Multimodal fusion via cortical network inspired losses

Shiv Shankar
University of Massachusetts
sshankar@cics.umass.edu

Abstract

Information integration from different modalities is an active area of research. Human beings and, in general, biological neural systems are quite adept at using a multitude of signals from different sensory perceptive fields to interact with the environment and each other. Recent work in deep fusion models via neural networks has led to substantial improvements over unimodal approaches in areas like speech recognition, emotion recognition and analysis, captioning and image description. However, such research has mostly focused on architectural changes allowing for fusion of different modalities while keeping the model complexity manageable. Inspired by neuroscientific ideas about multisensory integration and processing, we investigate the effect of introducing neural dependencies in the loss functions. Experiments on multimodal sentiment analysis tasks with different models show that our approach provides a consistent performance boost.

1 Introduction

Human beings perceive the world as a unified whole, not in individual sensory modalities. While traditionally different sensory models have been studied in isolation, it has been well recognized that perception operates via integration of information from multiple sensory modalities.

Research in multimodal fusion aims to achieve a similar goal in artificial models: extract and integrate all information from different input modalities. For example, if someone is sarcastic, the facial expression and voice intonation provide information not directly decipherable from the uttered words. If a model only looks at the text of the interaction, then it is unlikely to classify this interaction currently. Current research in deep multimodal fusion primarily deals with architectural improvements to create complex feature-rich, yet efficient representations (Zadeh et al., 2017; Liu et al., 2018; Hazarika et al., 2020). The hope is that more complex models will be able to integrate the complementary information from different unimodal representations into a unified common representation. Learning such unified representations, however, is a challenging task. Different modalities can present the same information in radically different ways with emphasis on different aspects of the content. These heterogeneities across different modalities mean that learning multimodal representations must deal with feature shifts, distributional effects, nuisance variation and a variety of related challenges (Baltrušaitis et al., 2018).

Inspiring from work in multisensory neural processing, we define a loss regularizer that we call *synergy* to train these models. Synergy has a specific meaning in information-theoretic literature (Cover, 1999). The synergy between random variables X and Y refers to the unique mutual information that X provides about Y. While our loss function is not the same as information theoretic synergy, the intuition behind our proposed loss is the same as actual synergy; to try to maximize dependencies between the representations. As our method uses neural networks or kernel-based methods to capture distributional divergences, we expect that this method will allow our model to capture complex dependencies which cannot be captured via techniques like subspace alignment.

We test our proposed training loss on different multimodal fusion architectures including LFN(Zadeh et al., 2017), MFN (Zadeh et al., 2018a), MAGBERT(Rahman et al., 2020) and MIM (Han et al., 2021). Our experiments show that training with synergy maximization improves the result by a significant margin.

2 Preliminaries

In this section, we give an overview of the basic ideas relevant to this work; primarily mutual information, and existing work on deep multimodal fusion and neural synergy.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1167 - 1178

2.1 Multimodal Fusion

The problem in the most abstract terms is a supervised learning problem. We are provided with a dataset of N observations $\mathcal{D} = (x_i, y_i)_{i=1}^N$. All x_i come from a space $\mathcal{X}$ and y_i from $\mathcal{Y}$. We are provided a loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ which is the task loss. Our goal is to learn a model $\mathcal{F}_\theta : \mathcal{X} \to \mathcal{Y}$ such that the total loss $\mathcal{L} = \sum_i L(\mathcal{F}(x_i), y_i)$ is minimized. In multimodal fusion the space of inputs $\mathcal{X}$ naturally decomposes into K different modalities $\mathcal{X} = \prod_{j=1}^K \mathcal{X}_j$. We use X_j to represent random variables which form the individual modality specific components of the input random variable X.

A common way to learn such a multimodal function is to decompose it into two components: a) an embedding component E which fuses information into a high dimensional vector in $\mathbb{R}^d$ and b) a predictive component P which maps vector from $\mathbb{R}^d$ to $\mathcal{Y}$. Furthermore since the different modalities are often no directly compatible with each other (for eg text and image), E itself is decomposed into a) modality specific readers $F_i\mathcal{X}_i \to \mathbb{R}^{d_i}$ which are specifically designed for each individual modality $\mathcal{X}_i$ and b) a fusion component $F : \prod_i \mathbb{R}^{d_i} \to \mathbb{R}^d$ which fuses information from eah individual modality embedding. F is provided with uni-modal representations of the inputs $X_i = (X_1, X_2, \ldots X_K)$ obtained through embedding networks f_i. F has to retain both unimodal dependencies (i.e relations between features that span only one modality) and multi-modal dependency (i.e relationships between features across multiple modalities).

