SELECT-ADDITIVE LEARNING: IMPROVING GENERALIZATION IN MULTIMODAL SENTIMENT ANALYSIS

选择性加法学习: 提高多模态情感分析的泛化能力

Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency and Eric P. Xing
Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency 和 Eric P. Xing
Language Technologies Institute
语言技术研究所
School of Computer Science
计算机科学学院
Carnegie Mellon University
卡内基梅隆大学
{haohanw, aaksham, morency, epxing}@cs.cmu.edu
{haohanw, aaksham, morency, epxing}@cs.cmu.edu

ABSTRACT

摘要

Multimodal sentiment analysis is drawing an increasing amount of attention these days. It enables mining of opinions in video reviews which are now available aplenty on online platforms. However, multimodal sentiment analysis has only a few high-quality data sets annotated for training machine learning algorithms. These limited resources restrict the generalizability of models, where, for example, the unique characteristics of a few speakers (e.g., wearing glasses) may become a confounding factor for the sentiment classification task. In this paper, we propose a Select-Additive Learning (SAL) procedure that improves the generalizability of trained neural networks for multimodal sentiment analysis. In our experiments, we show that our SAL approach improves prediction accuracy significantly in all three modalities (verbal, acoustic, visual), as well as in their fusion. Our results show that SAL, even when trained on one dataset, achieves good generalization across two new test datasets.

多模态情感分析近年来引起了越来越多的关注。它使得在如今在线平台上大量存在的视频评论中挖掘意见成为可能。然而，多模态情感分析仅有少量高质量的数据集被标注用于训练机器学习算法。这些有限的资源限制了模型的泛化能力，例如，少数说话者的独特特征 (例如，佩戴眼镜) 可能成为情感分类任务的混淆因素。本文提出了一种选择性加法学习 (SAL) 程序，旨在提高训练神经网络在多模态情感分析中的泛化能力。在我们的实验中，我们展示了 SAL 方法在所有三种模态 (语言、声学、视觉) 及其融合中显著提高了预测准确性。我们的结果表明，即使在一个数据集上训练，SAL 也能在两个新的测试数据集上实现良好的泛化。

Index Terms- multimodal, sentiment analysis, cross-datasets, generalization, cross-individual
关键词- 多模态，情感分析，跨数据集，泛化，跨个体

1. INTRODUCTION

1. 引言

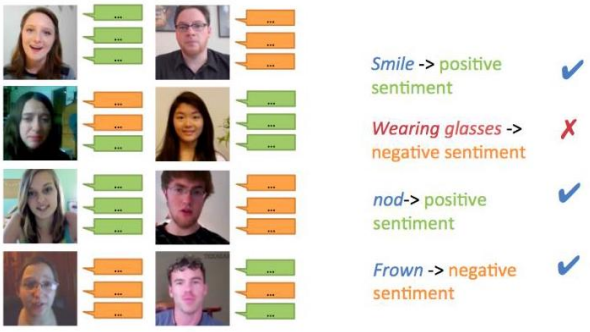
Sentiment analysis is the automatic identification of the private state of a human mind with a focus on determining whether this state is positive, negative or neutral [1]. It has been extensively studied in the last few decades [2], primarily based on textual data. With the recent proliferation of online avenues for sharing multimedia content, people are posting more and more videos with opinions. The opinions are expressed through the spoken word (verbal modality), how these words are spoken (acoustic modality) and what gestures and facial expressions accompany the spoken words (visual modality). Multimodal sentiment analysis extends traditional textual sentiment analysis by analyzing all three modalities present in online videos, including acoustic and visual modalities [3, 4].

情感分析是自动识别人类心理状态的过程，重点在于确定该状态是积极的、消极的还是中性的 [1]。在过去几十年中，这一领域得到了广泛研究 [2]，主要基于文本数据。随着在线多媒体内容分享渠道的迅速增加，人们发布的意见视频也越来越多。这些意见通过口头表达 (语言模态)、这些话语的表达方式 (声学

模态) 以及伴随口语的手势和面部表情 (视觉模态) 来表达。多模态情感分析通过分析在线视频中存在的所有三种模态，包括声学和视觉模态，扩展了传统的文本情感分析 [3, 4]。

To foster research in this area, a few datasets have been created with quality annotations for sentiment [1, 5, 6], but unfortunately the total number of annotations is still in the order of thousand samples. These limited-size resources make it challenging for conventional machine learning algorithms to generalize well across datasets. In these limited data scenarios, a unique characteristic of a few speakers in the training dataset (e.g., wearing glasses) can end up creating a confounding effect with the sentiment classification task. Fig. 1(a) shows one illustrative example where limited data can bring in learning and generalization challenges. Since in this example all individuals with glasses happen to be expressing negative sentiments, the classifier ends up learning an association between visual appearance of wearing glasses and negative sentiment (see Fig. 1(b)) .

为了促进该领域的研究，已经创建了一些具有高质量情感标注的数据集 [1, 5, 6]，但不幸的是，标注的总数量仍然在千个样本的范围内。这些有限规模的资源使得传统机器学习算法在跨数据集的泛化上面临挑战。在这些有限数据的场景中，训练数据集中少数说话者的独特特征 (例如，佩戴眼镜) 可能会对情感分类任务产生混淆效应。图 1(a) 展示了一个说明性的例子，其中有限的数据可能带来学习和泛化的挑战。由于在这个例子中，所有佩戴眼镜的个体恰好都表达了消极情感，因此分类器最终学习到了佩戴眼镜的视觉外观与消极情感之间的关联 (见图 1(b))。



(a) An illustrative data set (b) Rules learned

Fig. 1: An illustrative data set demonstrating the "wearing glass" as a confounding factor. Due to the limited amount of data, the model learns that wearing glasses means negative sentiment, which is only applicable to this training data set. (Orange denotes negative sentiment; green denotes positive sentiment; blue denotes correct rules & red denotes incorrect rules).

