April 2018

Models for automatic learner engagement estimation

Eli Sanborn Skeggs
Worcester Polytechnic Institute

Models for automatic learner engagement estimation

Advisor:

Professor Jacob Whitehill

Written By:

Eli S. Skeggs

A Major Qualifying Project
Worcester Polytechnic Institute

Submitted to the Faculty of the
Worcester Polytechnic Institute
in partial fulfillment of the requirements
for the Degree of Bachelor of Science in
Computer Science & Mathematical Sciences.

January 10th, 2018 - April 26th, 2018

Abstract

Automatic estimation of student engagement [1–6] can help computer-based learning systems adapt to individual learners [7]. Linear models trained on Gabor features established cutting-edge yet sub-human accuracy on this task [1], while convolutional neural networks (CNNs) [8, 9] overfit [10] to the dataset's few subjects [9, 11]. We found that transfer learning [12–14] enabled linear ridge regression to leverage CNN features learned for image recognition [15, 16] and face re-identification [17, 18] tasks. Our best model achieved a four-fold cross-validated correlation of $r = 0.581$, significantly outperforming [1] ($r = 0.522$). Our information strength metric correlated with model accuracy (FaceNet, $r = 0.755$; ImageNet, $r = 0.077$), inviting future study of feature utility prediction.

Table of Contents

List of Tables

LIST OF FIGURES

FIGURE Page

Chapter 1

Introduction

Computer-based learning systems such as intelligent tutoring systems, educational games, and massive online courses all serve to supplement a traditional curriculum based on lectures and problem sets [19–22]. Such systems often fall short in adapting instruction to fit the needs of their users. To better fulfill their purpose, computer-based learning systems must come to understand their students, particularly in the moment-to-moment engagement of the individual. Indeed, some computer-based learning systems attempted to use sensors or algorithms to judge the engagement of the learner [2–4], and adjusted the content or presentation to improve engagement such as in [7]. Engagement estimation ranks users on a well-defined scale of engagement. This ranking can be manual, derived from self-reports or teacher reports, or it can be automatic, via computer modeling such as [2, 23].

Several different tools have arisen to tackle automatic engagement recognition. Engagement tracing infers engagement from patterns in response timing and correctness [5], but its narrow view of the user limits inherently limits its power and capacity to generalize. Similarly, neurological and physiological sensor readings yielded signals that help determine engagement [6]. This approach required specialized equipment, limiting its adoption to small-scale research. A third variety of automatic estimation explicitly or implicitly analyses the pose and expression of the subject using computer vision techniques [1, 3]. This technique is unobtrusive compared to physical sensors, and potentially more accurate than engagement tracing [1]. The computer vision-based approach is well-positioned to see the broadest implementation, thanks to the proliferation of commodity and on-device digital cameras.

Computer vision engagement estimation requires a workable definition of engagement. For machine learning, the subjective decision-making of individual human labelers determines this definition, implicitly embedding it within the labels. [1] provided human labelers with still images containing subjects, and a definition for an engagement scale with four levels. Importantly, the human labelers agreed on the appropriate engagement labels for given images—both on a coarse scale (low vs high engagement, with Cohen's $\kappa = 0.96$), and on a granular scale (1-4, with Cohen's $\kappa = 0.56$). This agreement showed that [1] defined engagement well enough for it to be a measurable function of the images. This finding established automatic recognition of engagement as a tenable problem. Additionally, [1] found that engagement labels of constituent frames predicted the label of their containing video clip, so an accurate frame-level estimator would easily apply to video as well. Together, these results showed that frame-level automatic engagement estimation is both tractable and applicable.

Prior work in computer vision engagement estimation has not achieved human-level accuracy [1]. Approaches from traditional computer vision, such as linear combinations of Gabor filters, perform adequately. However, these approaches don't leave much room for experimentation. Unlike deep convolutional neural networks (CNNs), they cannot be easily augmented to introduce greater representational power. CNNs offer significant room for exploration in the choice of architecture design, hyper-parameters [24], and training procedures. Efforts to train deep CNNs to recognize engagement have struggled to match the accuracy of traditional computer-vision techniques [10]. To become viable, deep models need a means to sidestep their training difficulties, which are rooted in overfitting on small engagement datasets [9].

Transfer learning adapts knowledge in the form of parameters from one machine learning model to another [12, 25]. With transfer learning, deep convolutional networks can take valuable features from one domain, and use them to better understand a related domain. In particular, this pre-training works well even (and especially) if the related domain has less available data [12]. Chapter 2.3 includes further detail on transfer learning theory and practice.

We sought to understand whether transfer learning would provide a means for the engagement estimation problem to benefit from deep convolutional networks. To accomplish this, we applied transfer learning to train models that estimate the apparent engagement of a subject from still images. We designed a set of experiments to measure the efficacy of models that use features extracted from pre-trained deep computer vision models. We also designed an information strength metric, as a computational tool for deciding which layer(s) in pre-trained networks offer the most utility for engagement recognition.

Organization of the report. Chapter 2 describes relevant research. In Chapter 3, we detail the design of our transfer-learning architecture and experiments. Chapter 4 documents our experimental results, and briefly discusses findings that arose during implementation. Chapter 5 explains the results and findings. Chapter 6 talks about related future work and offers concluding remarks on the project.

CHAPTER 2

LITERATURE REVIEW

Most relevant literature for this project comes from the fields of deep learning, automatic engagement estimation, and transfer learning.

2.1 Deep learning

Deep learning research investigates challenges in designing and training deep neural networks (DNNs). For particularly deep networks (more than a dozen or so layers), conventional gradient descent for parameter optimization fails to converge [26, 27]. Deep learning includes solutions such as convolutional layers, residual connections, parameter initialization, clever architecture design, and vast datasets.

2.1.1 Convolutional networks

To solve this, convolutional layers exploit spatial information in images to help the network learn spatially-invariant features [9]. They learn restricted mappings from small windows into the input space [8, 9, 28], whereas conventional feed-forward networks learn parameters for every pair of neurons in adjacent layers [9]. Convolutional layers match patterns to subsets of the input image or spatial features. Pattern matching involves applying a set of filters to well-defined regions in the input space; each layer builds up a more sophisticated representation of the input by learning a growing number of more abstract features. Convolutions tend to learn features that are translation invariant, and that generalize to across the input space. These qualities make allow convolutional layers to boost accuracy and mitigate overfitting [8, 9].

2.1.2 Residual networks

Residual connections can solve optimization instability in DNNs [27]. They build on the intuition that optimizing an identity mapping with a residual component should be simpler than building a complete mapping that carries sufficient upstream information. Because deeper networks perform worse than their shallower counterparts, adding skip connections to deeper networks should allow them to perform at least as well as comparable shallower networks. That is, a shallow network can be trivially deepened with identity mappings $H(x) = x$, and the residual $H(x) = F(x) + x$ adds a learned term that allows the network to build more sophisticated representations.

