AN INFORMATION-THEORETIC APPROACH TO TRANSFERABILITY IN TASK TRANSFER LEARNING

Yajie Bao[1⋆] *Yang Li*[1⋆] *Shao-Lun Huang*[1] *Lin Zhang*[1] *Lizhong Zheng*[2] *Amir Zamir*[3,4] *Leonidas Guibas*[3]

[1] Tsinghua-Berkeley Shenzhen Institute [2] Massachusetts Institute of Technology
[3] Stanford University [4] University of California, Berkeley

ABSTRACT

Task transfer learning is a popular technique in image processing applications that uses pre-trained models to reduce the supervision cost of related tasks. An important question is to determine task transferability, i.e. given a common input domain, estimating to what extent representations learned from a source task can help in learning a target task. Typically, transferability is either measured experimentally or inferred through task relatedness, which is often defined without a clear operational meaning. In this paper, we present a novel metric, *H-score*, an easily-computable evaluation function that estimates the performance of transferred representations from one task to another in classification problems using statistical and information theoretic principles. Experiments on real image data show that our metric is not only consistent with the empirical transferability measurement, but also useful to practitioners in applications such as source model selection and task transfer curriculum learning.

Index Terms— Task transfer learning, Transferability, H-Score, Image recognition & classification

1 Introduction

Transfer learning is a learning paradigm that exploits the relatedness between different learning tasks in order to gain certain benefits, e.g. reducing the demand for supervision ([1]). In *task transfer learning*, we assume that the input domain of the different tasks are the same. Then for a target task $\mathcal{T}_T$, instead of learning a model from scratch, we can initialize the parameters from a previously trained model for some related source task $\mathcal{T}_S$ (Figure 1). For example, deep convolutional neural networks trained for the ImageNet classification task have been used as the source network in transfer learning for target tasks with fewer labeled data [2], such as medical image analysis [3] and structural damage recognition in buildings [4].

An imperative question in task transfer learning is *transferability*, i.e. *when a transfer may work* and *to what extent*. Given a metric capable of efficiently and accurately measuring transferability across arbitrary tasks, the problem of task transfer learning, to a large extent, is simplified to search procedures over potential transfer sources and targets as quantified by the metric. Traditionally, transferability is measured purely empirically using model loss or accuracy on the validation set ([5, 6, 7]). There have been theoretical studies that focus on *task relatedness* ([8, 9, 10, 11]). However, they either cannot be computed explicitly from data or do not directly explain task transfer performance. In this study, *we aim to estimate transferability analytically, directly from the training data*.

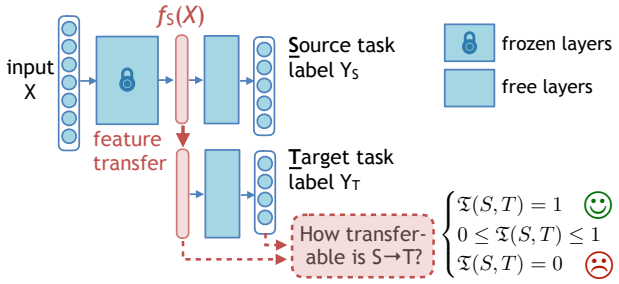


Fig. 1: A generic model of task transfer learning. The proposed transferability metric $\mathfrak{T}(S,T)$ can predict the target task's performance without training the task transfer network.

We quantify the transferability of feature representations across tasks via an approach grounded in statistics and information theory. The key idea of our method is to show that the expected log-loss of using a feature of the input data to predict the label of a given task under the probabilistic model can be characterized by an analytically expression, which we refer as the *H-score* of the feature. H-score is particularly useful to quantify feature transferability among tasks. Using this idea, we define *task transferability* as the normalized H-score of the optimal source task feature with respect to the target task.

As we demonstrate in this paper, the advantage of our transferability metric is threefold. (i) it is theoretically driven and has a strong operational meaning rooted in statistics and information theory; (ii) it can be computed directly and efficiently from the input data, with fewer samples than those needed for empirical learning; (iii) it can be shown to be strongly consistent with empirical transferability measurements.

2 Measuring Feature Effectiveness

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output space respectively. Denote the transferred feature representation by $f : \mathcal{X} \to \mathbb{R}^k$.