This decomposition has two advantages a) the individual modality reader can be pre-trained on the task at hand or even from a larger dataset (for example BERT (Devlin et al., 2018) for language, Resnet (He et al., 2016) for images) which allows us to leverage wider modality specific information and b) often but not always each individual modality is in principle enough to correctly predict the output

2.2 Distributional Divergences

Divergence is a functional which characterizes the distance or "discrepancy" between two probability distributions on the same space. Divergence however is a different notion than distance because divergences are not necessarily symmetric. A common measure of discrepancy between two distributions is the Kullback-Liebler divergence (KL divergence) (Cover, 1999). The KL divergence of the density p relative to the density q is given by

$$d(p; q) = \mathbb{E}_{x\sim p}\left[\log \frac{p(x)}{q(x)}\right]$$

This divergence is often also used implicitly for estimating dependence between two random variables. Mutual information (MI) is a measures of dependence between two random variable X and Y capable of incorporating multiple types of relationships between them. If we have variables X and Y, then the mutual information between them is given by

$$I(X; Y) = KL\left[p_{XY}(x, y) \| p_X(x)p_Y(y)\right]$$

where p_{XY} is the joint probability density of the pair (X, Y), and p_X, p_Y are the marginal probability densities of X, Y respectively.

Estimation of Divergence Estimating entropic differences between two distributions purely from their samples is a difficult task (Kinney and Atwal, 2014). As such there have been multiple types of divergences proposed over the years (Gretton et al., 2005; Studenỳ and Vejnarová, 1998). Moreover in recent years, several estimators have been proposed for entropic divergences based on variational methods (Belghazi et al., 2018; Hjelm et al., 2018; Amjad and Geiger, 2019). These estimators use flexible neural networks as a contrast function and optimize a variational bound. We describe two such methods which are used in our experiments

- **Neural Mutual Information** (Belghazi et al., 2018) is a variational method to estimate the KL divergence between two distributions. It is estimated via gradient ascent on the Donsker-Varadhan bound (Donsker and Varadhan, 1985). The Donsker Varadhan bound shows that:

$$KL(P, Q) \geq \sup_g \mathbb{E}_{X\sim P}[g(X)] - \mathbb{E}_{X\sim Q}[\exp^{g(X)}]$$

  The Young-Fenchel duality shows that the gap is zero; i.e. at the optima the right side of the above expression matches the KL divergence. Instead of a global maximization over all functions one can instead use a family of

functions parameterized via neural networks. The bound obtained thus is necessarily lower than the actual KL, but now one can use gradient descent to optimize the network.

- **Maximum Mean Discrepancy** or MMD (Gretton et al., 2012) is a kernel based estimator of divergence between distributions. Mathematically the MMD between two distributions P and Q is given by the norm of the difference of the mean embeddings of P and Q in the RKHS space of the chosen kernel. Further extensions to MMD have been developed based on neural networks which provide non-universal but more powerful kernel based tests (Liu et al., 2020).

$$\mathcal{MMD}(P, Q) = ||\mu_P^\phi - \mu_Q^\phi||$$

The above formula can be estimated purely via samples by using the Kernel matrix $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ where ϕ represents the corresponding RKHS embedding function The final monte carlo estimator is given by:

$$\begin{aligned}\mathcal{MMD}(P, Q) = &\sum_{p_i, p_j \sim P} K(p_i, p_j) \\ &+ \sum_{q_i, q_j \sim Q} K(q_i, q_j) \\ &- \sum_{p_i, q_j \sim P, Q} 2K(p_i, q_j)\end{aligned}$$

2.3 Kurtosis

Kurtosis is a statistical measure which is used to categorize the behavior of the distribution tails. It is more sensitive to rare events and hence is used for distributions with "fatter tails". For univariate variables, kurtosis is the standardized fourth moment i.e

$$\frac{\mathbb{E}[(X - \mu)^4]}{(\mathbb{E}[(X - \mu)^2])^2}$$

It is often used to measure deviations from normality. Mardia (1970) defined a measure of multivariate kurtosis as follows:

$$\mathbb{E}[((X - \mu])^T \Sigma^{-1} (X - \mu))^2]$$

where X is a $p \times 1$ dimensional random vector and μ, Σ are the mean and covariance matrix of X respectively. Multivariate cokurtosis between random variables is also sometimes used as a measure of dependence between them. It is one of the metrics used by Rosas et al. (2019); Barrett and Seth (2011) to analyze neural complexity and brain functional connectivity.

2.4 Other works on multimodal fusion

Earlier work on neural fusion models primarily relied on an early fusion of features. These approaches simply concatenated inputs of different modalities and used simple models to combine requisite information. Despite their simplicity, such models often perform well and are robust (Narayanan et al., 2019). More modern methods, however, deploy fancier methods to induce information aggregation. One set of models used gradient descent to try to force different feature networks to learn about each other and embed information jointly. This process can be enhanced by adding specific forms of regularization such as reconstruction loss (Mai et al., 2020), or auxiliary task loss (Chen et al., 2017; Yu et al., 2021). Another family of models uses linear algebra based methods to combine unimodal representations. Methods like those of Liu et al. (2018); Chen and Mitra (2018); Chachlakis et al. (2019) try to fuse information via tensor decomposition of high dimensional product tensors of individual unimodal representations. Other methods use subspace alignment (Lee et al., 2019; Yu et al., 2012) or correlation loss (Sun et al., 2020; Hazarika et al., 2020) to merge different representations. However, in some form or other, these models rely primarily on architectural changes. We, on the other hand, do not want to focus on such changes. Instead, our goal was to use insights from neuroscience to provide a methodology that can be deployed atop any standard multimodal fusion model.