图 1: 一个说明性数据集，展示了 "佩戴眼镜" 作为混淆因素。由于数据量有限，模型学习到佩戴眼镜意味着消极情感，这仅适用于该训练数据集。(橙色表示消极情感；绿色表示积极情感；蓝色表示正确规则；红色表示错误规则)。

The role that the visual appearance plays here is statistically known as a confounding factor [7, 8]. To generalize across datasets and individuals, a robust multimodal sentiment classifier should not include features from a confounding factor. In other words, the prediction of the sentiment polarity should not be dependent on the speaker's unique characteristic, namely, identity of the speaker.

视觉外观在这里所扮演的角色在统计上被称为混杂因素 [7, 8]。为了在数据集和个体之间进行概括，一个稳健的多模态情感分类器不应包含来自混杂因素的特征。换句话说，情感极性的预测不应依赖于说话者的独特特征，即说话者的身份。

Before going further in this research agenda, we studied if the confounding factor also exists in the real-world multimodal sentiment analysis datasets. In the MOSI multimodal sentiment analysis data set [6], we tested for the null hypothesis that sentiment is independent of an individual's identity. Chi-square test obtains a p-value of 1.202×10^{-19} , which strongly suggests the dependence between individual identities and the expressed sentiment. Consequently, naively applying machine learning algorithms on this dataset will most likely result in a suboptimal model that misinterprets an individual's identity as prescient information for sentiment.

在进一步推进这一研究议程之前，我们研究了混杂因素是否也存在于真实世界的多模态情感分析数据集中。在 MOSI 多模态情感分析数据集 [6] 中，我们检验了情感与个体身份独立的零假设。卡方检验得到了 p 值为 1.202×10^{-19} ，这强烈表明个体身份与表达的情感之间存在依赖关系。因此，天真地在该数据集上应用机器学习算法很可能会导致一个次优模型，将个体身份误解为情感的先验信息。

In this paper, we propose a Select-Additive Learning (SAL) procedure that addresses the confounding factor problem, specifically for neural architectures such as convolutional neural networks. Our proposed

SAL approach is a two-phase procedure with the (Selection phase and Addition phase. During the Selection phase, SAL identifies the confounding factors from the latent representation learned by neural networks. During the Addition phase, SAL forces the original model to discard (or rather, give less importance to) the confounding elements by adding Gaussian noises to these representations. We conduct extensive experiments to test the performances of state-of-art neural-based models enhanced by SAL. All our experiments are performed in a person-independent setting, where subjects in the test set are different from the training and validation sets. We test the generalization with both, within-data and across-datasets experiments.

在本文中，我们提出了一种选择-加法学习 (SAL) 程序，专门解决混杂因素问题，特别是针对卷积神经网络等神经架构。我们提出的 SAL 方法是一个两阶段程序，包括选择阶段和加法阶段。在选择阶段，SAL 从神经网络学习的潜在表示中识别混杂因素。在加法阶段，SAL 通过向这些表示添加高斯噪声，迫使原始模型丢弃 (或更确切地说，降低对) 混杂元素的重要性。我们进行了广泛的实验，以测试通过 SAL 增强的最先进的基于神经网络的模型的性能。我们所有的实验都是在一个与人无关的设置中进行的，其中测试集中的受试者与训练集和验证集中的受试者不同。我们通过数据内和跨数据集的实验来测试泛化能力。

2. RELATED WORK

2. 相关工作

Multimodal data has been studied for a variety of applications to analyze human behaviors, including person detection and identification [9, 10], human action recognition [11, 12], face recognition [13, 14], as well as sentiment analysis.

多模态数据已被研究用于分析人类行为的各种应用，包括人脸检测和识别 [9, 10]、人类动作识别 [11, 12]、面部识别 [13, 14] 以及情感分析。

Originating from analysis of the textual modality, sentiment analysis has been carried out at the word level [15], phrase level [16] and sentence level [17]. [18] performed sentiment analysis on audio data by first transcribing the spoken words and then performing sentiment analysis. Related to audio-based sentiment analysis is the task of estimating emotional state of the speaker from audio input [19]. For the visual modality, the Facial Action Coding System [20] laid the groundwork for analyzing facial expressions and emotions. Recently, convolutional neural networks were used to discover the affective regions for sentiment on still images

情感分析起源于对文本模态的分析，已在词级 [15]、短语级 [16] 和句子级 [17] 进行。[18] 通过首先转录口语并随后进行情感分析，对音频数据进行了情感分析。与基于音频的情感分析相关的任务是从音频输入中估计说话者的情感状态 [19]。对于视觉模态，面部动作编码系统 [20] 为分析面部表情和情感奠定了基础。最近，卷积神经网络被用于发现静态图像中情感的影响区域。[21].

The fusion of textual, acoustic and visual modalities for sentiment analysis has drawn increasing attention lately [1]. A variety of methods have been proposed and extensively discussed in recent years [22, 23, 24]. The state-of-the-art performance is achieved with a Convolutional Neural Network

文本、声学和视觉模态在情感分析中的融合最近引起了越来越多的关注 [1]。近年来提出并广泛讨论了多种方法 [22, 23, 24]。最先进的性能是通过卷积神经网络实现的。[25].