⋆ Joint-first authors

For a classification task, let $h_f : \mathcal{X} \times \mathcal{Y} \to [0,1]^{|\mathcal{Y}|}$ be a predictor function with the log-loss function $L(f(x), y)$ for a given (x, y) sample. The traditional machine learning approach uses stochastic gradient descent to minimize $L(h) = \mathbb{E}_{X,Y}[L(f(x), y)]$. We will show that the optimal log loss when f is given can be characterized analytically using concepts in information theory and statistics.

Definition 1. *The* Divergence Transition Matrix (DTM) *of discrete random variables X and Y is a $|\mathcal{Y}|$ by $|\mathcal{X}|$ matrix $\hat{B}$ with entries $\hat{B}_{y,x} = \frac{P_{XY}(x,y)}{\sqrt{P_X(x)}\sqrt{P_Y(y)}} - \sqrt{P_Y(y)}\sqrt{P_X(x)}$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.*

Given m training examples $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$, $L(h)$ can be written as

$$L(f,\theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{|\mathcal{Y}|} 1\{y^{(i)} = k\} \log \frac{e^{-\theta_k^T f(x^{(i)})}}{\sum_{j=1}^{|\mathcal{Y}|} e^{-\theta_j^T f(x^{(i)})}}$$

Using concepts in Euclidean information geometry, it is shown in [12] that under a local assumption, for a given feature dimension k,

$$\underset{f,\theta}{\operatorname{argmin}}\, L(f,\theta) = \underset{\Psi\in\mathbb{R}^{|\mathcal{X}|\times k}, \Phi\in\mathbb{R}^{|\mathcal{Y}|\times k}}{\operatorname{argmin}} \frac{1}{2}\|\tilde{B} - \Psi\Phi^{\mathrm{T}}\|_F^2 + o(\epsilon^2) \tag{1}$$

Let $\phi(x)$ represent row vectors of Φ for any $x \in \mathcal{X}$. By defining a one-to-one mapping $f(x) \leftrightarrow \phi(x)$ such that $\phi(x) = \sqrt{P_X(x)} f(x)$, Eq.(1) reveals a close connection between the optimal log-loss and the modal decomposition of $\tilde{B}$. In consequence, it is reasonable to measure the classification performance with $\|\tilde{B} - \Psi\Phi^{\mathrm{T}}\|_F^2$ given $f(X)$. i.e. Since Φ is fixed, we can find the optimal Ψ , Ψ^* by taking the derivative of the objective function with respect to Ψ:

$$\Psi^* = \tilde{B}\Phi(\Phi^{\mathrm{T}}\Phi)^{-1} \tag{2}$$

Substituting (2) in the Objective of (1), we can derive the following close-form solution for the log loss:

$$\|\tilde{B}\|_F^2 - \|\tilde{B}\Phi(\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\|_F^2 \tag{3}$$

The first term in (3) does not depend on $f(X)$, therefore it is sufficient to use the second term to estimate classification performance with transferred feature f. We can further rewrite $||\tilde{B}\Phi(\Phi^T\Phi)^{-\frac{1}{2}}||_F^2$ as follows and denote it as the *H-score*.

Definition 2. *Given data matrix $X \in \mathbb{R}^{m\times d}$ and label Y, let $f(X)$ be a k-dim, zero-mean feature function. The H-Score of f with respect to a task with joint probability P_{XY} is:*

$$\mathcal{H}(f) = \mathrm{tr}(\mathrm{cov}(f(X))^{-1}\mathrm{cov}(\mathbb{E}_{P_{X|Y}}[f(X)|Y])) \tag{4}$$

The derivation of (4) can be found in Section 1 of the Supplementary Material[1]. This formulation can be intuitively interpreted from a nearest neighbor perspective. i.e. a high H-score implies the inter-class variance $\mathrm{cov}(\mathbb{E}_{P_{X|Y}}[f(X)|Y])$ of f is large, while feature redundancy $\mathrm{tr}(\mathrm{cov}(f(X)))$ is small. Comparing to finding the optimal log-loss through gradient descent, H-score can be computed analytically and only requires estimating the conditional expectation $\mathbb{E}[f(X)|Y]$ from sample data. Moreover, $\mathcal{H}(f)$ has an operational meaning that characterizes the asymptotic error probability of using $f(X)$ to estimate Y in the hypothesis testing context. (See Section 2 in the Supplementary Material for details).