3 Dependency Coding in Multisensory Processing

A common and vital feature of nervous systems is the integration of information arriving simultaneously from multiple sensory pathways. The underlying neural structures have been found to be related in both vertebrates and invertebrates. The classic understanding of this process is that different sensory modalities are processed individually and then combined in various multimodal convergence zones, including cortical and subcortical regions (Ghazanfar and Schroeder, 2006), as well as

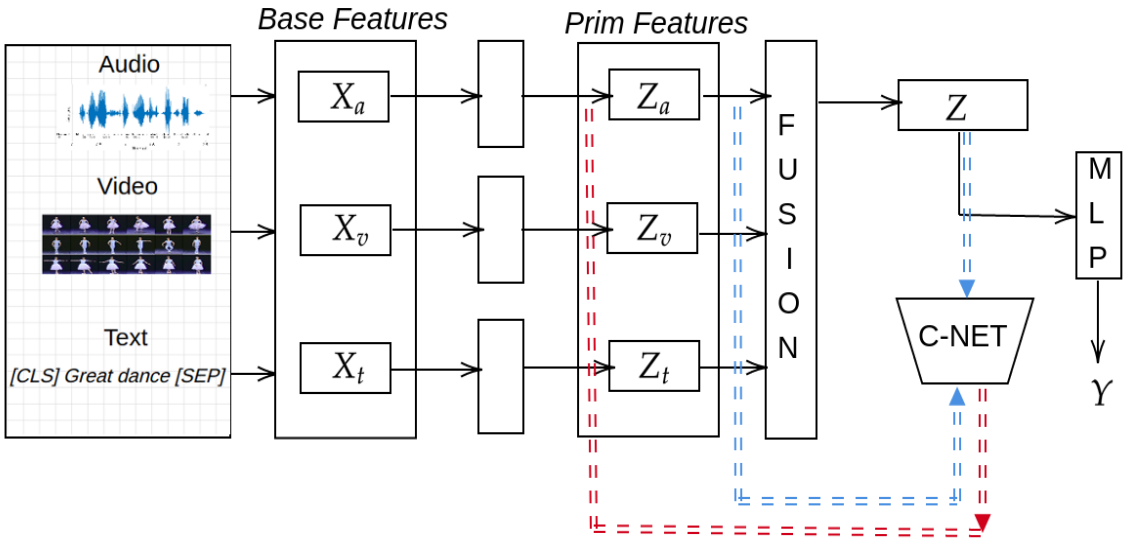

Figure 1: A general multimodal fusion Architecture. We depict in colour the additional components proposed in this work viz the proposed cortical network (C-Net) and its connection to individual layers

multimodal association areas (Rauschecker et al., 1995). Studies in the superior colliculus (Meredith et al., 1987) showed that multiple sensory modalities are processed in this brain stem region, with some neurons being exclusively unimodal and others being multimodal. Hypotheses of encoding of multimodal information include changes in neuronal firing rates (Pennartz, 2009) or a combinatorial code in population of neurons (Osborne et al., 2008; Rohe and Noppeney, 2016).

Evidence shows that while multimodal representations are distinct from unimodal ones, there is sufficient overlap between the set of neurons that process different sensory modalities. For example, Follmann et al. (2018) show that even in a simple crustacean organism, more than half the neurons in the commissural ganglion are multimodal. Moreover, they show that in 30% of these multimodal neurons, responses to one modality were predictive of responses to other modalities. *Both these facts suggest that the neural representations across different modalities have high information about each other.*

Studies of multisensory collicular neurons suggest that their crossmodal receptive fields (RF) often overlap (Spence et al., 2004). This pattern is also found in multisensory neurons present in other brain regions. As such, a spatiotemporal hypothesis of multisensory integration has been suggested: superadditive multimodal processing is observed when information from different modalities comes from spatiotemporally overlapping receptive fields (Recanzone, 2003; Wallace et al., 2004; Stanford et al., 2005). Since multimodal cortical neurons are generally downstream of modality-specific regions, the information about RF overlap is present in their input unimodal neural representations. Moreover, the sensory-specific nuclei of the thalamus have been shown to feed multisensory information to primary sensory specific-cortices (Kayser et al., 2008). *This suggests the existence of explicit feedback connection from the multimodal representations to unimodal representations.*

Cortical and subcortical networks often contain clusters of strongly connected neurons. Functionally the existence of such cliques imply highly integrated pyramidal cells that handle a disproportionately large amount of traffic (Harriger et al., 2012). In cortical circuits, around 20% of the neurons account for 80% of the information propagation (Nigam et al., 2016; Van Den Heuvel and Sporns, 2011). Timme et al. (2016); Faber et al. (2019) demonstrate that multimodal computation tends to concentrate in such local cortical clusters. They also found significantly lower kurtosis in such clusters and that dependence between oscillations was proportional to the amount of information flow. Sherrill et al. (2020) show that highly kurtotic neural activity positively related when multiple external stimuli are provided. Thus, kurtosis in neural firings is a representation of the dependence between inputs. *This suggests that when input kurtosis is high there is more significant cognitive processing and information flow required to extract relevant*

information.