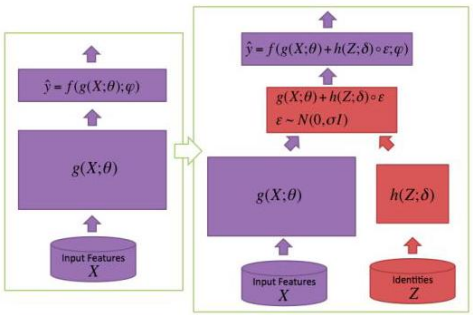


Fig. 2: The SAL architecture is achieved by a simple extension of a general deep learning discriminative classifier. The purple part is the original deep learning model. The red part is the extension SAL introduces. The extension network is connected to the original network via a Gaussian Sampling Layer.

图 2:SAL 架构是通过对一般深度学习判别分类器的简单扩展实现的。紫色部分是原始深度学习模型。红色部分是 SAL 引入的扩展。扩展网络通过高斯采样层与原始网络连接。

Our proposed Select-Additive Learning (SAL) procedure improves the generalizability of neural networks. Our experiments show improved prediction accuracy for all three modalities (verbal, acoustic and visual) as well as for multimodal fusion. The following section introduces our proposed Select-Additive Learning procedure.

我们提出的选择-加性学习 (SAL) 过程提高了神经网络的泛化能力。我们的实验显示，在所有三种模态 (语言、声学和视觉) 以及多模态融合方面，预测准确性都有所提高。以下部分介绍我们提出的选择-加性学习过程。

3. SELECT-ADDITIVE LEARNING

3. 选择-加性学习

The main goal of our work is to increase the generalizability of multimodal sentiment prediction models by encouraging the model to consider sentiment-associated features (i.e. people are smiling while expressing positive sentiment) more than the identity-related features (i.e. wearing glasses).

我们工作的主要目标是通过鼓励模型更多地考虑与情感相关的特征 (即人们在表达积极情感时微笑) 而非与身份相关的特征 (即佩戴眼镜)，来提高多模态情感预测模型的泛化能力。

We formalize the problem by defining an input feature matrix X of size $n \times p$ that encodes the p features for n utterances. In the multimodal scenario, p will be the total number of verbal, acoustic and visual features. We also define a vector y of size $n \times 1$ which represents the sentiment of each utterance. Finally, we define a new matrix Z which encodes for each utterance the speaker identity in a one-hot matrix of size $n \times m$ where m represents the total number of unique individuals in the dataset.

我们通过定义一个输入特征矩阵 X ，其大小为 $n \times p$ ，来形式化这个问题，该矩阵编码了 p 特征用于 n 发言。在多模态场景中，p 将是口头、声学和视觉特征的总数。我们还定义了一个大小为 $n \times 1$ 的向量 y ，它表示每个发言的情感。最后，我们定义了一个新的矩阵 Z ，该矩阵为每个发言编码说话者身份，采用大小为 $n \times m$ 的独热矩阵，其中 m 表示数据集中独特个体的总数。

3.1. Select-Additive Learning Architecture

3.1. 选择-加性学习架构

Our proposed SAL procedure is designed to enhance a preexisting (i.e. pre-trained model) discriminative neural network to be more robust against confounding factors. To formally introduce our SAL approach, we define two main components present in most discriminative neural network classifiers (e.g., Convolutional Neural Network, CNN): a representation learner component and a classification component. To simplify the notation, we use $g(\cdot;\theta)$ denotes the representation learner component and θ stands for its parameters. Our hypothesis is that confounding factors will be constrained to a subset of dimensions present in $g(\cdot;\theta)$. Similarly, we use $f(\cdot;\phi)$ to denote the classification component and ϕ denotes the parameters. Therefore, a full neural network classifier is denoted as $f(g(\cdot;\theta);\phi)$. In our SAL approach, $g(\cdot;\theta)$ from identity-related features as identity related confounding dimensions. Our SAL approach can be summarized as first identifying these dimensions and then reduce the impact of these dimensions by adding noise to them.

我们提出的 SAL 程序旨在增强一个预先存在的 (即预训练模型) 判别神经网络，使其对混淆因素更具鲁棒性。为了正式介绍我们的 SAL 方法，我们定义了大多数判别神经网络分类器中存在的两个主要组件 (例如，卷积神经网络 CNN): 一个表示学习组件和一个分类组件。为了简化符号表示，我们用 $g(\cdot;\theta)$ 表示表示学习组件，用 θ 表示其参数。我们的假设是，混淆因素将被限制在 $g(\cdot;\theta)$ 中存在的一个维度子集内。同样，我们用 $f(\cdot;\phi)$ 表示分类组件，用 ϕ 表示其参数。因此，一个完整的神经网络分类器表示为 $f(g(\cdot;\theta);\phi)$ 。在我们的 SAL 方法中，$g(\cdot;\theta)$ 来自与身份相关的特征，作为与身份相关的混淆维度。我们的 SAL 方法可以总结为首先识别这些维度，然后通过向它们添加噪声来减少这些维度的影响。

To select identity-related confounding dimensions, SAL introduces a simple neural network (denoted by $h(\cdot;\delta)$, where δ stands for its parameters). This is to predict identity-related confounding dimensions from individual identities Z , by minimizing the difference between $h(Z;\delta)$ and $g(X;\theta)$. Therefore, $h(Z;\delta)$ will effectively pinpoint the identity-related confounding dimensions in $g(X;\delta)$. Figure 3a shows an overview of this Selection Phase.