The upper bound of $\mathcal{H}(f)$ is obvious from its first definition: $\max_\Phi ||B\Phi(\Phi^T\Phi)^{-\frac{1}{2}}||_F^2 = ||\tilde{B}||_F^2$. We call features that achieve this bound the *minimum error probability features* for a given task.

3 Transferability

Next, we apply H-score to efficiently measure the effectiveness of task transfer learning. We will use subscripts S and T to distinguish variables for the source and the target tasks.

Definition 3 (Task transferability)**.** *Given source task $\mathcal{T}_S$, target task $\mathcal{T}_T$ and pre-trained source feature $f_S(x)$, the* transferability from $\mathcal{T}_S$ to $\mathcal{T}_T$ *is $\mathfrak{T}(S,T) \triangleq \frac{\mathcal{H}_T(f_S)}{\mathcal{H}_T(f_{T_{\text{opt}}})}$, where $f_{T_{\text{opt}}}(x)$ is the minimum error probability feature of the target task.*

This definition implies $0 \le \mathfrak{T}(S,T) \le 1$. With a known f, computing H-score from m sample data only takes $O(mk^2)$ time, where k is the dimension of $f(x)$ for $k < m$. The majority of the computation time is spent on computing the sample covariance matrix $\mathrm{cov}(f(X))$.

The remaining question is how to obtain $\mathcal{H}_T(f_{T_{opt}})$ efficiently. This question has been addressed in [13], which shows that $||\tilde{B}_T||_F^2 = \mathbb{E}[f(X)^{\mathrm{T}} g(Y)]$, where f and g are the solutions of the HGR-Maximum Correlation problem.

$$\rho(X;Y) = \sup_{\substack{f:\mathcal{X}\to\mathbb{R}^k, g:\mathcal{Y}\to\mathbb{R}^k \\ \mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0 \\ \mathbb{E}[f(X)f(X)^T] = I}} \mathbb{E}[f(X)^T g(Y)] \tag{5}$$

Eq. (6) can be solved efficiently using the Alternating Conditional Expectation (ACE) algorithm [14] for discrete $\mathcal{X}$, or using the neural network approach based on Generalized Maximal HGR Correlation [15] for a generic $\mathcal{X}$. The sample complexity of ACE is only $1/k$ of the complexity of estimating P_{YX} directly [16]. This result also applies to the Generalized HGR problem due to their theoretical equivalence.

A common technique in task transfer learning is fine-tuning, which adds before the target classifier additional free layers, whose parameters are optimized with respect to the target label. For the operational meaning of transferability to hold exactly, we require the fine tuning layers consist of only linear transformations. Nevertheless, later we will demonstrate empirically that this transferability metric can still be used for comparing the *relative* task transferability with fine-tuning. In many cases though, the computation of $H_T(f_{opt})$ can even be skipped entirely, such as the problem below:

Definition 4 (Source task selection)**.** *Given N source tasks $\mathcal{T}_{S_1}, \ldots, \mathcal{T}_{S_N}$ with labels $Y_{S_1}, \ldots, Y_{S_N}$ and a target task $\mathcal{T}_T$ with label Y_T. Let $f_{S_1}, \ldots, f_{S_N}$ be optimal representations*

[1] Supplementary materials, data and code are available at http://yangli-feasibility.com/home/ttl.html

for the source tasks. Find the source task $\mathcal{T}_{S_i}$ that maximizes the testing accuracy of predicting Y_T with feature f_{S_i}.

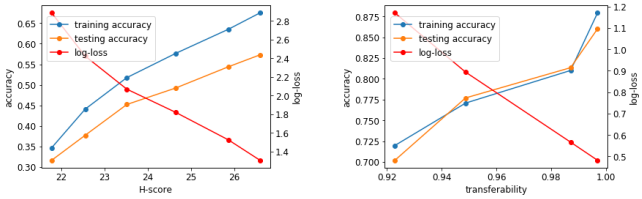
We can solve this problem by selecting the source task with the largest transferability to $\mathcal{T}_T$. In fact, we only need to compute the numerator in the transferability definition since the denominator is the same for all source tasks, i.e. $\text{argmax}_i\, \mathfrak{T}(S_i, T) = \text{argmax}_i\, \mathcal{H}(f_{S_i})$.

