Supplementary Material for
‘Visual Causal Feature Learning’

Krzysztof Chalupka
Computation and Neural Systems
California Institute of Technology
Pasadena, CA, USA

Pietro Perona
Electrical Engineering
California Institute of Technology
Pasadena, CA, USA

Frederick Eberhardt
Humanities and Social Sciences
California Institute of Technology
Pasadena, CA, USA

A PROOF OF THE CAUSAL COARSENING THEOREM

We first define the relevant notions from basic set theory.

Definition 1 (Partition, Coarsening)**.** *Let A be a set and Π a set of disjoint non-empty subsets of A. We say that Π is a* partition *of A if $\cup_{P \in \Pi} = A$. Let Π_1, Π_2 be two partitions of A. We say that Π_1 is a* coarsening *of Π_2 if each $P \in \Pi_2$ is a subset of some $Q \in \Pi_1$. In addition, Π_1 is a* proper coarsening *of Π_2 if it is a coarsening and $\Pi_1 \neq \Pi_2$.*

Before we prove the Causal Coarsening Theorem, we prove its less general version in order to split the rather complex proof of CCT into two parts. This Auxiliary Theorem can be proven using simpler techniques, however here we deliberately use techniques that transfer directly to the proof of the CCT.

Auxiliary Theorem *Among all the generative models of the form discussed in Fig. 2 (in the main text), the subset of distributions $P(T, H, I)$ for which the causal partition is not a coarsening (proper or improper) of the observational partition is Lebesgue measure zero.*

Proof. Our proof is inspired by a proof used by Meek [1995] to prove that almost all distributions compatible with a given causal graph are faithful. The proof strategy is thus first to express the proposition that for a given distribution, the observational partition does not refine the causal partition, as a polynomial equation on the space of all distributions compatible with the model. We then show that this polynomial equation is not trivial, i.e. there is at least one distribution that is not its root. By a simple algebraic lemma, this will prove the theorem. We extend Meek's proof technique in our usage of Fubini's Theorem for the Lebesgue integral. It allows us to “split” the polynomial constraint into multiple different constraints along several of the distribution parameters. This allows for additional flexibility in creating useful assumptions (in our proof, the assumption that the datapoints have well-defined causal classes, but the observational class can still vary freely).

Assume that T is binary and $H = (H_1, \cdots, H_M)$, I are discrete variables (say $|H_i| = K_i, |I| = N$, though N can be very large; we will use the notation $K \triangleq K_1 \times \cdots \times K_M$ for simplicity). We can factorize the joint as $P(T, H, I) = P(T \mid H, I)P(I \mid H)P(H)$. $P(T \mid H, I)$ can be parametrized by $|H_1| \times \cdots \times |H_M| \times |I| = K \times N$ parameters, $P(I \mid H)$ by $(N-1) \times K$ parameters, and $P(H)$ by another K parameters, all of which are independent. Call the parameters, respectively,

$$\alpha_{h,i} \triangleq P(T = 0 \mid H = h, I = i)$$
$$\beta_{i,h} \triangleq P(I = i \mid H = h)$$
$$\gamma_h \triangleq P(H = h)$$

We will denote parameter vectors as

$$\alpha = (\alpha_{h_1,i_1}, \cdots, \alpha_{h_K,i_N}) \in \mathbb{R}^{K \times N}$$
$$\beta = (\beta_{i_1,h_1}, \cdots, \beta_{i_{N-1},h_K}) \in \mathbb{R}^{(N-1) \times K}$$
$$\gamma = (\gamma_{h_1}, \cdots, \gamma_{h_K}) \in \mathbb{R}^{K},$$

where the indices are arranged in lexicographical order. This creates a one-to-one correspondence of each possible joint distribution $P(T, H, I)$ with a point $(\alpha, \beta, \gamma) \in P[\alpha, \beta, \gamma] \subset \mathbb{R}^{K^3 \times N \times (N-1)}$, where $P[\alpha, \beta, \gamma]$ is the $K^3 \times N \times (N-1)$-dimensional simplex of multinomial distributions.