为了选择与身份相关的混杂维度，SAL 引入了一个简单的神经网络 (用 $h(\cdot;\delta)$ 表示，其中 δ 代表其参数)。这是为了通过最小化 $h(Z;\delta)$ 和 $g(X;\theta)$ 之间的差异，从个体身份 Z 中预测与身份相关的混杂维度。因此，$h(Z;\delta)$ 将有效地确定 $g(X;\delta)$ 中与身份相关的混杂维度。图 3a 显示了这一选择阶段的概述。

To force the model to discard identity-related confounding dimensions, SAL introduces Gaussian noise to these dimensions while minimizing prediction error, so that $f(\cdot;\phi)$ learns to neglect noised representation. The noise is added through a Gaussian Sampling Layer [26]. Figure 3b shows an overview of this addition phase.

为了迫使模型舍弃与身份相关的混杂维度，SAL 在最小化预测误差的同时向这些维度引入高斯噪声，以便 $f(\cdot;\phi)$ 学会忽略噪声表示。噪声是通过高斯采样层 [26] 添加的。图 3b 显示了这一添加阶段的概述。

Figure 2 shows how SAL assembles $g(\cdot;\theta)$, $f(\cdot;\phi)$ and $h(\cdot;\delta)$ together via a Gaussian Sampling Layer.

图 2 显示了 SAL 如何通过高斯采样层将 $g(\cdot;\theta)$, $f(\cdot;\phi)$ 和 $h(\cdot;\delta)$ 组合在一起。

3.2. Select-Additive Learning Algorithm

3.2. 选择-加法学习算法

A pre-requisite to our Select-Additive Learning (SAL) approach is first learn a discriminative neural classifier. On our experiments, we achieve this goal by minimizing the following lost function:

我们的选择-加法学习 (SAL) 方法的前提是首先学习一个区分性的神经分类器。在我们的实验中，我们通过最小化以下损失函数来实现这一目标:

$$\arg\min_{\phi,\theta} \frac{1}{2}(y - f(g(X;\theta);\phi))^2$$

The same loss function is often used in discriminative neural networks [27].

相同的损失函数通常用于区分性神经网络 [27]。

3.2.1. Selection Phase

3.2.1. 选择阶段

Once the original representation $g(X;\theta)$ is learned, the selection phase optimizes a new loss function to discover the identity related confounding dimensions. This selection phase is operationalized by tuning the parameters λ using the following loss function (as illustrated in Figure 3(a)):

一旦学习了原始表示 $g(X;\theta)$，选择阶段就会优化一个新的损失函数，以发现与身份相关的混杂维度。此选择阶段通过使用以下损失函数调整参数 λ 来实现 (如图 3(a) 所示):

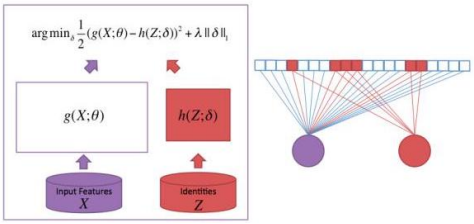
$$\arg\min_{\delta} \frac{1}{2}(g(X;\theta) - h(Z;\delta))^2 + \lambda \parallel \delta \parallel_1 \quad (1)$$

where λ is a scalar that controls the weight of the sparsity regularizer. In this phase, both X and Z are available, but only δ is tuned, as shown in Fig. 3(a).

其中 λ 是一个标量，用于控制稀疏正则化器的权重。在此阶段，X 和 Z 都可用，但仅调整 δ，如图 3(a) 所示。

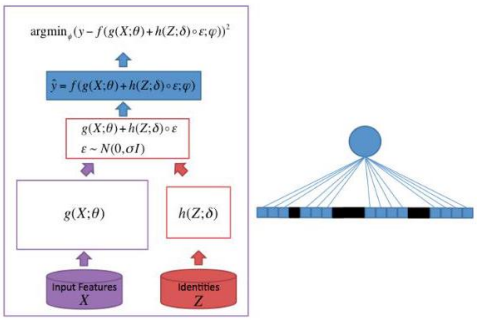
The goal of this phase is to select identity-related confounding dimensions from the original representation. To achieve this, we tune δ to minimize the difference between $g(X;\theta)$ and $h(Z;\delta)$. As Z only encodes identity information,

本阶段的目标是从原始表示中选择与身份相关的混杂维度。为此，我们调整 δ 以最小化 $g(X;\theta)$ 和 $h(Z;\delta)$ 之间的差异。由于 Z 仅编码身份信息，



(a) Selection Phase: SAL forces $h(Z;\delta)$ to identify identity-related confounding dimensions. Right side figure shows that $h(Z;\delta)$ selects these dimensions.

(a) 选择阶段:SAL 强制 $h(Z;\delta)$ 识别与身份相关的混杂维度。右侧图形显示 $h(Z;\delta)$ 选择了这些维度。



(b) Addition Phase: SAL forces the model to focus on other dimensions by adding Gaussian noise to identity-related confounding dimensions. Right side shows that the model shifts focus because these dimensions are contaminated/noisy.

(b) 添加阶段:SAL 强制模型通过向与身份相关的混杂维度添加高斯噪声来关注其他维度。右侧显示模型的关注点发生了变化，因为这些维度受到污染/噪声影响。

Fig. 3: Illustration of SAL. On the left, network structure and training objective is presented. On the right, circles denote neurons. Squares denote dimensions of representation. the minimum of difference will be achieved when $h(Z;\delta)$ is matched to the identity-related confounding dimensions of $g(X;\theta)$. L1 regularization of δ is necessary to avoid overfit-ting as output dimension of $h(\cdot;\delta)$ is typically significantly higher than input dimension.