4 Experiments

In this section, we present validation results and potential application of our transferability metric on real image data. (For implementation details, see Section 3 of the Supplementary Material.)

4.1 Validation of transfer performance

We validate H-score and transferability definitions in a transfer learning problem from ImageNet 1000-class classification (ImageNet-1000) to Cifar 100-class classification (Cifar-100). Figure 2.a compares the H-score and empirical performance of transferring from five different layers (4a-4f) of the ResNet-50 model pretrained on ImageNet1000. As H-score increases, log-loss of the target network decreases almost linearly while the training and testing accuracy increase, which validates the relationship between the expected log-loss and H-score. The training and testing accuracy are also positively correlated with H-score. It also shows that H-score can be applied for selecting the most suitable layer for fine-tuning in transfer learning.



(a) (b)

Fig. 2: H-score and transferability vs. the empirical transfer performance measured by log-loss, training and testing accuracy. a.) Performance of ImageNet-1000 features from layers 4a-4f for Cifar-100 classification. b.): Transferability from ImageNet-1000 to 4 different target tasks based on Cifar-100.

We further tested our transferability metric for selecting the best target task for a given source task. In particular, we constructed 4 target classification tasks with 3, 5, 10, and 20 object categories from the Cifar-100 dataset. We then computed the task transferability from ImageNet-1000 (using the feature representation of layer 4f) to the target tasks. In Figure2.b, we observe a similar behavior as the H-score in the case of a single target task in Figure 2.a, showing that transferability can directly predict the empirical transfer performance.

4.2 Task transfer for 3D scene understanding

Next, we apply our transferability metric to solve the source task selection problem among 8 image-based recognition tasks for 3D scene understanding using the Taskonomy dataset[6]. We also compared the task transferability ranking based on H-score with the ranking using *task affinity*, an empirical transferability metric proposed by [6] with non-linear fine tuning.

For a fair comparison, we use the same trained encoders in [6] to extract source features with dimension $k = 2048$. It's worth noting that, six of eight tasks have images as their output. To compute the transferability for these pixel-to-pixel tasks, we cluster the pixel values of the output images in the training data into a palette of 16 colors and then compute the H-score of the source features with respect to each pixel. The transferability of the task is computed as the average of the H-scores over all pixels. For larger images, H-score can be evaluated on super pixels instead of pixels to improve efficiency.

Table. 1: List of image scene understanding tasks

Classification tasks:	Object Class., Scene Class.
Pixel-to-pixel tasks:	Keypint2D, Edge3D, Keypoint2D, Edge2D, Reshading, Depth

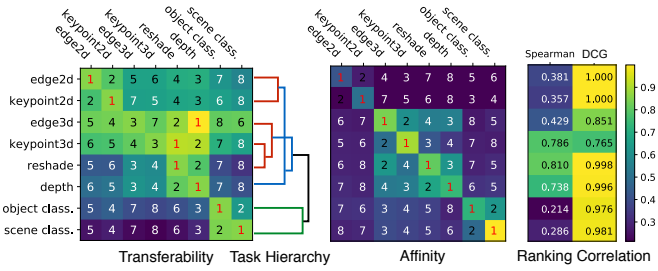


Fig. 3: Ranking comparison between transferability and affinity score.

Pairwise Transfer Results. Source task ranking results using transferability and affinity are visualized side by side in Figure 3, with columns representing source tasks and rows representing target tasks. For classification tasks (the bottom two rows in the transferability matrix), the top two transferable source tasks are identical for both methods. Similar observations can be found in 2D pixel-to-pixel tasks (top two rows). A slightly larger difference between the two rankings can be found in 3D pixel-to-pixel tasks, especially 3D Occlusion Edges and 3D Keypoints. Though the top four ranked tasks of both methods are exactly the four 3D tasks. It could indicate that these low level vision tasks are closely related to each other so that the transferability among them are inherently ambiguous. We also computed the ranking correlations between transferability and affinity using Spearman's R and Discounted Cumulative Gain (DCG). Both criterion show positive correlations for all target tasks. The correlation is especially strong with DCG as higher ranking entities are given larger weights.