2.1.3 Inception networks

The Inception networks arose from a careful investigation of architecture design and optimization [29]. They identified inefficiencies in deep networks and devised patterns—called Inception blocks—to increase efficiency without sacrificing performance. Inception blocks emulate sparse local connectivity using fast dense primitives [30], in a style called split-transform-merge. Inception networks further apply dimensionality reduction convolutions following the Network In Network pattern [31], which improves the representational strength of local feature extraction and increases computational efficiency. [15] proposed Inception-ResNet-v1 and Inception-ResNet-v2. These architectures explored fusing residual connections with the prior design strategies of the Inception networks, establishing a new state-of-the-art in image classification.

2.1.4 Pre-trained networks

Due to the extreme computational requirements for training deep networks from scratch, some researchers publicly release complete copies of their networks, including the learned parameters. The TensorFlow models repository [16] includes networks trained on the ILSVRC-2012-CLS dataset that perform image classification. An arbitrary selection of other models includes:

1. A model that analyzed neuroscience data [32]
2. A network that modeled natural language [33]
3. An end-to-end image recognition model that transcribes street names [34]
4. And a similar attention-based model to transcribe street names [35]

Advancements in metric learning [36–39] have facilitated the success of face re-identification models [17, 40–44]. These models learn to distinguish the identity of photographed subjects, ignoring confounding factors such as lighting, context, clothing, hairstyle, or expression. State-of-the-art face re-identification has reached accuracies upwards of 99.7%. Some of these networks have pre-trained models, such as OpenFace, VGGFace, SphereFace and NormFace, and a third-party implementation of FaceNet [18] (which achieved a competitive accuracy of 99.2% on the Labeled Faces in the Wild dataset [45]).

2.2 Automatic engagement estimation

Prior approaches have attempted to estimate engagement from images of test subjects [1, 3, 4, 10, 46–48]. Many approaches used Gabor filters and linear regression to perform expression recognition [1, 46–48]. In [1], this model achieved a correlation coefficient of $r = 0.5216$. [10] extended these findings by showing that using the label distribution from the human labelers significantly improved the performance of both a classifier and a regressor on the Gabor features.

[3] proposed a deep neural network approach that determined whether individual students were engaged based on labeled features extracted from images of students in a lecture hall. Such a system could be used as the basis for an automatic teaching assistant, which would propose specific strategies to maintain student interest during lectures [7]. This approach managed a 59% subject-independent cross-validated accuracy on their dataset, which did not test their solution on subjects in other contexts or experimental conditions [3]. [3] also did not directly report a human baseline for their results, leaving the reported accuracy with no context.

While [3] demonstrated the value of features beyond simple Gabor features, none of these concepts have attained human-level engagement recognition accuracy.

2.3 Transfer learning

Transfer learning—sometimes called domain adaptation—involves the adaptation of data or parameters from one domain to a related target domain [25]. For instance, a machine-learning model trained to transcribe spoken English to text could be adapted to transcribe spoken German, as both require an understanding of speech dynamics and involve similar frequency bands. In a sense, transfer learning compensates for relative weaknesses inherent to the design of the model. In the transcription example, transfer learning would be most useful if the corpus of labeled data were significantly larger for English than German: the German model would benefit from the volume of data from the English dataset. We focused on a type of transfer learning called inductive transfer learning, which applies when labeled data is available in the target domain [25]. Other other types of transfer learning tackle problems where there is no labeled data for the target domain, and perform a kind of unsupervised or semi-supervised learning [25, 49, 50].

Prior research has already studied the combination of transfer learning and deep learning. [51] gauged the effects of several common techniques on the transferability of models, including the impact of network depth, truncation, early-stopping, and fine-tuning. [14] showed that transfer learning enables state-of-the-art results on face attribute detection using deep features from pre-trained image classification models. Indeed, prior work has demonstrated both that transfer learning on pre-trained deep neural networks yields competitive results [13]. For example, [52] and [53] showed state-of-the-art performance using linear models on deep features. They suggested that the linear models were not just viable, but necessary to ensure generalization.

Transfer learning performs well both in improving overall accuracy and generalization beyond the conditions of the dataset. The availability of pre-trained networks has substantially increased the utility of transfer learning, as it lowers the barrier to entry. A similar approach might similarly improve test accuracy on automatic engagement estimation, where prior approaches have not been able to make the most of convolutional networks.

CHAPTER 3

METHODOLOGY

We examined whether transfer learning could enable deep networks to estimate engagement. To assess this, we devised experiments to evaluate models built on features extracted from two pre-trained deep computer vision models.

3.1 Design

First, we defined criteria for the selection of pre-trained models. The selected models needed to capture knowledge about visual patterns and relationships, and faces and expressions. To judge the models, we researched and defined the general approach, and developed the experimental design and procedures.

3.1.1 Pre-trained model selection

To allow us to concentrate on the core problem, we determined the following criteria for model selection.

1. **image-based**, not for neuroscience data analysis [32] or natural language modeling [33]
2. **relevant**: either a general image model or a human-centric one; not a specialized model such as one for street sign recognition [34, 35], to ensure we're considering the effects of relevant features
3. **single-pass**: we were interested in finding simple approaches to engagement estimation, and the adaptation of multi-pass [54] and cascading [55] models would have introduced additional complexity

4. available in **TensorFlow**, to avoid spending time researching other frameworks or migrating weights between frameworks
5. **comparable** between models, or at least similar in architecture, so the differences in results are not merely due to superior architecture

Given those requirements, we selected two networks based on the Inception-ResNet networks [15]. We selected a pre-trained Inception-ResNet-v2 model from the TensorFlow models repository [16]. We also selected [18], an open-source implementation of FaceNet [17]. Sandberg's implementation, when trained on a cleaned version of [56], achieved a competitive face re-identification accuracy of 99.2% on the LFW dataset [45]. We motivated these selections by observing that the image classification network may encode general features useful for classifying many objects, while the face re-identification network may encode specialized features that pick up on subtle facial cues. However, we hypothesized that FaceNet may have learned to erase information such as pose, expression, and lighting that would confound identity recognition. As our estimator needs expression information, this could result in a decrease in performance of our estimator. While they would have met our requirements, we did not select VGGNet [57] and VGGFace [42] because they require greater computation resources for the same accuracy. That said, a colleague did report anecdotal evidence that VGGNet encodes superior features.

FaceNet used a slightly modified version of the Inception-ResNet-v1 architecture [15], which closely resembled the architecture used in the ImageNet network. The image classification network exhibited a model capacity of about 5.6M trainable parameters as opposed to FaceNet's 2.8M. Moreover, the ImageNet model required more computation due to the higher number of layers in the network [15], and due to FaceNet's reduced input image size [17]. As a result, differences in the performance of the final estimator may be due to differences in the exact formulation of the network.

3.1.2 Estimator

[58] and [59] argued that some tasks benefit from concrete and localized features in shallower layers, and others benefit from abstract and de-localized nature of deeper layers. This argument receives further credence by [60]'s investigations into abstraction as a function of depth. Thus, different layers may yield a better representation to learn an approximation of engagement. We established our estimator protocol to explore this effect.