图 3:SAL 的示意图。左侧展示了网络结构和训练目标。右侧，圆圈表示神经元，方块表示表示的维度。当 $h(Z;\delta)$ 与 $g(X;\theta)$ 的与身份相关的混杂维度匹配时，将实现差异的最小值。对 δ 进行 L1 正则化是必要的，以避免过拟合，因为 $h(\cdot;\delta)$ 的输出维度通常显著高于输入维度。

The result of this selection phase is shown on the righthand-side of Fig. 3(a). All the weights of original model (purple circle) are active and connected to every dimension while only some weights of $h(\cdot,\delta)$ (red circle) are active and connected to the identity-related confounding dimensions

选择阶段的结果显示在图 3(a) 的右侧。原始模型的所有权重 (紫色圆圈) 都是活跃的并连接到每个维度，而只有一些 $h(\cdot,\delta)$ 的权重 (红色圆圈) 是活跃的并连接到与身份相关的混杂维度。

3.2.2. Addition Phase

3.2.2. 添加阶段

After the selection phase $h(\cdot,\delta)$ should be pointing at the identity-related confounding dimensions. Our remaining step is to learn a new neural network classifier where the confounding dimensions have "masked". We achieve this by adding Gaussian noise. Our addition phase defines the following loss function to achieve this goal:

在选择阶段之后，$h(\cdot,\delta)$ 应该指向与身份相关的混杂维度。我们剩下的步骤是学习一个新的神经网络分类器，其中混杂维度已被 "屏蔽"。我们通过添加高斯噪声来实现这一点。我们的添加阶段定义了以下损失函数以实现这一目标:

$$\arg\min_{\phi} \frac{1}{2}(y - f(g(X;\theta) + h(Z;\delta)\circ\epsilon;\phi))^2 \tag{2}$$

where $\epsilon \sim N(0,\sigma I)$ and $\circ$ stands for element-wise product, as showed in Fig. 3(b).

其中 $\epsilon \sim N(0,\sigma I)$ 和 $\circ$ 表示逐元素乘积，如图 3(b) 所示。

In this phase, parameter ϕ is tuned. The input representation of $f(\cdot;\phi)$ consists of the representation learned from $g(X;\theta)$ and the $h(Z;\delta)$ -selected identity-related confounding dimensions with Gaussian noise added. The noise ensures that identity-related confounding dimensions are no longer informative so that $f(\cdot;\phi)$ can be trained to ignore them.

在此阶段，参数 ϕ 被调整。$f(\cdot;\phi)$ 的输入表示由从 $g(X;\theta)$ 学习到的表示和添加了高斯噪声的 $h(Z;\delta)$ 选择的与身份相关的混淆维度组成。噪声确保与身份相关的混淆维度不再具有信息性，从而使得 $f(\cdot;\phi)$ 可以被训练以忽略它们。

As illustrated on the right side of Fig. 3(b), identity-related confounding dimensions are contaminated with addition of noise. Therefore, the model learns to discard these non-informative dimensions, and its weights get optimized to focus on the rest of the dimensions.

如图 3(b) 右侧所示，与身份相关的混淆维度在添加噪声后受到污染。因此，模型学习丢弃这些非信息性维度，其权重被优化以关注其余维度。

4. EXPERIMENTS

4. 实验

In this section, we perform an extensive set of experiments on three different data sets to see whether SAL can help improve the generalizability of a discriminative neural classifier. Generalizability is tested by performing across-dataset experiments where two of the dataset are kept exclusively for testing. All our experiments follow a person independent methodology where none of the subject from the training data are present in the test datasets.

在本节中，我们对三个不同的数据集进行了广泛的实验，以观察 SAL 是否能帮助提高判别神经分类器的可泛化性。可泛化性通过进行跨数据集实验来测试，其中两个数据集专门用于测试。我们所有的实验遵循一种与人无关的方法论，训练数据中的任何受试者均不出现在测试数据集中。

4.1. Models

4.1. 模型

We compare the following models:

我们比较以下模型:

⋆**CNN** : The state-of-the-art seven layer convolutional neural network architecture used previously for multimodal sentiment analysis [25].

⋆**CNN**：之前用于多模态情感分析的最先进的七层卷积神经网络架构 [25]。

*SAL-CNN: After the state-of-the-art CNN is fully trained, we use SAL to increase its generalizability and predict sentiment. $h(\cdot,\delta)$ is a neural perceptron [27].

*SAL-CNN: 在最先进的 CNN 完全训练后，我们使用 SAL 来提高其可泛化性并预测情感。$h(\cdot,\delta)$ 是一个神经感知器 [27]。

4.2. Datasets

4.2. 数据集

We performed our experiment on three multimodal sentiment analysis data sets:

我们在三个多模态情感分析数据集上进行了实验:

MOSI: This dataset consists of 93 videos obtained from YouTube channels. Each video contains the opinions from one unique individual. The dataset has 2199 utterances manually segmented from online videos of movie reviews. Each utterance was also manually annotated for sentiment label [6].

MOSI: 该数据集由 93 个从 YouTube 频道获得的视频组成。每个视频包含来自一个独特个体的观点。该数据集包含 2199 个从电影评论的在线视频中手动分割的发言。每个发言也经过手动标注以获取情感标签 [6]。

⋆ YouTube: This dataset consists of 47 opinion videos with 280 utterances with manually annotated sentiment labels [1]. Each video contains the opinions from one unique individual.