To show the task relatedness, we represent each task with a vector consisting of H-scores of all the source tasks for the

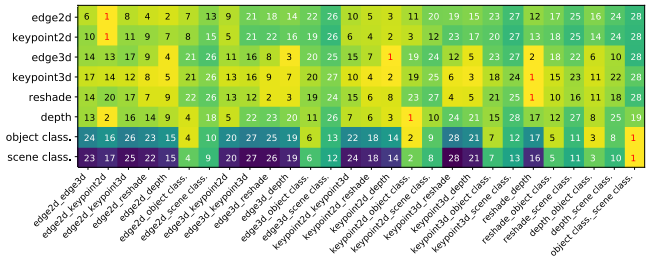


Fig. 4: Ranking of 2nd-order transferability for all tasks

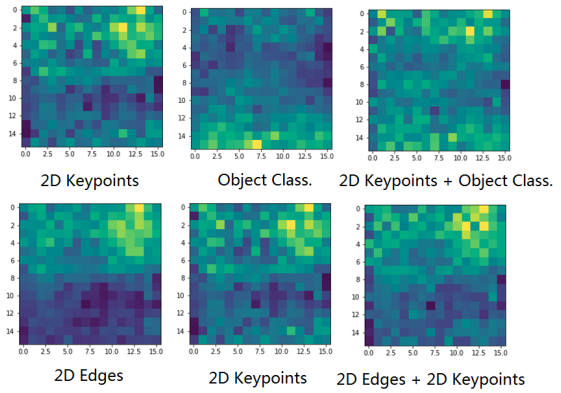


Fig. 5: 1st and 2nd order pixel-wise transferability to Depth.

given task, then apply agglomerative clustering over the task vectors. As shown in the dendrogram in Figure 3, 2D tasks and most 3D tasks are grouped into different clusters, but on a higher level, all pixel-to-pixel tasks are considered one category compared to the classifications tasks.

Higher Order Transfer. A common way for higher order transfer is to concatenate features from multiple models in deep neural networks. Our transferability definition can be easily adapted to such problems. Figure 4 shows the ranking results of all combinations of source task pairs for each target task. For all tasks except for Edge3D and Depth, the best seond-order source feature is the combination of the top two tasks of the first-order ranking. We examine the exception in Figure 5, by visualizing the pixel-by-pixel H-scores of first and second order transfers to Depth using a heatmap (lighter color implies a higher H-score). Note that different source tasks can be good at predicting different parts of the image. The top row shows the results of combining tasks with two different "transferability patterns" while the bottom row shows those with similar patterns. Combining tasks with different transferability patterns has a more significant improvement to the overall performance of the target task.

4.3 Task transfer learning curriculum

A potential application of our transferability metric is developing an optimal task transfer curriculum, a directed acyclic graph over tasks that specifies the order in which to obtain labeled data for each task. For each task in the curriculum, an optimal feature representation can be learned using both its raw input and the representations of its parent tasks to improve

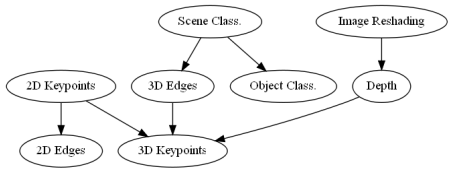


Fig. 6: Minimum spanning tree of task transferability.

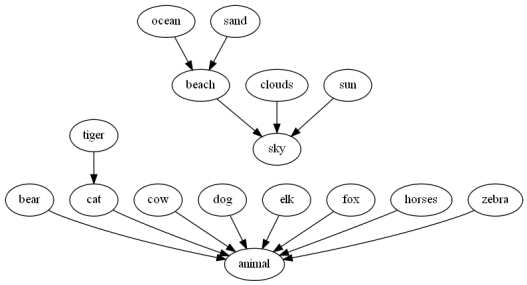


Fig. 7: Minimum spanning trees of binary image classification tasks using the NUS-WIDE multi-label dataset.

training efficiency. We use a heuristic based on the minimum spanning tree of a task graph, whose edge weights are inversely correlated with the larger transferability score between two tasks. Fig. 7.a shows the task curriculum for the eight tasks in Section 4.2. Furthermore, we did a similar experiments on a collection of binary object classification tasks using the NUS-WIDE multi-label dataset [17] (Fig. 7.b). We set a threshold to find the most salient transfers and the resulting curriculum is in line with human perception.

