《Are Multimodal Transformers Robust to Missing Modality?》

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, Xi Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18177-18186

论文链接：
https://openaccess.thecvf.com/content/CVPR2022/html/Ma_Are_Multimodal_Transformers_Robust_to_Missing_Modality_CVPR_2022_paper.html

摘要

由于缺失模态，从现实世界收集的多模态数据通常不完美。因此，对模态不完整数据具有鲁棒性的多模态模型是非常优选的。最近，transformer模型在处理多模态数据方面取得了巨大成功。然而，现有工作局限于架构设计或预训练策略；很少有人研究transformer模型是否对缺失模态的数据具有天然的鲁棒性。在本文中，我们首次开展了此类工作，全面研究了存在模态不完全数据时transformer的行为。不出所料，我们发现transformer模型对缺失模态敏感，而不同的模态融合策略将显著影响鲁棒性。让我们惊讶的是，即使对于相同的transformer模型，最佳融合策略也与数据集有关；不存在在一般情况下有效的通用策略。基于这些发现，我们提出了一种原理方法，通过自动搜索输入数据的最优融合策略来提高变压器模型的鲁棒性。在三个基准上的实验验证支持了所提出方法的优越性能。

1. 引言

由于隐私或安全限制，模态的完整性在现实世界中可能并不总是成立的。因此，transformer模型对于缺失模态的数据具有鲁棒性是很重要的，即模型的性能不会发生很显著的退化。

作者根据对实验结果的观察提出了两个问题：

1. Are Transformer models robust against missing-modal data? （表1）
2. Will the fusion strategy affect Transformer robustness against modal-incomplete data?（表2）

Table 1. Evaluation of the Transformer robustness against missing-modal data on MM-IMDb, UPMC Food-101, and Hateful Memes. We use ViLT [18] as the backbone. Note that the multimodal performance is *even worse* than the unimodal one, when modality is missing severely (results are highlighted in shaded gray). *The reported evaluation scores are F1-Macro (MM-IMDb), Accuracy (UPMC Food-101), and AUROC (Hateful Memes). Higher scores indicate better results

Dataset	Training		Testing		Evaluation*	Δ ↓
	Image	Text	Image	Text		
MM-IMDb [2]	100%	100%	100%	100%	55.3	0%
	100%	100%	100%	30%	31.2	**43.6%**
	100%	0%	100%	0%	35.0	36.7%
UPMC Food-101 [43]	100%	100%	100%	100%	91.9	0%
	100%	100%	100%	30%	65.9	**28.3%**
	100%	0%	100%	0%	71.5	22.2%
Hateful Memes [17]	100%	100%	100%	100%	70.2	0%
	100%	100%	100%	30%	60.2	14.2%
	100%	0%	100%	0%	56.3	**19.8%**

Table 2. Evaluation of the Transformer models under different fusion strategies on MM-IMDb and Hateful Memes. *Early* fusion refers to fusion at the first layer; *Late* fusion refers to fusion at the last layer. Different fusion strategies affect model robustness against the missing-modal data.

Dataset	Train		Test		Fusion Strategy	
	Image	Text	Image	Text	Early	Late
MM-IMDb	100%	100%	100%	100%	**55.3**	54.9
UPMC Food-101	100%	100%	100%	100%	**91.9**	91.8
Hateful Memes	100%	100%	100%	100%	**70.2**	64.5
MM-IMDb	100%	100%	100%	30%	**31.2**	31.0
UPMC Food-101	100%	100%	100%	30%	65.9	**69.1**
Hateful Memes	100%	100%	100%	30%	**60.2**	57.8

从表1中可以看出在面对缺失模态的数据时transformer模型的性能退化的很显著，特别是在MM-IMDb和UPMC Food-101数据集上，缺失30%的text模态的结果甚至比单模态（仅image）的结果还要差。

从表2中可以看出不同的融合策略对模型鲁棒性的影响很大，并且最优的融合策略与数据集有关，并不存在一种通用的策略来应对缺失模态的情况。从表中可以看出，在缺失模态情况下，对于MM-IMDb和Hateful Memes数据集来说early fusion是最优融合策略，而对于UPMC Food-101数据集来说，late fusion是最优融合策略。

