情感强度。这些示例表明，MAG-XLNet能够成功地将非语言模态与文本信息进行整合。

在Example-1和Example-2中，XLNet正确预测了显示情感的极性。然而，声学和视觉领域中存在额外的信息，XLNet无法利用。在这些情况下，MAG-XLNet能够更好地预测显示的情感强度。

虽然Example-3中的文本情感可以被描绘为稍微积极，但紧张的声音和皱眉表情帮助MAG-XLNet扭转了预测情感的极性。类似地，Example-4中的文本大部分是中性的，但MAG-XLNet可以通过讽刺的语调和沮丧的面部表情预测出负面情感。

7 结论

本文介绍了一种用于高效微调大型预训练Transformer模型的多模态语言方法。通过使用提出的多模态适应门（MAG），在存在视觉和声学模态的情况下成功微调了BERT和XLNet。MAG基本上将非语言行为看作是一个具有轨迹和幅度的向量，随后用于在预训练的Transformer模型中转移词汇表示。MAG的一个独特特点是它不对BERT或XLNet的原始结构进行任何更改，而是作为这两个模型的附件。我们的实验证明了MAG-BERT和MAG-XLNet的优越性能。MAG-BERT和MAG-XLNet的代码在此处公开可用(https://github.com/WasifurRahman/BERT_multimodal_transformer)。