4 Model

For our purposes we will limit ourselves to talk about tasks similar to the MOSI dataset. In this setting the input has three modalities viz audio (a), visual (v), and textual language (l). The fusion problem involves learning a representation $\mathcal{M}_f$ that combined the uni-modal representations of the inputs $X_{a,v,l} = (X_a, X_v, X_l)$.

4.1 Dependency Coding and C-Network

We modify the base neural architecture to incorporate the global structure explained in the last section. We propose a way to incorporate such changes without major architectural change into current baseline designs. The key component is the additional network (colored in red) in Figure 1 which we shall call as C-network. The C-network takes as input the individual unimodal representations and the fused representation and attempts to force a specific form of dependency as explained below.

C-Network The purpose of the C-Network is to try to enforce on the model the three primary characteristics of real neural circuits explained in the earlier section. We list them here and describe how we attempt to incorporate those characteristics in a more standard model.

- Individual uni-modal representations should be predictive of other uni-modal representations. We try to achieve this by simply predicting on modality representation by the combination of others. Q_i refers to a modality associated neural network which attempts to reconstruct the unimodal representation Z_i from the other representations Z_{-i}. The error between the two is penalized in the form of a reconstruction loss between modalities i.e. we add a penalty of the form:

$$\mathcal{L}_{L2} = ||Q_i(Z_{-i}) - Z_i||^2$$

- Multimodal representation should be feedback into input neurons to align and capture information between them. Providing feedback during inference time from the multimodal representation would be ideal. However this would make the overall prediction recurrent, something fundamentally different from most current architectures. Moreover given current high dimensional encoders; doing such processing would be extremely resource intensive. As such we aim to achieve this feedback by treating the multimodal representation and unimodal representation spaces as different domains and adding a loss of the form:

$$\mathcal{L}_d = d(p(g_i(Z_i)), p(g_i(Z)))$$

The purpose of the aforementioned loss is to align the distributions of the features in the same embedding space of the mapping from the multimodal and unimodal domains. d represents a measure that captures the discrepancy between the distributions, g_i refers to neural networks for projecting and aligning the combined representation Z with unimodal representations Z_i, and p denotes the empirical/sample distribution of the corresponding features. In our experiments, for d we use the MMD discrepancy (Gretton et al., 2012) and KL divergence as the metric; though other divergences can also be used. Note that this loss by itself can be minimized by forcing the g functions to ignore their inputs. We prevent this by first doing a random projection of the features[1] into a smaller dimensional vector space and then apply an invertible neural network. Such alignment losses have been used in works on domain adaptation (Motiian et al., 2017) under the name semantic loss or confusion loss. We refer the readers to Motiian et al. (2017); Li et al. (2019) for more details on semantic losses.

Note that instead of aligning the features via some kind of embedding based distributional distance, one could try to maximize mutual information between the embeddings as well. We experiment with one such model in our experiment and as the results show, found it to be slightly worse than using MMD based alignment loss.

- Individual unimodal and multimodal representations should have low kurtosis. To ensure this condition we estimate the multivariate kurtosis by plugging in standard estimators for the mean and covariates. The final kurtosis estimator used is given by:

[1] similar to Johnson Lindenstrauss projections (Landweber et al., 2016)

$$\kappa = \frac{1}{n}\sum_{i}^{n}[((z_i - \bar{z})^T S^{-1}(z_i - \bar{z}))^2]$$

where z_i here are samples from the Z features in the model (where Z can be unimodal features like Z_a or fused final feature Z). $\bar{z}$ refers to the empirical mean feature $\bar{z} = \frac{\sum_{i}^{n} z_i}{n}$ and S is the empirical covariance matrix $S = \frac{\sum_{i}^{n}(z_i - \bar{z})(z_i - \bar{z})^T}{n}$.

An important thing to note here is that high dimensional kurtosis values can be highly sensitive to outliers. As such we regularize the estimate by doing three things: a) We cap the max norm of the difference vectors during estimation. b) We scale up the diagonal of the covariance matrix to reduce its condition number c) Finally the covariance matrix itself is computed via a decaying moving average over a window of multiple batches to produce smoother estimates before the inversion operation.

During training we add the regularization penalties described earlier along with the usual maximum likelihood based objective. The different loss components are weighted with seperate hyperparameters. Note that the C-Network is purely a training time addition, and is not invoked during inference. Hence the additional network invoke zero additional time during testing. An algorithmic description of the full method is presented in the Appendix D

5 Experiments

5.1 Datasets

We empirically evaluate our methods on two commonly used datastes for multimodal training viz CMU-MOSI and CMU-MOSEI.