基于对上面实验的观察总结，本文提出通过自动获得针对不同数据集的最佳融合策略来提高transformer的鲁棒性。具体来说，本文的想法是通过多任务优化，结合模态完整数据和模态不完

同数据集的最优融合策略。

总的来说，本文的贡献主要有以下几点：

1. 这篇文章是第一个探索transformer对模态不完整数据的鲁棒性的研究。
2. 本文观察到transformer模型在处理缺失模态数据时性能会显著退化。并且最优的融合策略与数据集有关，并不存在一种通用的策略来应对缺失模态的情况。
3. 本文通过多任务优化来提高transformer模型的鲁棒性。为了进一步提高鲁棒性，本文应用一种可微的算法来得到最优融合策略。
4. 在MM-IMDb、UPMC Food-101和Hateful Memes数据集上进行了广泛的实验和消融实验来支持上述的发现并验证本文提出的方法可以对抗缺失模态的情况。

2. 相关工作

多模态学习

不同的形式，例如自然语言、视觉信号或声音信号，通常在内容上是完整的，而在共同的概念上是重叠的。多模态学习旨在利用每种模态的完整信息来提高各种计算机视觉任务的性能。**多模态学习的一个关键方面是探索多模态融合的有效方法。** 本文指出有下列几种融合方式：

名称	文章
concatenation	《Convolutional mkl based multimodal emotion recognition and sentiment analysis》、《Select-additive learning: Improving generalization in multimodal sentiment analysis》
tensor fusion	《Tensor fusion network for multimodal sentiment analysis》
low-rank fusion	《Efficient low-rank multimodal fusion with modality-specific factors》

这些融合方式都非常依赖于模态的完整性，这就导致多模态融合不可能处理缺失模态的数据。因此，**多模态学习的另一个重要方向就是建立一个对缺失模态有很好鲁棒性的模型。**但是现有的模型通常对每一个模态都应用了一个特定于模态的子模型，如ResNet（for image） 和LSTM（for text），这可能导致更大的架构决策和训练参数集。

相反，本文采用transformer作为通用结构来联合建模每一个模态，这就能够简化设计并减少训练参数。

多模态transformer

transformer比传统的框架更灵活，训练负担更小。

传统的框架输出联合多模态表示通过利用concatenation、tensor fusion 或是其他融合机制来融合每一个模态的特征。然而这些明确的融合依赖于完整的模态，缺失任意模态都会破坏训练流程。相反，多模态transformer利用self-attention机制来生成一个所有模态的整体表示，这就允许任意模态的缺失。处理模态不完整样例时，他可以通过在注意力矩阵中应用mask来忽略这个缺失的模态。因此，多模态transformer可以灵活地处理缺失模态的数据。

因为传统的框架通常是由模态特定的子模型组成，这些子模型需要对每种模态独立训练，因此传统框架的训练负担会随着模态数量的增加而增加。相反，transformer模型可以用单一模型来同时处理所有模态，这大大减小了训练负担。

动态神经网络

这篇论文的工作还与动态神经网络有关，动态神经网络使网络结构适应不同的输入，从而显著提高了准确性、计算效率或灵活性。已经提出的很多方法都通过动态选择参考层来降低计算成本。受到AdaShare [1]的启发，它专注于学习一种在多任务学习中选择共享层的策略，主要想法是利用Gumbel Softmax Sampling在不依赖强化学习和额外的策略网络的情况下，去学习策略和网络参数。然而，将Gumbel Softmax Sampling直接应用于本文的问题会导致搜索空间很大，并且有许多无效策略。因此，本文开发了一种不使用Gumbel Softmax采样的有效方法。

3. 对多模态transformer的分析

背景

本文利用的是vision-language Transformer（ViLT）来对数据进行预处理。

输入文本通过单词嵌入码本和位置嵌入码本映射到单词嵌入中。输入图像首先被分割成块，然后被展平成矢量。利用线性投影与位置嵌入将这些矢量转换到潜在嵌入。最后，图像和文本嵌入与其对应的模态类型嵌入相结合。最后的多模式输入序列是视觉和文本嵌入的串联。

针对缺失模态的鲁棒性

Question: Are Transformer models robust again

Observation: Unsurprisingly, Transformer models degrade dramatically with modal-incomplete data.