Because both Inception-ResNet architectures share a superstructure, we identified each of the six high-level building elements as a *hook* and assigned each a number, where larger values correspond to deeper layers (those more distant from the input image). See Figure 3.1 for a high-level map of the Inception-ResNet network, and Table 3.1 for the exact mapping between superstructure elements and labeled *hooks*. The original softmax layer in Inception-ResNet maps directly to output classes, so per convention, we truncated the networks before their softmax

hook1	*hook2*	*hook3*	*hook4*	*hook5*	*hook6*
stem	block A	reduction A	block B	reduction B	block C

Table 3.1: *The* hook*-layer mapping for Inception-ResNet feature extraction. These blocks correspond to those defined in [15].*

layers and preceding dropout and average-pooling layers. Moreover, it doesn't make a lot of intuitive sense for a model to learn engagement from a linear combination of predictions that represent the degree to which an image looks like a bird or a car.

To facilitate comparisons between [1] and our results, we retained the input image size of 48x48 grayscale. We then upscaled the input to match each network's train and test conditions: 299x299 for ImageNet, and 160x160 for FaceNet. Following the upscaling operation, we duplicated the images over a new axis to reach the expected three color channels and performed appropriate per-image standardizations. To match ImageNet's train and test conditions, we linearly scaled the input color intensities from the interval $[0,1]$ to $[-1,1]$. For FaceNet, we standardized each image X by applying Equation 3.1, using TensorFlow's `tf.image.per_image_standardization` operation. Equation 3.1 linearly scales the color intensities of image X such that $\mu_X = 0$ and $\sigma_X \approx 1$. In a series of experiments, we considered the performance of linear models trained on feature vectors from individual *hooks*.

$$X_{norm} = \frac{X - \mu_X}{\max\left(\sigma_X, N_X^{-\frac{1}{2}}\right)} \tag{3.1}$$

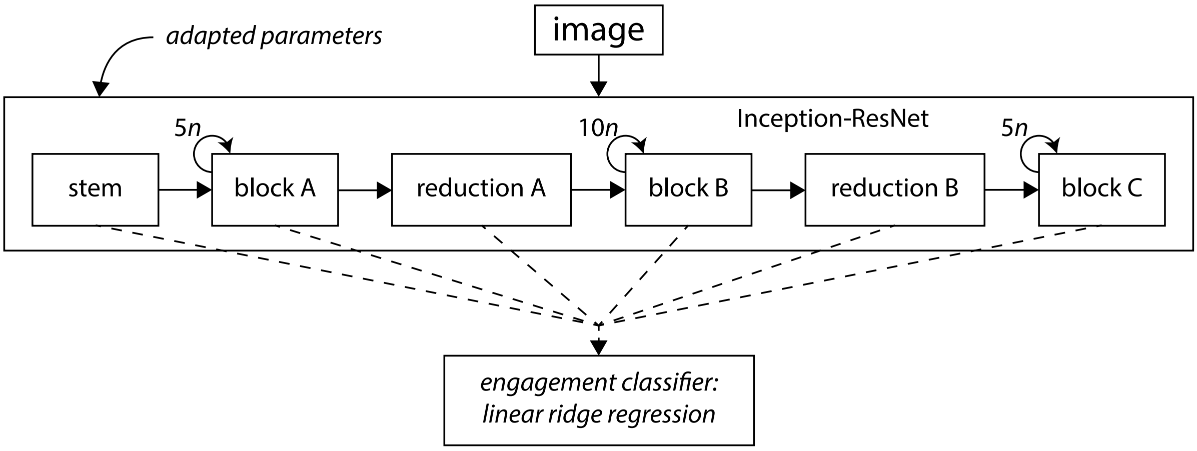

Figure 3.1: *Inception-ResNet attached to a simple linear model; applies to both FaceNet and ImageNet. The variable n is a scaling factor that corresponds to the version of the network: n = 1 for FaceNet and n = 2 for ImageNet. A single classifier operated on the outputs from just one of the hooks.*

3.2 Dataset

We followed the data annotation, and filtering procedures from [1], and reused their raw data. They collected their data from an experiment, wherein they captured video footage of participants as they used a set of cognitive skill training programs. Each image was cropped to the face bounding box as identified by [46].

The subjects in that experiment consisted of 26 undergraduate students at a Historically Black College/University (HBCU) in the southern United States. [1] asked their team of labelers to view static images from the experiment, and rate the engagement of the participant on a scale of 1-4, or X, with X being no subject/unclear. Each labeler labeled set of images that overlapped with the sets for other labelers, but the labelers did not label every image in the complete image set. This assignment balanced the need for multiple engagement data points per image with the desire to label many images. To ensure reasonably consistent labels, the labelers only considered the apparent engagement, as opposed to imagining the context. [1] used the following definition for their engagement labels:

> 1. Not engaged at all – e.g., looking away from computer and obviously not thinking about task, eyes completely closed.
>
> 2. Nominally engaged – e.g., eyes barely open, clearly not "into" the task.
>
> 3. Engaged in task – student requires no admonition to "stay on task".
>
> 4. Very engaged – student could be "commended" for his/her level of engagement in task.
>
> X. The clip/frame was very unclear, or contains no person at all.

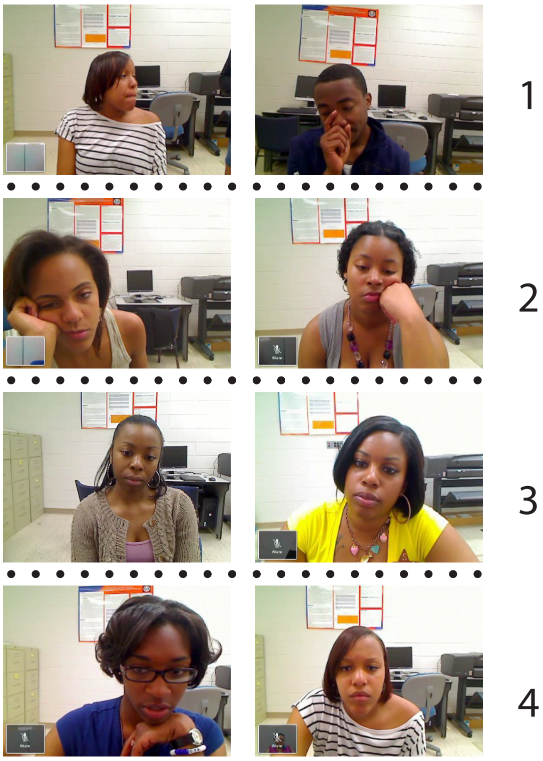

Figure 3.2: *Images of subjects from each engagement level, where labelers considered the top two images engagement level 1, and the bottom two engagement level 4 [1].*

From the labeled images, we used the following data selection procedure from [1] to eliminate problematic examples:

> We started with a pool of 13584 [image] frames from the HBCU dataset. We then applied the following procedure to select training and testing data:
>
> 1. If the minimum and maximum label given to an image differed by more than 1 (e.g., one labeler assigns a label of 1 and another assigns a label of 3), then the image was discarded. This reduced the pool from 13584 to 9796 images.
> 2. If the automatic face detector (from CERT [46]) failed to detect a face, or if the largest detected face was less than 36 pixels wide (usually indicative of [an] erroneous face detection), the image was discarded. This reduced the pool from 9796 to 7785 images.
> 3. For each of the labeled images, we considered the set of all labels given to that image by all the labelers. If any labeler marked the frame as X (no face, or very unclear), then the image was discarded. This reduced the pool from 7785 to 7574 images.
> 4. Otherwise, the "ground truth" label for each image was computed by rounding the average label for that image to the nearest integer (e.g., 2.4 rounds to 2; 2.5 rounds to 3).