CMU-MOSI (Wöllmer et al., 2013) is sentiment prediction taks on a set of short youtube video clips. CMU-MOSEI (Zadeh et al., 2018b) is a similar dataset consisting of around 23k review videos taken from YouTube. The output in both cases is a sentiment score in $[-3, 3]$. For each dataset, three modalities are available; audio, visual frames, and language. Preliminary features on each modality is obtained as follows:

- Audio: Features are extracted from the sund recordings using the method of Degottex et al. (2014).
- Language: The video transcripts are converted to word embeddings using BERT (Devlin et al., 2018) or Glove (Pennington et al., 2014)
- Visual: Visual features are extracted using FACET (iMotion) which provides facial action units vectors.

5.2 Models

We run our experiments with the following architectures:

- FLSTM (Narayanan et al., 2019) is the baseline early fusion LSTM architecture used by Zadeh et al. (2017)
- Tensor Fusion Network or TFN (Zadeh et al., 2017) combined information via pooling of a high dimensional tensor representation of multimodal features. More specifically it does a multimodal Hadamard product of the aggregated features with RNN based language features.
- Memory Fusion Network or MFN (Zadeh et al., 2018a) incorporate gated memory-units to store multiview representations. It then performs an attention augmented readout over the memory units to combine information into a single representation.
- MAGBERT (Rahman et al., 2020) is a transformer based architecture that uses the Wang gate (Wang et al., 2019). The multimodal information is send to the multimodal gate to compute modified embeddings which are passed to a BERT (Devlin et al., 2018) based model. This model achieves state-of the-art results on multimodal sentiment benchmark MOSI (Wöllmer et al., 2013) and MOSEI (Zadeh et al., 2018c).
- MIM (Han et al., 2021) is a recent near SOTA architecture. It combined BERT based text embeddings with modality specific visual and acoustic LSTMs (Hazarika et al., 2020).
- Recently Colombo et al. (2021) conducted experiments introducing a information regularizer on existing architectures. The main differences between the our method and their

method are a) our method focuses on synergy terms whereas their proposal is optimizing joint mutual information between different unimodal representations; and b) they experiment with variational measures of information. We replicate our experiments with their best performing model and present the results with the label I_{Was}.

Split	CMU-MOSI	CMU-MOSEI
Train	1284	16326
Validation	229	1871
Test	686	4659
All	2199	22856

Table 1: Dataset summary

5.3 Evaluation

We report both the Mean Absolute Error (MAE) and the correlation of model predictions with true labels. In the literature, the regression task is also turned into a binary classification task for polarity prediction. We follow Rahman et al. (2020) Accuracy Acc_7 denotes accuracy on 7 classes and Acc_2 the binary accuracy) of our best performing models. We also report the Mean Absolute Error (MAE) and the correlation of model intensity predictions with true values.

5.4 Results

We present and discuss here the results obtained in our experiments. Results on MOSI are presented in Table 2 while Table 3 present results for MOSEI dataset. We trained each of the models with the standard cross entropy loss (labeled as NLL); and with cross entropy loss regularized with the synergy penalty discussed earlier. On both datasets, regularization via synergy leads to performance improvement. For example, a MFN on CMU-MOSI trained with MMD based synergy (NLL+S_{MMD}) outperforms by more than 4 points on Acc_7 than standard likelihood training. On CMU-MOSEI too the gains are significant when trained with synergy regularization. In general training via MMD synergy tends to be better than via KL synergy. This might be the inherent behavior of the MMD dependency which is always well defined; or it might reflect the hardness of information estimation. For example it is well known that good bounds on standard mutual information are difficult to obtain (Kinney and Atwal, 2014); while MMD estimator are asymptotically consistent (Gretton et al., 2012)

	Acc_7	Acc_2	MAE	CORR
	FLSTM			
NLL	31.2	75.9	1.01	0.64
NLL+S_{KL}	31.6	76.3	1.01	0.66
NLL+S_{MMD}	33.6	76.4	0.98	0.66
	MFN			
NLL	31.3	76.6	1.01	0.62
NLL+S_{KL}	32.5	76.6	0.94	0.65
NLL+S_{MMD}	35.9	77.4	0.95	0.66
NLL+I_{Was}	35.1	77.1	0.97	0.63
	LFN			
NLL	31.9	76.9	1.01	0.64
NLL+S_{KL}	32.6	77.6	0.97	0.64
NLL+S_{MMD}	35.4	77.9	0.97	0.67
NLL+I_{Was}	32.4	77.6	0.97	0.64
	MAGBERT			
NLL	40.2	83.7	0.79	0.80
NLL+S_{KL}	41.9	84.1	0.76	0.82
NLL+S_{MMD}	41.9	85.6	0.76	0.82
NLL+I_{Was}	41.8	84.2	0.76	0.82
	MIM			
NLL	46.3	83.7	0.77	0.76
NLL+S_{KL}	46.4	83.7	0.74	0.75
NLL+S_{MMD}	46.7	84.2	0.72	0.79
NLL+I_{Was}	46.6	84.2	0.75	0.79

Table 2: Results on sentiment analysis on CMU-MOSI. Acc_7 denotes accuracy on 7 classes and Acc_2 the binary accuracy. MAE denotes the Mean Absolute Error and Corr is the Pearson correlation