To proceed with the proof, we first pick any point in the $P(T \mid H, I) \times P(H)$ space: that is, we fix the values of α and γ. The only free parameters are now $\beta_{i,h}$ for all values of i, h; varying these values creates a subset of the space of all the distributions which we will call

$$P[\beta; \alpha, \gamma] = \{(\alpha, \beta, \gamma) \mid \beta \in [0, 1]^{(N-1) \times K}\}.$$

$P[\beta; \alpha, \gamma]$ is a subset of $P[\alpha, \beta, \gamma]$ isometric to the $[0, 1]^{(N-1) \times K}$-dimensional simplex of multinomials. We will use the term $P[\beta; \alpha, \gamma]$ to refer both the subset of $P[\alpha, \beta, \gamma]$ and the lower-dimensional simplex it is isometric to, remembering that the latter comes equipped with the Lebesgue measure on $\mathbb{R}^{(N-1) \times K}$.

Now we are ready to show that the subset of $P[\beta; \alpha, \gamma]$ which does not satisfy the Causal Coarsening constraint is of measure zero with respect to the Lebesgue measure. To see this, first note that since α and γ are fixed, each image i has a well-defined causal class $C(i) = \sum_h \alpha_{h,i}\gamma_h$. The Causal Coarsening constraint says "For every pair of images i, j such that $P(T \mid i) = P(T \mid j)$ it holds that $C(i) = C(j)$." The subset of $P[\beta; \alpha, \gamma]$ of all distributions that do not satisfy the constraint consists of the $P(T, H, I)$ for which for some i, j it holds that

$$P(T = 0 \mid i) = P(T = 0 \mid j) \quad \text{and} \quad C(i) \neq C(j).$$

Take any pair i, j for which $C(i) \neq C(j)$ (if such a pair does not exist, then the Causal Coarsening constraint holds for all the distributions in $P[\beta; \alpha, \gamma]$). We can write

$$\begin{aligned} P(T = 0 \mid i) &= \sum_h P(T = 0 \mid h, i)P(h \mid i) \\ &= \frac{1}{P(i)} \sum_h P(T = 0 \mid h, i)P(i \mid h)P(h). \end{aligned}$$

Since the same equation applies to $P(T = 0 \mid j)$, the constraint $P(T \mid i) = P(T \mid j)$ can be rewritten

$$\begin{aligned} &\frac{1}{P(i)} \sum_h P(T = 0 \mid h, i)P(i \mid h)P(h) \\ &\quad = \frac{1}{P(j)} \sum_h P(T = 0 \mid h, j)P(j \mid h)P(h) \\ &\iff P(j) \sum_h P(T = 0 \mid h, i)P(i \mid h)P(h) \\ &\quad - P(i) \sum_h P(T = 0 \mid h, j)P(j \mid h)P(h) = 0, \end{aligned}$$

which we can rewrite in terms of the independent parameters (after defining $\alpha_{0,h,i} = \alpha_{h,i}$ and $\alpha_{1,h,i} = 1 - \alpha_{h,i}$) and further simplify as

$$\begin{aligned} &\left(\sum_{t \in \{0,1\}} \sum_h \alpha_{t,h,j}\gamma_h\beta_{j,h} \right) \sum_h \alpha_{0,h,i}\gamma_h\beta_{i,h} - \\ &\quad - \left(\sum_{t \in \{0,1\}} \sum_h \alpha_{t,h,i}\gamma_h\beta_{i,h} \right) \sum_h \alpha_{0,h,j}\gamma_h\beta_{j,h} = 0 \\ &\iff \left(\sum_h \alpha_{1,h,j}\gamma_h\beta_{j,h} \right) \sum_h \alpha_{0,h,i}\gamma_h\beta_{i,h} - \\ &\quad - \left(\sum_h \alpha_{1,h,i}\gamma_h\beta_{i,h} \right) \sum_h \alpha_{0,h,j}\gamma_h\beta_{j,h} = 0 \\ &\iff \left(\sum_h (1 - \alpha_{h,j})\gamma_h\beta_{j,h} \right) \sum_h \alpha_{h,i}\gamma_h\beta_{i,h} - \\ &\quad - \left(\sum_h (1 - \alpha_{h,i})\gamma_h\beta_{i,h} \right) \sum_h \alpha_{h,j}\gamma_h\beta_{j,h} = 0 \\ &\iff \left(\sum_h \gamma_h\beta_{j,h} \right) \sum_h \alpha_{h,i}\gamma_h\beta_{i,h} - \\ &\quad - \left(\sum_h \gamma_h\beta_{i,h} \right) \sum_h \alpha_{h,j}\gamma_h\beta_{j,h} = 0, \qquad (1) \end{aligned}$$