⋆ YouTube: 该数据集由 47 个意见视频组成，包含 280 个带有手动注释情感标签的发言 [1]。每个视频包含一个独特个体的观点。

*MOUD: This dataset consists of 498 Spanish opinion utterances from 55 unique individuals [5].

*MOUD: 该数据集由来自 55 个独特个体的 498 个西班牙语意见发言组成 [5]。

Although, majority of the data originate from YouTube, they differ in recording quality and the processing done after curation. The verbal features in the MOUD dataset need one extra step of translation from Spanish to English. These three datasets are good candidates to evaluate across-dataset generalization.

尽管大多数数据来自 YouTube，但它们在录制质量和后期处理上有所不同。MOUD 数据集中的语言特征需要额外一步将西班牙语翻译为英语。这三个数据集是评估跨数据集泛化的良好候选者。

Table 1: Within data set experiments

表 1: 数据集内实验

		CNN	SAL-CNN
Unimodal	Verbal	0.678	0.732
	Acoustic	0.588	0.618
	Visual	0.572	0.636
Bimodal	Verbal+Acoustic	0.687	0.725
	Verbal+Visual	0.706	0.73
	Acoustic+Visual	0.661	0.621
All Modalities		0.715	0.73

		卷积神经网络	SAL-CNN
单模态	语言	0.678	0.732
	声音	0.588	0.618
	视觉	0.572	0.636
双模态	语言 + 声音	0.687	0.725
	语言 + 视觉	0.706	0.73
	声音 + 视觉	0.661	0.621
所有模态		0.715	0.73

4.3. Feature Extraction

4.3. 特征提取

We extracted an embedding for each word using a word2vec dictionary pre-trained on a Google News corpus [28]. The text feature of each utterance was formed by concatenating the word embeddings for all the words in the sentence and padding them with the appropriate zeros to have the same dimension. We set the maximum length as 60 and discarded additional words'. For YouTube dataset, we extracted the transcripts using the IBM Bluemix' s speech2text API For MOUD dataset, we translated Spanish transcripts into English transcripts. We used openSMILE [29] to extract the low-level audio descriptors for each spoken utterance. These audio descriptors included the Mel-frequency cepstral coefficients, pitch and voice quality. We processed every frame in each video and used the audio-visual synchrony to identify which frames happen during a specific utterance. We used the CLM-Z library [30] for extracting facial characteristic points.

我们使用在 Google News 语料库上预训练的 word2vec 字典为每个单词提取了嵌入 [28]。每个发言的文本特征是通过连接句子中所有单词的嵌入并用适当的零填充以保持相同维度来形成的。我们将最大长度设置为 60，并丢弃额外的单词。对于 YouTube 数据集，我们使用 IBM Bluemix 的 speech2text API 提取了转录文本。对于 MOUD 数据集，我们将西班牙语转录文本翻译为英语转录文本。我们使用 openSMILE [29] 提取了每个口语发言的低级音频描述符。这些音频描述符包括梅尔频率倒谱系数、音高和音质。我们处理了每个视频中的每一帧，并利用音视频同步来识别特定发言发生时的帧。我们使用 CLM-Z 库 [30] 提取面部特征点。

4.4. Experiment Setup

4.4. 实验设置

We remove the netural utterances out of the data set. The first 62 individuals in the MOSI data set are selected as training/validation set. There are around 1250 utterances in total. These utterances are shuffled and then 80% are used for training and 20% used for validation. We have three test datasets. 1) MOSI: 546 utterances from the remaining 31 individuals. 2) YouTube: 195 utterances from 47 individuals

and 3) MOUD: 450 utterances from 55 individuals. We use MOSI as training set because it is the largest and most recent dataset among all three.

我们从数据集中移除了中立的发言。MOSI 数据集中的前 62 个人被选为训练/验证集。总共有大约 1250 个发言。这些发言被打乱，然后 80% 用于训练，20% 用于验证。我们有三个测试数据集。1) MOSI: 来自剩余 31 个人的 546 个发言。2) YouTube: 来自 47 个人的 195 个发言，3) MOUD: 来自 55 个人的 450 个发言。我们使用 MOSI 作为训练集，因为它是三者中最大和最新的数据集。

4.5. Experiment Results

4.5. 实验结果

4.5.1. Within data set

4.5.1. 数据集内

Table 1 shows the results for CNN and SAL-CNN tested on the remaining 31 individuals' data of MOSI. The results indicate that SAL could help to increase the generalizability of the trained model.

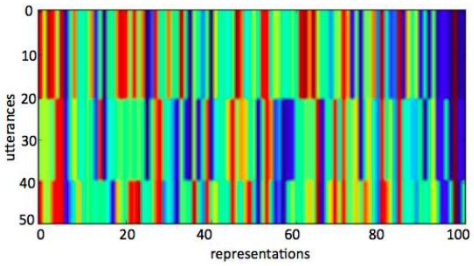
表 1 显示了在剩余 31 个人的 MOSI 数据上测试的 CNN 和 SAL-CNN 的结果。结果表明，SAL 可以帮助提高训练模型的泛化能力。

Table 2: Across data set experiments

表 2: 跨数据集实验

	Youtube		MOUD	
	CNN	SAL-CNN	CNN	SAL-CNN
Verbal	0.605	0.657	0.522	0.569
Acoustic	0.441	0.564	0.455	0.549
Visual	0.492	0.549	0.555	0.548
Ver+Acou	0.642	0.652	0.515	0.574
Ver+Vis	0.642	0.667	0.542	0.574
Acou+Vis	0.452	0.559	0.533	0.554
All	0.611	0.667	0.531	0.574

	YouTube		MOUD	
	CNN	SAL-CNN	CNN	SAL-CNN
语言	0.605	0.657	0.522	0.569
声学	0.441	0.564	0.455	0.549
视觉	0.492	0.549	0.555	0.548
语言 + 声学	0.642	0.652	0.515	0.574
语言 + 视觉	0.642	0.667	0.542	0.574
声学 + 视觉	0.452	0.559	0.533	0.554
所有	0.611	0.667	0.531	0.574



[1] only around 0.5% utterances in our datasets have more than 60 words

[1] 我们的数据集中只有大约 0.5% 个发言超过 60 个单词

[2] https://www.ibm.com/watson/developercloud/speech-to-text.html

[2] https://www.ibm.com/watson/developercloud/speech-to-text.html

Fig. 4: Confounding factors identified in the Selection phase for first 50 utterances (rows), first 100 representation values (columns) in the training set.