评价鲁棒性的标准：

- 设置："full" test with modal-complete data，"missing" test with modal-incomplete data
- 评价：比较在两种设置下模型的性能差，相差越小，鲁棒性越好。

Table 3. Multi-label classification scores (%) on the *MM-IMDb* [2] under different settings: train and test with full modality (100% Image + 100% Text); train and test with single modality (100% Image or 100% Text). † indicate our implementation.

Method	Modality Image	Modality Text	F1 Micro	F1 Macro	F1 Weighted	F1 Samples
MFAS [30]	✓		47.8	25.6	42.1	48.4
		✓	60.2	48.9	58.5	60.6
CentralNet [41]	✓			33.5	49.2	
		✓		45.9	57.5	
ViLT [18]†	✓		51.8	35.0	48.0	51.1
		✓	63.3	52.5	62.0	62.9
MFAS [30]	✓	✓		55.7	62.5	
CentralNet [41]	✓	✓	63.9	56.1	63.1	63.9
ViLT [18]†	✓	✓	64.7	55.3	64.4	64.6

Table 4. Classification accuracy (%) on the *UPMC Food-101* [43]. † indicate our implementation.

Method	Modality Image	Modality Text	Accuracy
BERT+LSTM [9]	✓		71.7
		✓	84.4
ViLT [18]†	✓		71.5
		✓	84.4
BERT+LSTM [9]	✓	✓	92.5
MMBT [15]	✓	✓	92.1
ViLT [18]†	✓	✓	92.0

Table 5. AUROC (%) on the unseen test set of *Hateful Memes* [17]. *denotes the results from hateful memes challenge [16]. † indicate our implementation.

Method	Modality		AUROC
	Image	Text	
Unimodal*	✓		54.6
		✓	62.7
ViLT [18]†	✓		56.3
		✓	58.3
MMBT-Grid [15]*	✓	✓	67.3
MMBT-Region [15]*	✓	✓	72.2
ViLBERT [26]*	✓	✓	73.4
ViLBERT CC [26]*	✓	✓	72.8
Visual BERT [24]*	✓	✓	73.2
ViLT [18]†	✓	✓	70.2

通过表3、4、5的结果观测到了模态的重要性有所不同。在MM-IMDb和UPMC Food-101数据集上text占主导地位，而在Hateful Memes数据集上，两种模态的地位相当。

Table 6. Evaluation on the overfitting issue of Transformer models on MM-IMDb and Hateful Memes. Transformer models tend to overfit to dominate modality.

Dataset	Training		Testing		Evaluation
	Image	Text	Image	Text	
MM-IMDb	100%	100%	100%	100%	55.3
	100%	100%	0%	100%	47.4
	100%	100%	100%	0%	35.0
Hatful Memes	100%	100%	100%	100%	70.2
	100%	100%	0%	100%	55.7
	100%	100%	100%	0%	54.9

根据表6中的结果可以看出，Transformer模型倾向于过度拟合主导模态。具体来说，首先用多模态数据训练模型，并用不同的单峰数据进行测试。然后检查单峰和多峰测试之间的性能差距——差距越大，过度拟合越严重。实验结果如表6所示。如图所示，对于MM-IMDb数据集，纯文本测试比纯图像测试性能更好，这意味着纯文本测试更接近于全模态测试。因此，纯文本测试比纯图像测试有更小的差距，这表明在此数据集上训练的模型往往过度适合文本模态。

最优融合策略

Question: Will the fusion strategy affect Transformer robustness against modal-incomplete data?

Observation: Different fusion strategies do affect the robustness of Transformer models. Surprisingly, the optimal fusion strategy is dataset-dependent; there does not exist a universal strategy that works in general cases.