which is a polynomial constraint on $P[\beta; \alpha, \gamma]$ (note that to keep the notation manageable, we have omitted the dependent term $1 - \sum_h \gamma_h$ from the equations). By a simple algebraic lemma [proven by Okamoto, 1973], if the above constraint is not trivial (that is, if there exists β for which the constraint does not hold), the subset of $P[\beta; \alpha, \gamma]$ on which it holds is measure zero.

To see that Eq. (1) does not always hold, note that if for *any* h^* we set $\beta_{i,h^*} = 1$ (and thus $\beta_{i,h} = 0$ for any $h \neq h^*$) and $\beta_{j,h^*} = 1$, the equation reduces to

$$(\gamma_{h^*})^2(\alpha_{h_i,i} - \alpha_{h_j,h}) = 0.$$

Thus if Eq. (1) was trivially true, we would have $\alpha_{h,i} = \alpha_{h,j}$ or $\gamma_h = 0$ for all h. However, this implies $C(i) = C(j)$, which contradicts our assumption.

We have now shown that the subset of $P[\beta; \alpha, \gamma]$ which consists of distributions for which $P(T \mid i) = P(T \mid j)$ (even though $C(i) \neq C(j)$) is Lebesgue measure zero. Since there are only finitely many pairs of images i, j for which $C(i) \neq C(j)$, the subset of $P[\beta; \alpha, \gamma]$ of distributions which violate the Causal Coarsening constraint is also Lebesgue measure zero. The remainder of the proof is a direct application of Fubini's theorem.

For each α, γ, call the (measure zero) subset of $P[\beta; \alpha, \gamma]$ that violates the Causal Coarsening constraint $z[\alpha, \gamma]$. Let $Z = \cup_{\alpha,\gamma} z[\alpha, \gamma] \subset P[\alpha, \beta, \gamma]$ be the set of all the joint distributions which violate the Causal Coarsening constraint. We want to prove that $\mu(Z) = 0$, where μ is the Lebesgue measure. To show this, we will use the indicator function

$$\hat{z}(\alpha, \beta, \gamma) = \begin{cases} 1 & \text{if } \beta \in z[\alpha, \gamma], \\ 0 & \text{otherwise.} \end{cases}$$

By the basic properties of positive measures we have

$$\mu(Z) = \int_{P[\alpha,\beta,\gamma]} \hat{z} \, d\mu.$$

It is a standard application of Fubini's Theorem for the Lebesgue integral to show that the integral in question equals zero. For simplicity of notation, let

$$\begin{aligned}\mathcal{A} &= \mathbb{R}^{K\times N}\\ \mathcal{B} &= \mathbb{R}^{N\times K}\\ \mathcal{G} &= \mathbb{R}^{K}.\end{aligned}$$

We have

$$\begin{aligned}\int_{P[\alpha,\beta,\gamma]} \hat{z}\; d\mu &= \int_{\mathcal{A}\times\mathcal{B}\times\mathcal{G}} \hat{z}(\alpha,\beta,\gamma)\, d(\alpha,\beta,\gamma)\\ &= \int_{\mathcal{A}\times\mathcal{G}} \int_{\mathcal{B}} \hat{z}(\alpha,\beta,\gamma)\, d(\beta)\; d(\alpha,\gamma)\\ &= \int_{\mathcal{A}\times\mathcal{G}} \mu(z[\alpha,\gamma])\; d(\alpha,\gamma) \qquad (2)\\ &= \int_{\mathcal{A}\times\mathcal{G}} 0\, d(\alpha,\gamma)\\ &= 0.\end{aligned}$$

Equation (2) follows as $\hat{z}$ restricted to $P[\beta;\alpha,\gamma]$ is the indicator function of $z[\alpha,\gamma]$.