图 4: 在训练集中识别的前 50 个发言 (行)、前 100 个表示值 (列) 选择阶段的混淆因素。

4.5.2. Across data sets

4.5.2. 跨数据集

Table 2 shows the results for CNN and SAL-CNN tested on YouTube and MOUD dataset. First, it is noteworthy that in some cases the performance of the CNN is worse than mere chance. This inferior performance substantiates the existence of the non-generalization problems we are targeting.

表 2 显示了在 YouTube 和 MOUD 数据集上测试的 CNN 和 SAL-CNN 的结果。首先，值得注意的是，在某些情况下，CNN 的表现甚至不如随机猜测。这种劣质表现证实了我们所针对的非泛化问题的存在。

Overall, Select-Additive Learning increases the robustness and performance of the previous models consistently (except only two cases: Video modality in MOUD and fusion of acoustic & visual in MOSI). Permutation Test rejects the null hypothesis (no improvement) with p-values 0.037,0.0003,0.0023respectively for MOSI, YouTube, and MOUD, indicating significant improvement. 3

总体而言，选择性加性学习持续提高了先前模型的鲁棒性和性能 (仅有两个例外:MOUD 中的视频模态和 MOSI 中的声学与视觉融合)。置换检验以 p 值 0.037、0.0003、0.0023 分别拒绝了零假设 (无改善)，表明显著改善。

4.6. Discussion

4.6. 讨论

To substantiate our proposed model and algorithm, we examine the learning process and verify that representation of confounding factors exists and our method can mitigate its effects. We demonstrate this with the visual modality as it intuitively contributes the most to confounding.

为了证实我们提出的模型和算法，我们检查学习过程并验证混杂因素的表示存在，我们的方法可以减轻其影响。我们通过视觉模态展示这一点，因为它直观上对混杂因素的贡献最大。

Figure 4 shows a plot of $h(Z, \delta)$ during the Selection phase. It is a zoomed-in figure for the first 50 utterances (rows) and first 100 values of the representation vector (columns). Blue indicates lowest values and red indicates highest values and other colors are linearly interpolated.

图 4 显示了选择阶段期间 $h(Z, \delta)$ 的图。它是前 50 个发言 (行) 和表示向量前 100 个值 (列) 的放大图。蓝色表示最低值，红色表示最高值，其他颜色为线性插值。

The representation of utterances forms clear clusters and each cluster belongs to one person. Despite each individual having their own pattern, some dimensions have generalized well across individuals. Our model learns to assign more weights to these dimensions after noise is introduced.

发言的表示形成了清晰的聚类，每个聚类属于一个人。尽管每个个体都有自己的模式，但某些维度在个体之间有很好的概括。我们的模型在引入噪声后学会对这些维度赋予更多权重。

In addition to these results, we calculated the inter-cluster distance over intra-cluster distance ratio for the representation learned under two situations: 1) clustered by category of sentiment and 2) clustered by individual's identity. We compared the ratios for CNN and SAL-CNN. The higher ratio indicates a clearer clustering structure.

除了这些结果外，我们计算了在两种情况下学习到的表示的聚类间距离与聚类内距离的比率:1) 按情感类别聚类，2) 按个体身份聚类。我们比较了 CNN 和 SAL-CNN 的比率。更高的比率表明更清晰的聚类结构。

After SAL, for representation clustered by category of sentiment, the ratio increased by $44\%, 15\%$ and 72% respectively for verbal, acoustic, and visual modality, while for representation clustered by individual's identity, the ratio increased by $9\%, 3\%$ and 13% , respectively. These numbers indicate SAL almost maintains the clustering structure of identity, but greatly improves the clustering structure of category of sentiment. This shows the effectiveness of SAL.

在 SAL 之后，对于按情感类别聚类的表示，口头、声学和视觉模态的比例分别增加了 $44\%, 15\%$ 和 72% ，而对于按个体身份聚类的表示，比例分别增加了 $9\%, 3\%$ 和 13% 。这些数字表明 SAL 几乎保持了身份的聚类结构，但大大改善了情感类别的聚类结构。这显示了 SAL 的有效性。

5. CONCLUSION

5. 结论

High-quality datasets required to train machine learning models for automatic multimodal sentiment analysis are only of the order of a few thousand samples. These limited resources restrict models' generalizability, leading to the issue of confounding factors. We proposed a Select-Additive Learning (SAL) procedure that can mitigate this problem. With extensive experiments, we have shown how SAL improves the generalizability of state-of-the-art models. We increased prediction accuracy significantly in all three modalities (verbal, acoustic, visual), as well as in their fusion. We also showed that SAL could achieve good prediction accuracy even when tested across data sets.