5.5 Modality Dropout

Zadeh et al. (2018a); Rahman et al. (2020) have demonstrated that while multimodal fusion does improve performance, the primary modality continues to be textual data. Hence in this experiment, we want to assess the effect of corruptions of text modality in our model. Following Colombo et al. (2021) we experiment with dropping the text modality either by itself (T) or with one of the other modalities (T+V or T+A). The results are presented in Table 4

Since the C-Networks forces a reconstruction and distributional divergence loss between the unimodal and multimodal representations, one would expect that models trained using our approach would be more resistant to modality errors. This is borne out in the experiments, where we see that

	Acc_7	Acc_2	MAE	CORR
	FLSTM			
NLL	44.1	75.1	0.72	0.52
NLL+S$_{KL}$	44.4	75.6	0.70	0.52
NLL+S$_{MMD}$	45.3	76.0	0.68	0.54
	MFN			
NLL	44.3	74.7	0.72	0.52
NLL+S$_{KL}$	44.3	74.8	0.72	0.56
NLL+S$_{MMD}$	46.2	75.1	0.69	0.56
NLL+I$_{Was}$	45.1	75.2	0.72	0.54
	LFN			
NLL	45.2	74.3	0.70	0.54
NLL+S$_{KL}$	46.1	75.3	0.69	0.56
NLL+S$_{MMD}$	46.3	75.3	0.67	0.56
NLL+I$_{Was}$	45.9	75.1	0.69	0.55
	MAGBERT			
NLL	46.9	83.9	0.59	0.77
NLL+S$_{KL}$	47.4	85.3	0.59	0.79
NLL+S$_{MMD}$	47.9	85.4	0.59	0.79
NLL+I$_{Was}$	47.2	85.0	0.59	0.78
	MIM			
NLL	53.3	79.6	0.54	0.75
NLL+S$_{KL}$	53.5	80.3	0.54	0.77
NLL+S$_{MMD}$	54.3	82.4	0.52	0.77
NLL+I$_{Was}$	53.5	82.1	0.53	0.77

Table 3: Results on sentiment analysis on CMU-MOSEI. Acc_7 denotes accuracy on 7 classes and Acc_2 the binary accuracy. MAE denotes the Mean Absolute Error and Corr is the Pearson correlation

training with synergy based loss has better performance than training with simple max-likelihood.

Note that the C-network itself is not active at test time; instead this effect is due to the alignment forced by the network during training. An interesting future direction would be to explicitly use the C-network outputs to ameliorate modality corruption.

Drop Modality	None	T	T+V	T+A
NLL	83.7	36.4	35.1	34.4
NLL+S$_{MMD}$	85.6	48.3	46.7	45.9
NLL+S$_{KL}$	84.1	46.8	45.9	45.5

Table 4: Modality corruptions results on sentiment analysis on CMU-MOSI. The results are the binary accuracies Acc_2^{corrupt}

5.6 Ablation Study

Our overall proposal has multiple components viz a) the reconstruction loss (also called $\mathcal{L}_{L2}$ loss); b) the distribution alignment loss (which we call $\mathcal{L}_d$ Loss); and c) the kurtosis loss $\mathcal{L}_\kappa$. As such we ran experiments to assess the importance of each component. Specifically we trained the model without each of the three loss components prescribed in our method, and assessed the test performance. The results are presented in Appendix A.

First we note the performance improvement by incorporating kurtosis in the regularization which shows the efficacy of this term. Second one can also note that removing any individual component leads to reduction in performance, suggesting all components act together in a synergistic way to improve the results.

6 Concluding Remarks

In this paper, we used the idea of regularizing via a term which we label neural synergy maximization. This regularizer is inspired by neural cicruit design in the vertebral cortex. We experimented with different measures of synergy based on discrepancy measures such as KL and MMD. We also show that training with synergy can produce benefit on even SOTA architectures.

Limitations The most prominent limitation of this approach, is that it is inherently limited by the architecture with which it is being used. While our additional loss did improve performance, one can observe that the final performance is dependent on the initial performance. For example, while we tested on four architectures, the final performance of each model was in the same range as the initial performance. An entirely different architecture can possibly improve over our results. On the other hand our approach is model agnostic and applicable on any model trained only via max-likelihood.

References

Rana Ali Amjad and Bernhard C Geiger. 2019. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2225–2239.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Adam B Barrett and Anil K Seth. 2011. Practical measures of integrated information for time-series data. *PLoS computational biology*, 7(1):e1001052.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. 2018. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*.

Dimitris G Chachlakis, Mayur Dhanaraj, Ashley Prater-Bennette, and Panos P Markopoulos. 2019. Options for multimodal classification based on l1-tucker decomposition. In *Big Data: Learning, Analytics, and Applications*, volume 10989, page 109890O. International Society for Optics and Photonics.

Junting Chen and Urbashi Mitra. 2018. A tensor decomposition technique for source localization from multimodal data. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4074–4078. IEEE.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 163–171.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloe Clavel. 2021. Improving multimodal fusion via mutual dependency maximisation.

Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

MD Donsker and SRS Varadhan. 1985. Large deviations for stationary gaussian processes. *Communications in Mathematical Physics*, 97(1):187–210.