This procedure yielded 10698 frames from the HBCU dataset. For reasons that weren't immediately clear, this number differed from the 7574 frames reported by [1]. Consequently, we reported results for both their approach and our approaches on the larger set of 10698 frames.

3.3 Metrics

To more accurately estimate the performance of our approaches given the small number of subjects and frames in our dataset, we employed four-fold subject-independent cross-validation. This allowed us to train and test the model over the entire dataset. We used identical folds to those used in [1] to enable direct comparison of results. Per Section 3.2, we filtered and partitioned the data, such that the images and labels for each subject were assigned to one and only one fold. We reported a cross-validated Pearson correlation coefficient to compare our approaches with Gabor filter techniques as in [1].

3.4 Hardware

We ran our experiments on a high-performance computing system acquired through NSF MRI grant DMS-1337943 to WPI. The wonderful Academic & Research Computing group at Worcester Polytechnic Institute supported our use of the computing system. We considered Microsoft's Azure cloud computing platform but found WPI's compute cluster provided the on-demand job scaling we needed to run our experiments in a time-efficient manner.

We used a mix of NVIDIA Tesla GPUs to run our experiments, including K20Xms, K40s, K80s, and V100s. We allocated between 10 and 128GiB of memory, and between 1 and 20 CPUs to each job to maximize training throughput. We allocated such a wide range of hardware because we saw a significant increase in throughput with multi-core scaling and caching of input data using TensorFlow's Dataset API, but didn't consistently need the higher number of cores or amount of memory.

3.5 Transfer learning experiment

We extracted off-the-shelf features from the ImageNet and FaceNet networks per Section 3.1. We normalized the feature vectors $v \rightarrow \hat{v}$ from each *hook*, such that $\mu_{\hat{v}} = 0$ and $\sigma_{\hat{v}} = 1$ for each example. For vector $\hat{v}$ from each *hook*, we trained a linear ridge regression model to predict engagement, using the three regularization strengths $\alpha = 0.1$, $\alpha = 1.0$, and $\alpha = 10.0$. Per Section 3.3, we reported the four-fold cross-validated correlation coefficient for each trial. As a baseline, the best prior result was $r = 0.5216$ [1]. Using a 1-sample (paired) t-test on the correlation coefficients from each fold, we tested the hypothesis that our experiments yielded a statistically significant improvement over [1].

3.5.1 Implementation details

We first ran our experiments without including the per-image normalization and zero-meaning the extracted feature vectors. This produced significantly sub-par results, and prompted us to use the normalization expected by the network and zero-mean the produced vectors. The models experienced a mean accuracy degradation of $\Delta r = -0.0714$ (up to $\Delta r = -0.1351$) without the prescribed normalization operations. Similarly, when we omitted the feature vector normalization step $v \rightarrow \hat{v}$, we saw warnings about singular matrices and numerical inaccuracy during solving. Given that the experiments did not include fine-tuning, which might have helped the network resolve its expectations for the data distribution.

3.5.2 Engagement-subject information strength metric

FaceNet was designed to differentiate people irrespective of lighting, pose, and expression [17]. Therefore, we hypothesized that the features from deeper layers in the network would tend to erase pose and expression information—information crucial to engagement estimation. In contrast, the ImageNet network should learn more abstract features of the scene in deeper layers [60].

To test our hypothesis, we needed to better understand the performance of models trained on the features from the *hooks* in the two networks. We devised the following information strength metric $S^{(h)}$ to approximate the relative strength of engagement-discriminating information and subject-identifying information.

Let $\vec{v}_{s,i}^{(h)}$ be the feature vector from *hook* h for the i^{th} frame of subject s.
Let $e_{s,i}$ be the engagement in the integer range $[1,4]$ for the i^{th} frame of subject s.

$$E^{(h)} = \mathbb{E}_{s,i,j}\left[\left\|\vec{v}_{s,i}^{(h)} - \vec{v}_{s,j}^{(h)}\right\|_2^2\right] \qquad s.t.\ \ e_{s,i} = 1,\ e_{s,j} = 4 \tag{3.2}$$

$$I^{(h)} = \mathbb{E}_{s,t,k,k'}\left[\left\|\vec{v}_{s,k}^{(h)} - \vec{v}_{t,k'}^{(h)}\right\|_2^2\right] \qquad \forall s \neq t \tag{3.3}$$

$$S^{(h)} = \frac{E^{(h)}}{I^{(h)}} \qquad \text{for } \textit{hook } h \tag{3.4}$$

In practice, we approximated the expectations in Equations 3.2 and 3.3 by sampling. For $E^{(h)}$, we sampled 32 frame pairs for every one of the 18 subjects. For $I^{(h)}$, we sampled 64 frame pairs from 32 subject pairs (ensuring each pair of subjects contained two different subjects). Feature vectors from *hooks* with a large information strength ratio $S^{(h)}$ should contain more information that helps discriminate between frames that have different levels of engagement compared to the information that helps determine the subjects' identities.

Chapter 4

Results

We implemented the model designs and executed the experimental procedures described in Chapter 3. Our results compare the accuracy of the transfer learning-based approaches to the approach by [1] based on Gabor filters, shown in Tables 4.1 and 4.2, and Figures 4.1 and 4.2. The cross-validated accuracy of [1]'s Gabor filter-based approach managed a correlation coefficient of $r = 0.5216$. The average four-fold cross-validated correlation accuracy using the *hooks* from ImageNet was $r = 0.5751$, and from FaceNet was $r = 0.5809$; both higher than using Gabor features.

	hook1	*hook2*	*hook3*	*hook4*	*hook5*	*hook6*
$\alpha = 0.1$	0.5600	0.5700	0.5566	0.5507	0.5655	0.5351
$\alpha = 1.0$	0.5629	0.5687	0.5670	**0.5809***	0.5634	0.5363
$\alpha = 10.0$	0.4836	0.4776	0.4705	0.4950	0.4926	0.4628