训练用于自动多模态情感分析的机器学习模型所需的高质量数据集仅在几千个样本的数量级。这些有限的资源限制了模型的泛化能力，导致混杂因素的问题。我们提出了一种选择性加法学习 (SAL) 程序，可以缓解这个问题。通过大量实验，我们展示了 SAL 如何提高最先进模型的泛化能力。我们在所有三种模态 (口头、声学、视觉) 及其融合中显著提高了预测准确性。我们还展示了 SAL 即使在跨数据集测试时也能达到良好的预测准确性。

6. REFERENCES

6. 参考文献

[1] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in Proceedings of the 13th international conference on multimodal interfaces. ACM, 2011.

[1] Louis-Philippe Morency, Rada Mihalcea, 和 Payal Doshi, " 朝向多模态情感分析: 从网络中收集意见," 载于第十三届国际多模态接口会议论文集。ACM, 2011。

[2] Bo Pang and Lillian Lee, "Opinion mining and sentiment analysis," Foundations and trends in information retrieval, 2008.

[2] Bo Pang 和 Lillian Lee, " 意见挖掘与情感分析," 信息检索的基础与趋势, 2008。

[3] Akshi Kumar and Mary Sebastian Teeja, "Sentiment analysis: A perspective on its past, present and future," International Journal of Intelligent Systems and Applications, 2012.

[3] Akshi Kumar 和 Mary Sebastian Teeja, " 情感分析: 对其过去、现在和未来的看法," 国际智能系统与应用期刊, 2012。

[4] Martin Wollmer, Felix Weninger, Timo Knaup, Bjorn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe

[4] Martin Wollmer, Felix Weninger, Timo Knaup, Bjorn Schuller, Congkai Sun, Kenji Sagae, 和 Louis-Philippe

Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," Intelligent Systems, IEEE, 2013.

[5] Verónica Pérez Rosas, Rada Mihalcea, and Louis-Philippe Morency, "Multimodal sentiment analysis of spanish online videos," IEEE Intelligent Systems,, no. 3, 2013.

[6] Amir Zadeh, "Micro-opinion sentiment intensity analysis and summarization in online videos," in ICMI. ACM, 2015.

[7] Robert M Ewers and Raphael K Didham, "Confounding factors in the detection of species responses to habitat fragmentation," Biological Reviews, 2006.

[8] Haohan Wang and Jingkang Yang, "Multiple confounders correction with regularized linear mixed effect models, with application in biological processes," in BIBM. IEEE, 2016.

[9] Lingxiang Wu, Jinqiao Wang, Guibo Zhu, Min Xu, and Hanqing Lu, "Person re-identification via rich color-gradient feature," in ICME. IEEE, 2016.

[10] Xiaoke Zhu, Xiao-Yuan Jing, Fei Wu, Weishi Zheng, Ruimin Hu, Chunxia Xiao, and Chao Liang, "Distance learning by treating negative samples differently and exploiting impostors with symmetric triplet constraint for person re-identification," in ICME. IEEE, 2016.

[3] Select-additive Learning implementation is available at

[3] 选择性加法学习的实现可在 https://github.com/HaohanWang/SelectAdditiveLearning

[11] Antonio Tejero-de Pablos, Yuta Nakashima, Tomokazu Sato, and Naokazu Yokoya, "Human action recognition-based video summarization for rgb-d personal sports video," in ICME. IEEE, 2016.

[12] Ying Zhao, Huijun Di, Jian Zhang, Yao Lu, and Feng Lv, "Recognizing human actions from low-resolution videos by region-based mixture models," in ICME. IEEE, 2016.

[13] Zhongjun Wu and Weihong Deng, "One-shot deep neural network for pose and illumination normalization face recognition," in ICME. IEEE, 2016.

[14] Binghui Chen and Weihong Deng, "Weakly-supervised deep self-learning for face recognition," in ICME. IEEE, 2016.

[15] Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal, "Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis," in AAAI. AAAI Press, 2014.

[16] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in EMNLP. Association for Computational Linguistics, 2005.

[17] Ellen Riloff and Janyce Wiebe, "Learning extraction patterns for subjective expressions," in EMNLP. Association for Computational Linguistics, 2003.

[18] Lakshmish Kaushik, Abhijeet Sangwan, and John HL Hansen, "Sentiment extraction from natural audio streams," in ICASSP. IEEE, 2013.

[19] Boya Wu, Jia Jia, Tao He, Juan Du, Xiaoyuan Yi, and Yishuang Ning, "Inferring usersemotions for human-mobile voice dialogue applications,".

[20] Paul Ekman and Wallace V Friesen, "Facial action coding system," 1977.

[21] Ming Sun, Jufeng Yang, Kai Wang, and Hui Shen, "Discovering affective regions in deep convolutional neural networks for visual sentiment prediction," in ICME. IEEE, 2016.

[22] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency, "Utterance-level multimodal sentiment analysis.," in ACL, 2013.

[23] Luca Casaburi, Francesco Colace, Massimo De Santo, and Luca Greco, "magic mirror in my hand, what is the sentiment in the lens?: An action unit based approach for mining sentiments from multimedia contents," Journal of Visual Languages & Computing.

[24] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," Neurocomputing.

[25] Soujanya Poria, Erik Cambria, and Alexander Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in EMNLP, 2015, pp. 2539-2544.

[26] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.

[27] Haohan Wang and Bhiksha Raj, "On the origin of deep learning,"arXiv preprint arXiv:1702.07800, 2017.

[28] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[29] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in International conference on Multimedia. ACM, 2010.

[30] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency, "3d constrained local model for rigid and nonrigid facial tracking," in CVPR. IEEE, 2012, pp. 2610- 2617.