Samantha P Faber, Nicholas M Timme, John M Beggs, and Ehren L Newman. 2019. Computation is concentrated in rich clubs of local cortical networks. *Network Neuroscience*, 3(2):384–404.

Rosangela Follmann, Christopher John Goldsmith, and Wolfgang Stein. 2018. Multimodal sensory information is represented by a combinatorial code in a sensorimotor system. *PLoS biology*, 16(10):e2004527.

Asif A Ghazanfar and Charles E Schroeder. 2006. Is neocortex essentially multisensory? *Trends in cognitive sciences*, 10(6):278–285.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(null):723–773.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Logan Harriger, Martijn P Van Den Heuvel, and Olaf Sporns. 2012. Rich club organization of macaque cerebral cortex and its role in network communication.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.

Christoph Kayser, Christopher I Petkov, and Nikos K Logothetis. 2008. Visual modulation of neurons in auditory cortex. *Cerebral Cortex*, 18(7):1560–1574.

Justin B Kinney and Gurinder S Atwal. 2014. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359.

Peter S. Landweber, Emanuel A. Lazar, and Neel Patel. 2016. On fiber diameters of continuous maps. *The American Mathematical Monthly*, 123(4):392.

John Lee, Max Dabagia, Eva L Dyer, and Christopher J Rozell. 2019. Hierarchical optimal transport for multimodal distribution alignment. *arXiv preprint arXiv:1906.11768*.

Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. 2019. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1446–1455.

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J Sutherland. 2020. Learning deep kernels for non-parametric two-sample tests. In *International Conference on Machine Learning*, pages 6316–6326. PMLR.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 164–172.

Kanti V Mardia. 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530.

M Alex Meredith, James W Nemitz, and Barry E Stein. 1987. Determinants of multisensory integration in superior colliculus neurons. i. temporal factors. *Journal of Neuroscience*, 7(10):3215–3229.

Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. 2017. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5715–5725.

Athma Narayanan, Avinash Siravuru, and Behzad Dariush. 2019. Temporal multimodal fusion for driver behavior prediction tasks using gated recurrent fusion units. *CoRR*, abs/1910.00628.

Sunny Nigam, Masanori Shimono, Shinya Ito, Fang-Chin Yeh, Nicholas Timme, Maxym Myroshnychenko, Christopher C Lapish, Zachary Tosi, Pawel Hottowy, Wesley C Smith, et al. 2016. Rich-club organization in effective connectivity among cortical neurons. *Journal of Neuroscience*, 36(3):670–684.

Leslie C Osborne, Stephanie E Palmer, Stephen G Lisberger, and William Bialek. 2008. The neural basis for combinatorial coding in a cortical population response. *Journal of Neuroscience*, 28(50):13522–13531.

Cyriel MA Pennartz. 2009. Identification and integration of sensory modalities: neural basis and relation to consciousness. *Consciousness and cognition*, 18(3):718–739.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

Josef P Rauschecker, Biao Tian, and Marc Hauser. 1995. Processing of complex sounds in the macaque nonprimary auditory cortex. *Science*, 268(5207):111–114.

Gregg H Recanzone. 2003. Auditory influences on visual temporal rate perception. *Journal of neurophysiology*, 89(2):1078–1093.

Tim Rohe and Uta Noppeney. 2016. Distinct computational principles govern multisensory integration in primary sensory and association cortices. *Current Biology*, 26(4):509–514.

Fernando E Rosas, Pedro AM Mediano, Michael Gastpar, and Henrik J Jensen. 2019. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Physical Review E*, 100(3):032305.

Samantha P Sherrill, Nicholas M Timme, John M Beggs, and Ehren L Newman. 2020. Correlated activity favors synergistic processing in local cortical networks in vitro at synaptically relevant timescales. *Network Neuroscience*, 4(3):678–697.

Charles Spence, Jon Driver, and Jonathan C Driver. 2004. *Crossmodal space and crossmodal attention*. Oxford University Press.

Terrence R Stanford, Stephan Quessy, and Barry E Stein. 2005. Evaluating the operations underlying multisensory integration in the cat superior colliculus. *Journal of Neuroscience*, 25(28):6499–6508.

Milan Studenỳ and Jirina Vejnarová. 1998. The multiinformation function as a tool for measuring stochastic dependence. In *Learning in graphical models*, pages 261–297. Springer.

Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8992–8999.

Nicholas M Timme, Shinya Ito, Maxym Myroshnychenko, Sunny Nigam, Masanori Shimono, Fang-Chin Yeh, Pawel Hottowy, Alan M Litke, and John M Beggs. 2016. High-degree neurons feed cortical computations. *PLoS computational biology*, 12(5):e1004858.

Martijn P Van Den Heuvel and Olaf Sporns. 2011. Rich-club organization of the human connectome. *Journal of Neuroscience*, 31(44):15775–15786.

Mark T Wallace, Thomas J Perrault, W David Hairston, and Barry E Stein. 2004. Visual experience is necessary for the development of multisensory integration. *Journal of Neuroscience*, 24(43):9580–9584.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019.

Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223.

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53.

Jun Yu, Feng Lin, Hock-Soon Seah, Cuihua Li, and Ziyu Lin. 2012. Image classification by multimodal subspace learning. *Pattern Recognition Letters*, 33(9):1196–1204.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *arXiv preprint arXiv:2102.04830*.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018b. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018c. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

A Ablation Study Results

,

In this section we run ablation experiments to assess the individual impact of each component of the overall synergy loss. For this purpose we use the MIM model on the MOSEI dataset, and the divergence type was chosen to be MMD. We activated each of the three loss components viz. ($\mathcal{L}_{L2}$, $\mathcal{L}_{d/MMD}$, $\mathcal{L}_{\kappa}$), trained the MIM model, and report the test accuracies on all the metrics.

	Acc_7	Acc_2	MAE	CORR
	MIM			
MLE	53.3	79.6	0.54	0.75
+ $\mathcal{L}_{L2}$	53.6	80.0	0.55	0.77
+ $\mathcal{L}_{d/MMD}$	53.1	80.3	0.57	0.73
+ $\mathcal{L}_{\kappa}$	53.9	79.9	0.54	0.76
+ S_{MMD}	54.3	82.4	0.52	0.77

Table 5: Ablation results on sentiment analysis on CMU-MOSEI. Acc_7 denotes accuracy on 7 classes and Acc_2 the binary accuracy. MAE denotes the Mean Absolute Error and Corr is the Pearson correlation

We note the performance improvement caused by adding the Kurtosis loss. We also note that directly adding the distributional divergence while mildly helpful can also degrade the performance. Each component overall has some value to add over others. We leave the exact nature of interactions between these terms for future work.

B Additional Experiments

We present results on the UR_FUNNY dataset which is another common affective sentiment prediction dataset. The model is evaluated on accuracy so higher numbers are better.

	NLL	NLL+S_{KL}	NLL+S_{MMD}
MISA	68.6	68.9	**69.6**
MFN	65.2	66.5	**67.2**
TFN	64.7	67.3	**67.8**

Table 6: Results on sentiment analysis on UR-FUNNY. The performance is evaluated in terms of accuracy

C Training details

We perform a grid-search for the best set of hyper-parameters: batch size in {32, 64}, learning rate

in {1e-2, 5e-3,1e-3, 5e-4, 1e-4}. We did gradient clipping with clip value of 5. Model selection was done following (Zadeh et al., 2017), by selecting the model with the best MAE on validation data. Optimization was done using the AdamW (Loshchilov and Hutter, 2018) optimizer. For both the Q and the g function we used a four layer MLP with LeakyRelu activation. The dataset statistics are given in Table 1. All our experiments were conducted on Nvidia Titan X GPUs.

D Algorithm

As a reminder the input the fusion problem involves learning a representation $\mathcal{M}_f$ that combined the uni-modal representations Z_i of the input $X = (X_1, X_2, .., X_k)$ where X_i are individual input modalities. We shall denote observations as X^j and the fused representations as Z^j. Q_i, g_i are multi-layer perceptrons.

Dataset $\mathcal{D} = \{(X_i), Y\}$, decay α, learning rates η_{MI}, η_{task}, hyper-parameters weights $\gamma_{L2}, \gamma_d, \gamma_\kappa$ Prediction $\hat{Y}$

for each training epoch **do**

  for minibatch $\mathcal{B} = \{(X_i^j, Y^j)\}_{j=1}^N$ sampled from $\mathcal{D}$ **do**

Encode X_i^j to Z_i^j

Compute fused vector Z^j from Z_i^j

Compute $\mathcal{L}_{L2} = \sum_j ||Q_i(Z_{-i}^j) - Z_i^j||^2$

Compute $\mathcal{L}_d = d(p(g_i(Z_i^j)), p(g_i(Z^j)))$ (d is a distribution divergence like MMD, and p refers just to the empirical distribution of these vectors)

Compute $Z_{diff}^j = clip(Z^j - \bar{Z})$

Update $S = \alpha * S + \frac{\sum_j^N (Z_{diff}^j)(Z_{diff}^j)^T}{N}$

Compute $\mathcal{L}_\kappa = \frac{1}{N} \sum_j^N [((Z^j - \bar{Z})^T (S + diag(c))^{-1} (Z^j - \bar{Z}))^2]$

Compute $\mathcal{L}_{reg} = \gamma_{L2}\mathcal{L}_{L2} + \gamma_d \mathcal{L}_d + \gamma_\kappa \mathcal{L}_\kappa$

Update $C - Network$ parameters : $\theta_C \leftarrow \theta_C - \eta_{MI} \nabla_{\theta_C} \mathcal{L}_{reg}$

Compute predictions $\hat{Y}^j$

Compute $\mathcal{L}_{task}$ (cross entropy, least square) from $\hat{Y}^j, Y^j$

Update all parameters in the model except θ_C: $\theta \leftarrow \theta - \eta_{task} \nabla_\theta [\mathcal{L}_{task} + \beta \mathcal{L}_{reg}]]$