Table 4.1: *The cross-validated correlation accuracy of the linear ridge regression models trained on the outputs of each fundamental building element from the FaceNet pre-trained network. The α parameter corresponds to the regularization strength for the linear ridge regression model. The entry in bold designates the hook and regularization that yielded the best accuracy.*

	hook1	*hook2*	*hook3*	*hook4*	*hook5*	*hook6*
$\alpha = 0.1$	0.5495	0.5560	**0.5751**	0.5526	0.4320	0.3974
$\alpha = 1.0$	0.5077	0.5510	0.5469	0.5399	0.4377	0.4306
$\alpha = 10.0$	0.4369	0.4768	0.4474	0.4761	0.4003	0.3947

Table 4.2: *The cross-validated correlation accuracy of the linear ridge regression models trained on the outputs of each fundamental building element from the ImageNet pre-trained network. The α parameter corresponds to the regularization strength for the linear ridge regression model. The entry in bold designates the hook and regularization that yielded the best accuracy.*

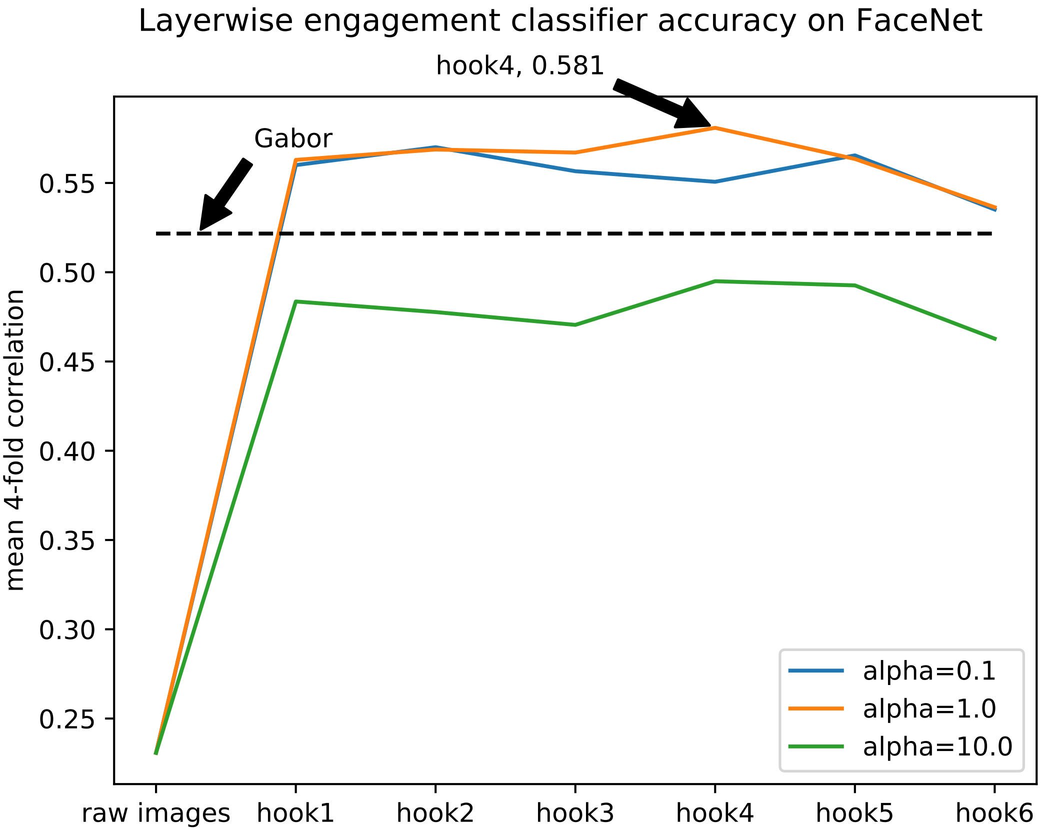

Figure 4.1: *A graph of linear ridge regression correlation accuracy trained on the outputs of each fundamental building element from the FaceNet pre-trained network. The α parameter corresponds to the regularization strength for the linear ridge regression model. The dashed line labeled Gabor represents the previous state-of-the-art from [1]. The raw images reference point represents linear regression on the raw 48x48 grayscale images.*

4.1 Analysis

We performed the 1-sample (paired) t-test described in Chapter 3.5 on the results. The test considered the pair of correlation coefficient from each fold as corresponding samples, taking samples from [1] and our results. We used the test to determine whether particular experiments with the transfer learning model outperformed the results of the Gabor approach. On the best result of FaceNet *hook4* features with an $\alpha = 1.0$, we found a T-value of 3.6245, which yielded a p-value of 0.0361. The 95% confidence interval for the improvement was (0.0072, 0.1113). This analysis showed that our model is a viable alternative to Gabor filters. We could not assert whether our approach is strictly better than Gabor filters, as we took the best of three regularization parameters, and did not have enough data to perform nested cross-validation.

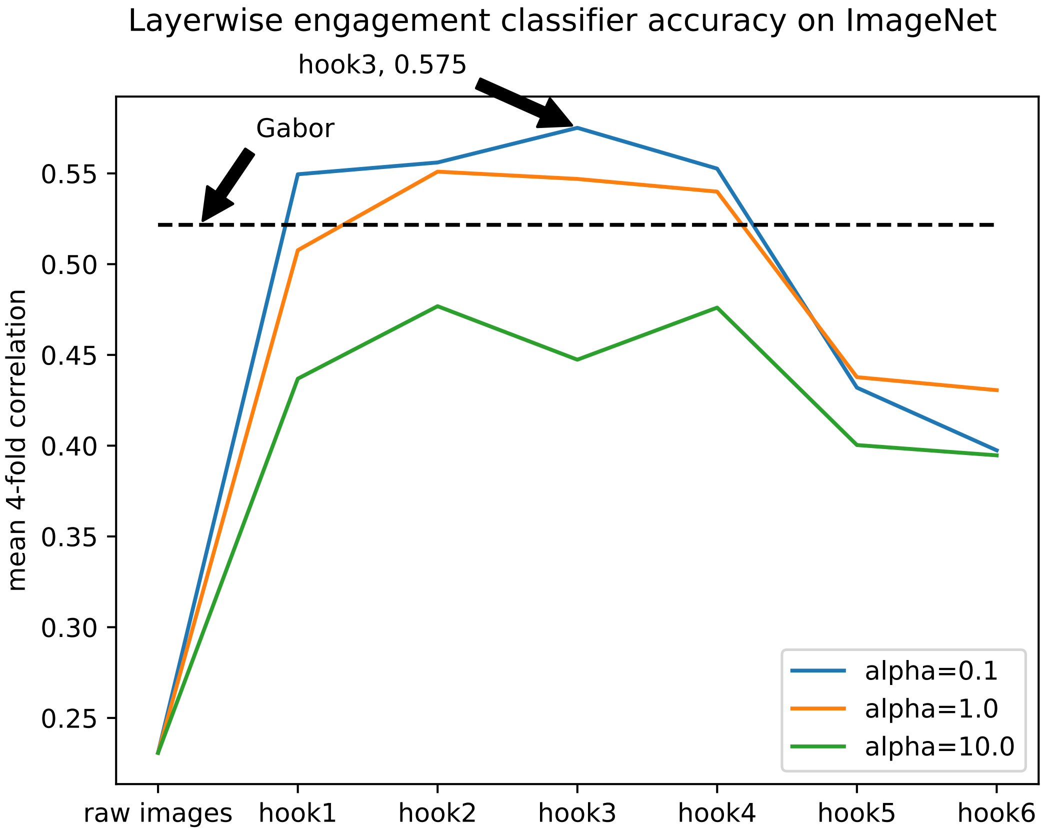

Figure 4.2: *A graph of linear ridge regression correlation accuracy trained on the outputs of each fundamental building element from the ImageNet pre-trained network. The α parameter corresponds to the regularization strength for the linear ridge regression model. The dashed line labeled Gabor represents the previous state-of-the-art from [1]. The raw images reference point represents linear regression on the raw 48x48 grayscale images.*