5 Conclusion

In this paper, we presented H-score, an information theoretic approach to estimating the performance of features when transferred across classification tasks. Then we used it to define a notion of task transferability in multi-task transfer learning problems, that is both time and sample complexity efficient. Our transferability score successfully predicted the performance for transfering features from ImageNet-1000 classification task to Cifar-100 task. Moreover, we showed how the transferability metric can be applied to a set of diverse computer vision and image-based recognition tasks using the Taskonomy and NUS-WIDE datasets. In future works, we will investigate properties of higher order transferability, developing more scalable algorithms that avoid computing the H-score of all task pairs. We also hope to design better task curriculum for task transfer learning in practical applications.

Acknowledgment

This research is funded by Natural Science Foundation of China 61807021, Shenzhen Science and Technology Research and Development Funds JCYJ20170818094022586, Innovation and Entrepreneurship Project for Overseas High-Level Talents of Shenzhen KQJSCX2018032714403783, a grant from the Toyota-Stanford Center for AI Research, a Vannevar Bush Faculty Fellowship, and NSF grant DMS-1546206.

6 References

[1] Lorien Y Pratt, "Discriminability-based transfer between neural networks," in *Advances in neural information processing systems*, 1993, pp. 204–211.

[2] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International conference on machine learning*, 2014, pp. 647–655.

[3] C. K. Shie, C. H. Chuang, C. N. Chou, M. H. Wu, and E. Y. Chang, "Transfer representation learning for medical image analysis," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug 2015, pp. 711–714.

[4] Yuqing Gao and Khalid M Mosalam, "Deep transfer learning for image-based structural damage recognition," *Computer-Aided Civil and Infrastructure Engineering*.

[5] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson, "How transferable are features in deep neural networks?," in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[6] Amir Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese, "Taskonomy: Disentangling task transfer learning," *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes, "Supervised learning of universal sentence representations from natural language inference data," *arXiv preprint arXiv:1705.02364*, 2017.

[8] Jonathan Baxter, "A model of inductive bias learning," *Journal of artificial intelligence research*, vol. 12, pp. 149–198, 2000.

[9] Andreas Maurer, "Transfer bounds for linear feature learning," *Machine learning*, vol. 75, no. 3, pp. 327–350, 2009.

[10] Anastasia Pentina and Christoph H. Lampert, "A pac-bayesian bound for lifelong learning," in *International Conference on International Conference on Machine Learning*, 2014, pp. II–991.

[11] Shai Ben-David, Reba Schuller, et al., "Exploiting task relatedness for multiple task learning," *Lecture notes in computer science*, pp. 567–580, 2003.

[12] Shao-Lun Huang, Anuran Makur, Gregory W. Wornell, and Lizhong Zheng, "On universal features for high-dimensional learning and inference," `http://allegro.mit.edu/~gww/unifeatures`, 2019.

[13] Shao-Lun Huang, Anuran Makur, Lizhong Zheng, and Gregory W Wornell, "An information-theoretic approach to universal feature selection in high-dimensional inference," in *Information Theory (ISIT), 2017 IEEE International Symposium on*. IEEE, 2017, pp. 1336–1340.

[14] Leo Breiman and Jerome H Friedman, "Estimating optimal transformations for multiple regression and correlation," *Journal of the American statistical Association*, vol. 80, no. 391, pp. 580–598, 1985.

[15] Lichen Wang, Jiaxiang Wu, Shao-Lun Huang, Lizhong Zheng, Xiangxiang Xu, Lin Zhang, and Junzhou Huang, "An efficient approach to informative feature extraction from multimodal data," *AAAI*, 2019.

[16] Anuran Makur, Fabián Kozynski, Shao-Lun Huang, and Lizhong Zheng, "An efficient algorithm for information decomposition and extraction," in *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*. IEEE, 2015, pp. 972–979.

[17] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng, "Nus-wide: A real-world web image database from national university of singapore," in *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*, Santorini, Greece., July 8-10, 2009.