给每一层设置一个策略参数α，融合策略从策略参数中采样得到。

4. 鲁棒的多模态transformer

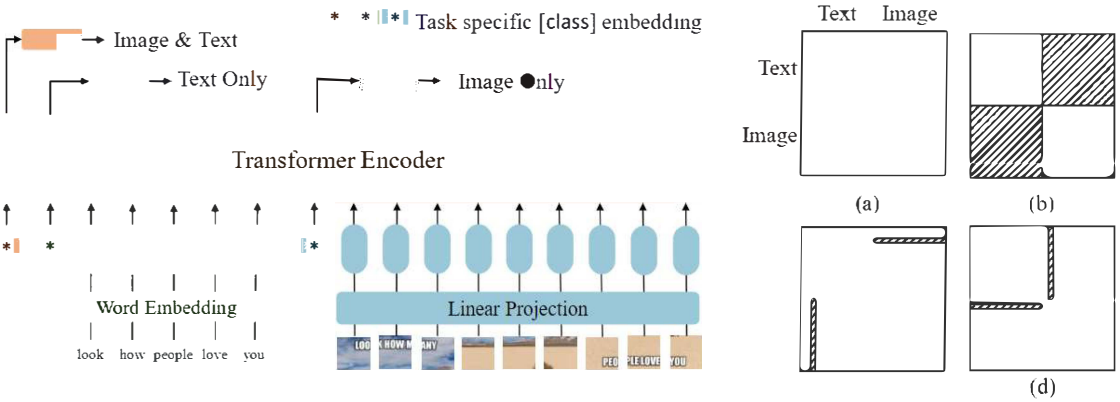

Figure 1. *Left*: Overview of our model. *Right*: Attention masks for different tasks: (a) Original attention without masking; (b) Mask-out cross-modal attention; (c) Mask-out image attention for text only [class] token; (d) Mask-out text attention for image only [class] token.

多任务学习

考虑三个不同的任务：

1. full-modal（image+text）task
2. image-only task
3. text-only task

并通过下列损失函数来优化transformer：

$$\mathcal{L} = \lambda_1 \mathcal{L}_{img}(\boldsymbol{x}^1; \boldsymbol{\theta}) + \lambda_2 \mathcal{L}_{txt}(\boldsymbol{x}^2; \boldsymbol{\theta}) + \lambda_3 \mathcal{L}_{it}(\boldsymbol{x}^1, \boldsymbol{x}^2; \boldsymbol{\theta})$$

transformer模型利用分类token来生成用于分类的嵌入。对于这三个任务，我们将三个分类token添加到tansformer模型中。每个分类token将为目标任务输出特定于任务的嵌入。模型概述如图1左侧所示。对于多任务学习，每个任务只能使用相应的模态进行分类，例如纯文本任务的文本模态。因此，我们在注意力矩阵上应用mask，确保每个分类token的输出嵌入仅包含来自相应模态的信息。例如，在纯文本任务中，我们掩盖了对图像的所有自我关注以及图像和文本之间的交叉关注。attention mask如图1右侧所示。

查找最优融合策略

将该查找看做一个两阶段优化问题：

$$\begin{aligned} \min \quad & \mathcal{L}^{val}(\boldsymbol{\theta}^*, \boldsymbol{\alpha}), \\ s.t. \quad & \boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \mathcal{L}^{tr}(\boldsymbol{\theta}, \boldsymbol{\alpha}^*). \end{aligned}$$

先在测试阶段中最小化$\mathcal{L}^{tr}$来得到最优的θ^*，在根据这个θ^*在验证阶段最小化$\mathcal{L}^{val}$来得到最优的α^*，再从中采样得到最优融合策略。上述过程伪代码如下：

Algorithm 1: Search for Optimal Fusion Policy.

Input: Multimodal dataset D^{tr}, D^{val}: inner-level learning rate γ; outer-level learning rate β; initialized policy parameter α; number of iterations K.

while *not converged* **do**
  $\{\mathbf{x}_i^1, \mathbf{x}_i^2, y_i\} \sim D^{tr}; \{\mathbf{x}_j^1, \mathbf{x}_j^2, y_j\} \sim D^{val}$
  $\theta_0 \leftarrow \boldsymbol{\theta}$
  Lower-Level:
  for $k = 0$ to $K - 1$ **do**
    Sample policy s with α using Eqn. 3
    $\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k - \gamma \nabla_{\boldsymbol{\theta}_k} \mathcal{L}^{tr}(\mathbf{x}_i^1, \mathbf{x}_i^2, s; \boldsymbol{\theta})$
  end
  $\theta^* \leftarrow \theta_K$
  Upper-Level:
  Sample policy s with $\boldsymbol{\alpha}$ using Eqn. 3
  $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \beta \nabla_{\boldsymbol{\alpha}} \mathcal{L}^{val}(\mathbf{x}_j^1, \mathbf{x}_j^2, s; \boldsymbol{\theta}^*)$
end