	FaceNet	ImageNet
$\alpha = 0.1$	0.786	0.022
$\alpha = 1.0$	0.911	0.164
$\alpha = 10.0$	0.567	0.045

Table 4.3: *The correlation between the model accuracy as reported in Tables 4.1 and 4.2, and the information strength metric reported in Table 4.4*

4.2 Information strength metric

We computed the engagement-subject information strength $S^{(h)}$ (defined in Chapter 3.5.2) for each *hook* and reported the strength values in Figure 4.3 and Table 4.4. We also computed the correlation between the information strength metric and the correlation accuracy on the two networks across the *hooks* and reported them in Table 4.3. FaceNet correlated strongly with the

Model	*hook1*	*hook2*	*hook3*	*hook4*	*hook5*	*hook6*
FaceNet	0.8439	0.8561	0.8605	0.8499	0.8145	0.6641
ImageNet	0.9083	0.9405	0.9804	0.9485	0.9330	0.9586

Table 4.4: *The information strength metric as defined in Chapter 3.5.2 for features extracted from the* hooks *of both models. The different values for the raw image depth correspond to the two expected input image sizes of the networks.*

metric ($r = 0.755$), whereas ImageNet did not correlate with it ($r = 0.077$).

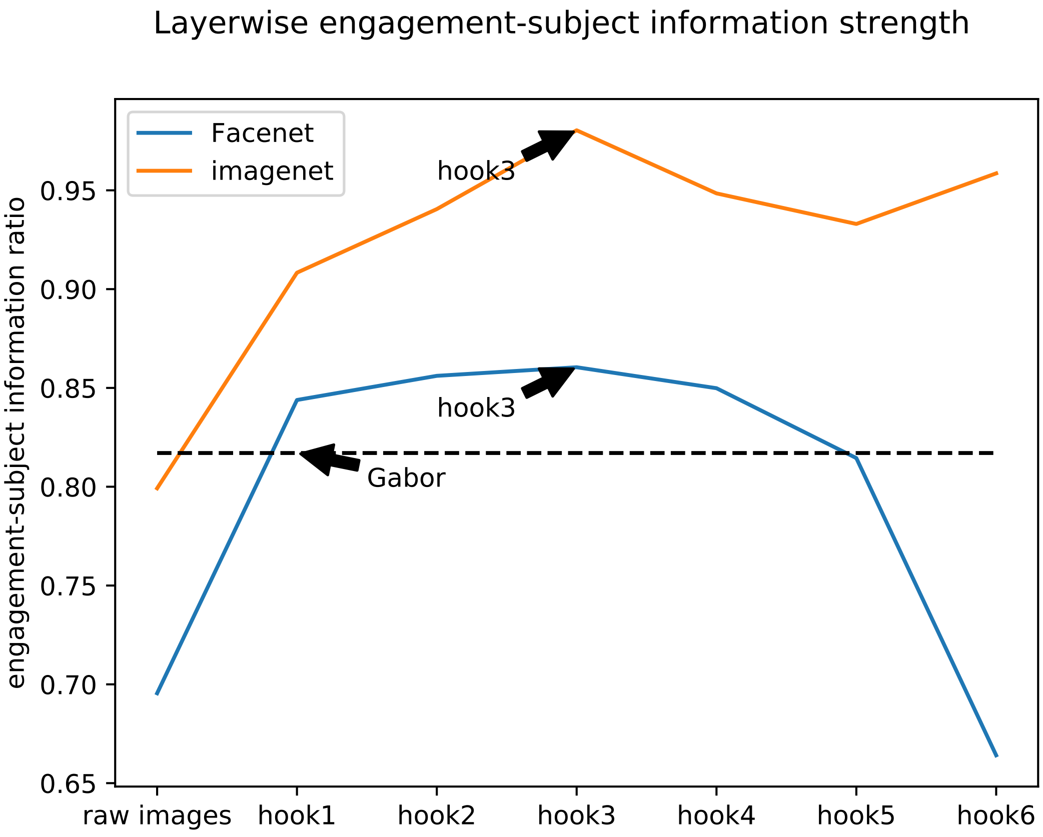

Figure 4.3: *A graph of engagement-subject information strength. Larger values correspond to relatively more engagement information than subject information. The dashed line labeled Gabor corresponds to the information strength metric for the features of the bank of Gabor filters from [1].*

Chapter 5

Discussion

We aimed to determine whether convolutional neural networks would benefit from transfer learning when estimating engagement. Our results demonstrated that transfer learning allows convolutional models, without any fine-tuning, to achieve a higher correlation accuracy than Gabor filters and linear regression. Moreover, we can achieve a significantly higher accuracy when we also optimize the hyper-parameter α in the same cross-validation experiment.

Ideally, hyper-parameters such as the regularization parameter α should be evaluated against validation data before being tested on the test data. Varying the regularization parameter without using a separate test set or nested cross-validation thus departed from best-practice. We did so to get a sense for the upper performance bound of the approach, and to allow comparison of results, as [1] partitioned the HBCU dataset into the same folds that we used. Moreover, because the size of the HBCU dataset required a technique such as cross-validation to estimate accuracy, the addition of a validation fold would have reduced the performance of the final classifier. The size of the dataset requires that experiments trade-off accuracy between optimizing the right factors and avoiding data contamination.

We weren't able to conclude whether general or face-derived deep features yield the best performance, due to the inherent differences in the ImageNet and FaceNet's complexity and capacity. In retrospect, we may have eliminated this inconsistency by training the models on features extracted from VGGNet-16 [57] and VGGFace [42]. These models use an identical architecture, and while VGGNet-16 is more inefficient and yields worse accuracy than VGGNet-19 and Inception-ResNet [15], that comparability would have allowed us to reach stronger conclusions.

We noted with interest that the information strength metric correlated strongly with FaceNet performance, but didn't correlate with ImageNet performance. This discrepancy may indicate nothing, due to the differences in the networks' design and implementation. Alternatively, this may indicate that one (or both) of the linear models underutilized their feature vectors.

The size of the input image likely limited the ability for the off-the-shelf models to extract good features. The network may have benefited from full-color images and an increase in input resolution. The pre-trained models may have also seen an improvement from data augmentation, whereas we used none.

Chapter 6

Future work & Conclusion

Although we focused on linear models due to their precedent of performing well on small datasets [52, 53], we hypothesize that this dataset might be large enough to allow slightly deeper models to learn good parameters from the *hooks*. Our pilot efforts to train deeper models on top of the *hooks* did not result in increased accuracy compared to just linear ridge regression. Future research may benefit from better tools such as [60] that help to visualize and debug intermediate features.

We did not fine-tune the pre-trained models. Fine-tuning has worked in other problem domains [12, 61–63]. While some research has indicated that fine-tuning on small datasets results in overfitting [12] or is unnecessary [11], other findings indicated that fine-tuning is an important step [61, 62]. Curiously, [63] found that for expression recognition, the resolution of the input image determines whether fine-tuning benefits the task. They linked this factor to the possibility that holistic processing may dominate facial expression recognition [64–66]. This evidence merely indicates that we need to further experiment with fine-tuning, to understand the conditions under which it improves accuracy.

We experimented solely with transfer learning. Multi-task learning has the same theoretical foundations as transfer learning, and may similarly help overcome poor generalization on small datasets.

We also propose that erasing confounding information as a preprocessing step or a form of network regularization may improve final model accuracy. The HBCU dataset included a subject that wore glasses, and the cross-validated fold that included the subject in its test partition performed worse than the other folds. The preprocessing or regularization might help eliminate information that identifies subject gender, skin color, or age, yielding a representation with fewer potential confounds. This could be potentially formulated as a minimax problem [67] or the optimization of a ratio of discriminabilities for different tasks (such as maximizing the discriminability of engagement while minimizing the discriminability of glasses) [68]. These approaches could serve as yet another way to amplify the value of small engagement datasets using supplementary data.

Future research might benefit from other complex model extensions, including end-to-end spatial transformations [69], and visual attention-based approaches [70]. Small datasets should also benefit from data augmentation, including varying the rotation, position, and size of bounding boxes, the addition of obscuring elements, and synthetic augmentation such as [71]. However, these particular extensions are more experimental and than our other recommendations, and we believe they would yield marginal improvements.

The models we proposed did not achieve human-level accuracy in engagement estimation. Our experiments showed that transfer learning enables automatic engagement estimation to benefit from convolutional networks. The utility of the information strength metric indicated that the features might be abstract enough that we could reliably determine useful feature layers for transfer learning. Our results provided further evidence that the transfer learning can overcome some limitations of small datasets.

Bibliography

[1] Jacob R. Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan.
The faces of engagement: Automatic recognition of student engagement from facial expressions.
IEEE Transactions on Affective Computing, 5(1):86–98, 2014.

[2] Brais Martinez and Michel F. Valstar.
Advances, challenges, and opportunities in automatic facial expression recognition.
In *Advances in Face Detection and Facial Image Analysis*, pages 63–100. Springer, 2016.

[3] Richard Klein and Turgay Celik.
The Wits Intelligent Teaching System: Detecting student engagement during lectures using convolutional neural networks.
In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2856–2860, Sept 2017.

[4] Hamed Monkaresi, Nigel Bosch, Rafael A. Calvo, and Sidney K. D'Mello.
Automated detection of engagement using video-based estimation of facial expressions and heart rate.
IEEE Transactions on Affective Computing, 8(1):15–28, 2017.

[5] Beck E. Joseph.
Engagement tracing: using response times to model student disengagement.
Artificial intelligence in education: Supporting learning through intelligent and socially informed technology, 125:88, 2005.

[6] Stephen H. Fairclough and Louise Venables.
Prediction of subjective states from psychophysiology: A multivariate approach.
Biological psychology, 71(1):100–110, 2006.

[7] Kate Forbes-Riley and Diane Litman.
Adapting to multiple affective states in spoken dialogue.
In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226. Association for Computational Linguistics, 2012.

[8] Yann LeCun and Yoshua Bengio.
Convolutional networks for images, speech, and time series.
The handbook of brain theory and neural networks, 3361(10):1995, 1995.

[9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville.
Deep Learning.
MIT Press, 2016.
`http://www.deeplearningbook.org`.

[10] Arkar Min Aung and Jacob R. Whitehill.
Harnessing label uncertainty to improve modeling: An application to student engagement recognition.
2018.

[11] 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*, pages 647–655, 2014.

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

[13] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson.
CNN features off-the-shelf: an astounding baseline for recognition.
In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 512–519. IEEE, 2014.

[14] Yang Zhong, Josephine Sullivan, and Haibo Li.
Face attribute prediction using off-the-shelf CNN features.
In *Biometrics (ICB), 2016 International Conference on*, pages 1–7. IEEE, 2016.

[15] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi.
Inception-v4, Inception-ResNet and the impact of residual connections on learning.
In *AAAI*, volume 4, page 12, 2017.

[16] TensorFlow-Slim image classification model library.
`https://github.com/tensorflow/models/tree/master/research/slim`, 2018.

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin.
FaceNet: A unified embedding for face recognition and clustering.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[18] David Sandberg.
Face recognition using TensorFlow.
`https://github.com/davidsandberg/facenet`, 2018.

[19] Kenneth R. Koedinger, John R. Anderson, William H. Hadley, and Mary A. Mark.
Intelligent tutoring goes to school in the big city.
1997.

[20] John R. Anderson.
Acquisition of cognitive skill.
Psychological review, 89(4):369, 1982.

[21] Kurt Vanlehn, Collin Lynch, Kay Schulze, Joel A. Shapiro, Robert Shelby, Linwood Taylor, Don Treacy, Anders Weinstein, and Mary Wintersgill.
The Andes physics tutoring system: Lessons learned.
International Journal of Artificial Intelligence in Education, 15(3):147–204, 2005.

[22] Neil T. Heffernan and Cristina Lindquist Heffernan.
The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching.
International Journal of Artificial Intelligence in Education, 24(4):470–497, 2014.

[23] Nigel Bosch, Sidney D'Mello, Ryan Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao.
Automatic detection of learning-centered affective states in the wild.
In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 379–388. ACM, 2015.

[24] Yoshua Bengio.
Practical recommendations for gradient-based training of deep architectures.
In *Neural networks: Tricks of the trade*, pages 437–478. Springer, 2012.

[25] Sinno Jialin Pan and Qiang Yang.
A survey on transfer learning.
IEEE Transactions on knowledge and data engineering, 22(10):1345–1359, 2010.

[26] Xavier Glorot and Yoshua Bengio.
Understanding the difficulty of training deep feedforward neural networks.
In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.
Deep residual learning for image recognition.

In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[28] Yann LeCun.
Generalization and network design strategies.
Connectionism in perspective, pages 143–155, 1989.

[29] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich.
Going deeper with convolutions.
Cvpr, 2015.