5. 实验

分析了提出的策略在三个多模态数据集上的性能，并旨在回答以下问题：

1. Transformer模型与模态完整数据的关系是否良好?
2. 提出的方法是否提高了主干对缺失模态数据的鲁棒性?
3. 为什么不同的数据集偏好不同的多模式融合层?
4. 哪些因素影响我们方法的有效性?

数据集

MM-IMDb有两种模态：图像和文本。目标任务是使用图像、文本或两者来预测电影的类型。这项任务是多标签分类，因为每部电影可能有多种类型。该数据集包含25956个图像文本对和23个类。

UPMC Food-101是一个由文本和图像组成的分类数据集。在UPMC Food-101中，图像和文本对是有噪声的，因为所有图像都是在不受控制的环境中获得的。该数据集包含90704个图像文本对和101个类。

Hateful Memes是另一个具有挑战性的多模态数据集，专注于识别meme中的hate speech。它被构造为依赖于单模态和多模态模型的失败模型可能表现良好：将具有挑战性的样本（“良性混杂因素”）添加到数据集中，以使依赖于单模态信号变得更加困难。仇恨模因正好包含10万个meme。

结果分析

1. 前面的分析回答了第一个问题。
2. 如下图所示，本文提出的方法在面对缺失模态的情况，性能明显优于baseline，特别是在$text$（$\eta = 10\%$）的时候，baseline的多模态融合性能甚至低于单模态性能。

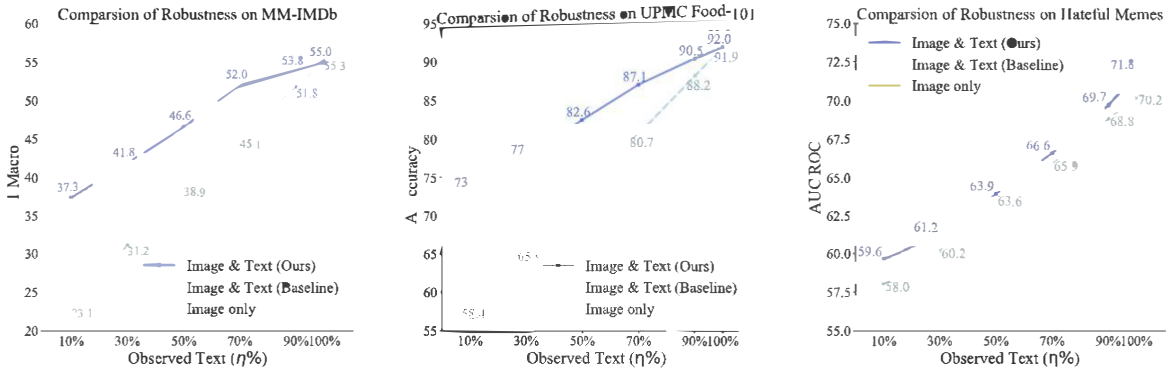

Figure 2. Comparison of Transformer robustness [18] on MM-IMDb [2] (*left*), UPMC Food-101 [43] (*middle*), and Hateful Memes dataset [17] (*right*). We adopt ViLT [18] as the backbone. Models are trained with 100% text + 100% image and tested with η% text + 100% image. "Image only" refers to the single modality setting – only image modality is used for training and testing. *Our method significantly improves model robustness, especially when the modality is severely missing*