[30] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma.
Provable bounds for learning some deep representations.
In *International Conference on Machine Learning*, pages 584–592, 2014.

[31] Min Lin, Qiang Chen, and Shuicheng Yan.
Network In Network.
arXiv preprint arXiv:1312.4400, 2013.

[32] Chethan Pandarinath, Daniel J. O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D. Stavisky, Jonathan C. Kao, Eric M. Trautmann, Matthew T. Kaufman, Stephen I. Ryu, Leigh R. Hochberg, Jaimie M. Henderson, Krishna V. Shenoy, Larry F. Abbott, and David Sussillo.
Inferring single-trial neural population dynamics using sequential auto-encoders.
bioRxiv, 2017.

[33] Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins.
Globally normalized transition-based neural networks.
arXiv preprint arXiv:1603.06042, 2016.

[34] Raymond Smith, Chunhui Gu, Dar-Shyang Lee, Huiyi Hu, Ranjith Unnikrishnan, Julian Ibarz, Sacha Arnoud, and Sophia Lin.
End-to-end interpretation of the French Street Name Signs dataset.
In *European Conference on Computer Vision*, pages 411–426. Springer, 2016.

[35] Zbigniew Wojna, Alex Gorban, Dar-Shyang Lee, Kevin Murphy, Qian Yu, Yeqing Li, and Julian Ibarz.
Attention-based extraction of structured information from Street View imagery.
arXiv preprint arXiv:1704.03549, 2017.

[36] Laurens Van Der Maaten and Kilian Weinberger.
Stochastic triplet embedding.
In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6. IEEE, 2012.

[37] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese.
Deep metric learning via lifted structured feature embedding.
In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 4004–4012. IEEE, 2016.

[38] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao.
A discriminative feature learning approach for deep face recognition.
In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

[39] Swami Sankaranarayanan, Azadeh Alavi, Carlos D Castillo, and Rama Chellappa.
Triplet probabilistic embedding for face verification and clustering.
In *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on*, pages 1–8. IEEE, 2016.

[40] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan.
OpenFace: A general-purpose face recognition library with mobile applications.
Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[41] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf.
DeepFace: Closing the gap to human-level performance in face verification.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[42] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman.
Deep Face Recognition.
In *BMVC*, volume 1, page 6, 2015.

[43] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song.
Sphereface: Deep hypersphere embedding for face recognition.
In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 2017.

[44] Feng Wang, Xiang Xiang, Jian Cheng, and Alan L. Yuille.
NormFace: L_2 hypersphere embedding for face verification.
arXiv preprint arXiv:1704.06369, 2017.

[45] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller.

Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments.
Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[46] Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian Fasel, Mark Frank, Javier Movellan, and Marian Bartlett.
The Computer Expression Recognition Toolbox (CERT).
In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 298–305. IEEE, 2011.

[47] Tingfan Wu, Marian S. Bartlett, and Javier R. Movellan.
Facial expression recognition using Gabor motion energy filters.
In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 42–47. IEEE, 2010.

[48] Ahmed Bilal Ashraf, Simon Lucey, and Tsuhan Chen.
Reinterpreting the application of Gabor filters as a manipulation of the margin in linear support vector machines.
IEEE transactions on pattern analysis and machine intelligence, 32(7):1335–1341, 2010.

[49] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y. Ng.
Self-taught learning: transfer learning from unlabeled data.
In *Proceedings of the 24th international conference on Machine learning*, pages 759–766. ACM, 2007.

[50] Grégoire Mesnil, Yann Dauphin, Xavier Glorot, Salah Rifai, Yoshua Bengio, Ian Goodfellow, Erick Lavoie, Xavier Muller, Guillaume Desjardins, David Warde-Farley, Pascal Vincent, Aaron Courville, and James Bergstra.
Unsupervised and transfer learning challenge: a deep learning approach.
In *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning workshop-Volume 27*, pages 97–111. JMLR. org, 2011.

[51] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson.
From generic to specific deep representations for visual recognition.
In *CVPRW DeepVision Workshop, June 11, 2015, Boston, MA, USA*. IEEE conference proceedings, 2015.

[52] Mercedes Torres Torres, Michel F. Valstar, Caroline Henry, Carole Ward, and Don Sharkey.
Small sample deep learning for newborn gestational age estimation.
pages 79–86, 2017.

[53] Durjoy Sen Maitra, Ujjwal Bhattacharya, and Swapan K. Parui.
CNN based common approach to handwritten character recognition of multiple scripts.
In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 1021–1025. IEEE, 2015.

[54] Koen E. A. Van de Sande, Jasper R. R. Uijlings, Theo Gevers, and Arnold W. M. Smeulders.
Segmentation as selective search for object recognition.
In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1879–1886. IEEE, 2011.

[55] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao.
Joint face detection and alignment using multitask cascaded convolutional networks.
IEEE Signal Processing Letters, 23(10):1499–1503, Oct 2016.

[56] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li.
Learning face representation from scratch.
arXiv preprint arXiv:1411.7923, 2014.

[57] Karen Simonyan and Andrew Zisserman.
Very deep convolutional networks for large-scale image recognition.
arXiv preprint arXiv:1409.1556, 2014.

[58] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa.
An all-in-one convolutional neural network for face analysis.
In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 17–24. IEEE, 2017.

[59] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa.
HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[60] Matthew D. Zeiler and Rob Fergus.
Visualizing and understanding convolutional networks.
In *European conference on computer vision*, pages 818–833. Springer, 2014.

[61] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman.
Return of the devil in the details: Delving deep into convolutional nets.
arXiv preprint arXiv:1405.3531, 2014.

[62] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik.
Rich feature hierarchies for accurate object detection and semantic segmentation.

In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[63] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler.
Deep learning for emotion recognition on small datasets using transfer learning.
In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 443–449. ACM, 2015.

[64] Akinyinka Omigbodun and Garrison Cottrell.
Is facial expression processing holistic?
In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35, 2013.

[65] Emily R. Prazak and E. Darcy Burgund.
Keeping it real: Recognizing expressions in real compared to schematic faces.
Visual Cognition, 22(5):737–750, 2014.

[66] Emilie Meaux and Patrik Vuilleumier.
Facing mixed emotions: analytic and holistic perception of facial emotion expressions engages separate brain networks.
NeuroImage, 141:154–173, 2016.

[67] Jihun Hamm.
Minimax filter: Learning to preserve privacy from inference attacks.
arXiv preprint arXiv:1610.03577, 2016.

[68] Jacob R. Whitehill and Javier R. Movellan.
Discriminately decreasing discriminability with learned image filters.
In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2488–2495. IEEE, 2012.

[69] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu.
Spatial Transformer Networks.
In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[70] Artsiom Ablavatski, Shijian Lu, and Jianfei Cai.
Enriched deep recurrent visual attention model for multiple object recognition.
In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 971–978. IEEE, 2017.

[71] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz.
mixup: Beyond empirical risk minimization.
arXiv preprint arXiv:1710.09412, 2017.