3. 结果表明，late fusion在MM-IMDb数据集上更受青睐，而early fusion在Hateful Memes数据集上更受欢迎。学习的策略与每个数据集的特征一致。融合层的深度影响着transformer建模跨模态关系的能力。融合层越深，容量越低。在MM-IMDb上，主要模态文本（情节描述）提供了比图像模态（海报）更多的电影生成细节。模型采用late fusion策略是合理的，因为预测任务可以很容易地利用主导模态进行处理，并且建模跨模态关系只会带来边际收益。相反，Hateful Memes数据集被构造为一个失败模型，它通过向数据集添加具有挑战性的样本（"良性混杂因素"）来依赖于单一模态。因此，为了处理这个数据集，模型应该有足够的能力来建模跨模态关系。对于一个依赖于两种模式进行准确预测的数据集，学习early fusion策略是合理的。

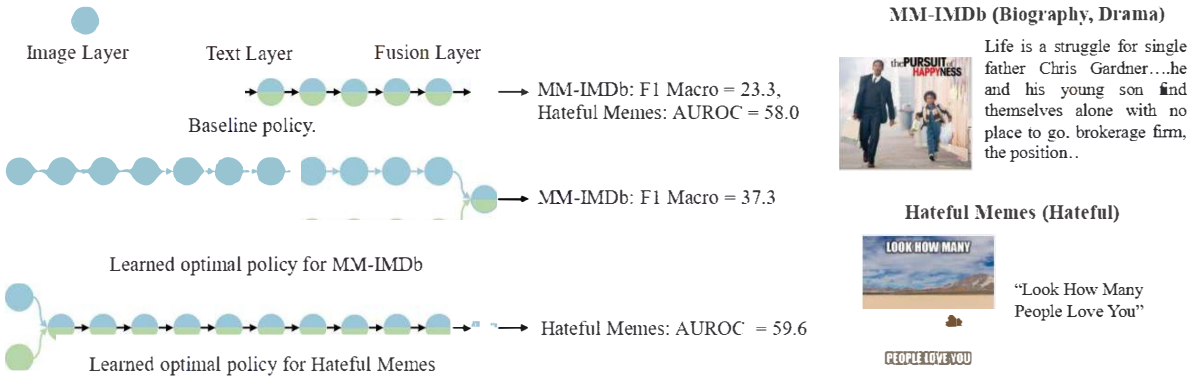

Figure 3. *Left*: Visualization of the learned policy. *Right*: Example sample from MM-IMDb and Hateful Memes. *Late fusion yields the best robustness on MM-IMDb, while early fusion leads to the most robust model on Hateful Memes.* The reported results were obtained using the following settings: training with 100% Image + 100% Text, testing with 100% Image + 10% Text.

4. 表7：把将缺失模态应用到训练中的策略看做一个新基线，进行比较。结果表明，这种简单的方法将不再起作用，其结果比单模态性能还差。

Table 7. Results of the new baseline: training and testing with 100% image + 30% text.

Method	MM-IMDb	Food-101	Hateful Memes
Image only	31.2	65.9	60.2
w baseline	40.4	44.3	59.7
Ours	**46.6**	**77.5**	**61.2**

表8：分析多任务学习和最优融合层。结果如表8所示。两个组件都提高了变压器的鲁棒性。此外，可以看到多任务学习比融合策略贡献更大。具体来说，当测试中只有10%的文本可用时，多任务学习比最优融合策略优30%。

able 8. Ablation study on multi-task learning and optimal fusion on MM-IMDb.

Method		Training		Testing		F1 Macro
Multi-task	Opt. Policy	Image	Text	Image	Text	
✓		100%	100%	100%	30%	31.2
	✓	100%	100%	100%	30%	28.6
✓	✓	100%	100%	100%	30%	**41.8**
✓		100%	100%	100%	10%	22.6
	✓	100%	100%	100%	10%	17.3
✓	✓	100%	100%	100%	10%	**37.3**

表9：attention mask分析。在本文的方法中，在注意力矩阵上应用掩mask，以强制分类token仅利用来自相应模态的信息。研究attention mask的效果如表9所示。可以看出重要的是确保每个分类token不会偷看其他模式的信息。

Table 9. Ablation study on the effect of attention mask for multi-task learning on MM-IMDb.

Method	Training		Testing		F1 Macro
	Image	Text	Image	Text	
without masking	100%	100%	100%	10%	23.0
with masking	100%	100%	100%	10%	37.3