This completes the proof that Z, the set of joint distributions over T, H and I that violate the Causal Coarsening constraint, is measure zero. □

We are now ready to prove the main theorem.

Theorem (Causal Coarsening Theorem) *Among all the generative models of the form discussed in Fig. 2 (in the main text) that have distributions $P(T,\mathbf{H},I)$ that induce some given observational partition Π_o,* almost all *induce a causal partition Π_c that is a coarsening of Π_o.*

Proof. Any variables that appear in this proof without definition are defined in the proof of the Auxiliary Theorem. We take the same α,β,γ parametrization of distributions. Fixing an observational partition means fixing a set of observational constraints (OCs)

$$\begin{gathered}P(T \mid i^1_1) = \cdots = P(T \mid i^1_{N_1}),\\ \vdots\\ P(T \mid i^L_1) = \cdots = P(T \mid i^L_{N_K}),\end{gathered}$$

where $1 \leq L \leq N$ is the number of observational classes. Since $P(T,H,I) = P(H \mid T,I)P(T \mid I)P(I)$, $P(T \mid i)$ is an independent parameter in the unrestricted $P(T,H,I)$, and the OCs reduce the number of independent parameters of the joint by $\sum_{l=1}^{L}(N_l - 1)$. We want to express this parameter-space reduction in terms of the α,β and γ parameterization and then apply the proof of the Auxiliary Theorem. To do this, for each observational class l, choose a representative image $\hat{\imath}^l$ such that

$$P(T \mid i^l_m) = P(T \mid \hat{\imath}^l) \quad \forall_{m\in 1\cdots N_k}.$$

Then for each $i^l_m \neq \hat{\imath}^l$ it holds that

$$P(T, i^l_m) = P(T \mid \hat{\imath}^l)P(i^l_m)$$

or

$$\sum_h P(T,h,i^l_m) = P(T \mid \hat{\imath}^l)\sum_h P(h,i^l_m).$$

Picking an arbitrary h_0, we can separate the left-hand side as

$$P(T,h_0,i^l_m) = P(T \mid \hat{\imath}^l)\sum_h P(h,i^l_m) - \sum_{h\neq h_0} P(T,h,i^l_m).$$

Finally, this equation can be rewritten in terms of α,β and γ as

$$\alpha_{h_0,i}\beta_{i,h_0}\gamma_{h_0} = P(T \mid \hat{\imath}^l)\sum_h \beta_{h,i^l_m}\gamma_h - \sum_{h\neq h_0} \alpha_{h,i^l_m}\beta_{i^l_m}\gamma_h,$$

or

$$\alpha_{h_0,i} = \frac{\left(P(T \mid \hat{\imath}^l)\sum_h \beta_{h,i^l_m}\gamma_h - \sum_{h\neq h_0}\alpha_{h,i^l_m}\beta_{i^l_m}\gamma_h\right)}{\beta_{i,h_0}\gamma_{h_0}}$$

for any $i^l_m \neq \hat{\imath}^l$. There are precisely $\sum_{l=1}^{L}(N_l - 1)$ such equations, altogether equivalent to the observational constraints. Thus we can express any $P(T,H,I)$ distribution that is consistent with a given observational partition in terms of the full range of β and γ parameters, and a restricted number of independent α parameters. The rest of the proof now follows similarly to the proof of the Auxiliary Theorem and shows that within this restricted parameter space, the parameters for which the (fixed) observational partition is not a refinement of the causal partition is measure zero. □

B CCT: EXAMPLES AND COUNTER-EXAMPLES

In Fig. 1 we provide examples of three distributions over binary variables H, T and three-valued I. The first model induces a causal partition that is a proper coarsening of the observational partition, and thus agrees with the CCT. The second model induces an observational partition that is a proper coarsening of the causal partition – CCT implies that this is a measure-zero case and that, after fixing the observational partition, we had to carefully tweak the parameters to align the causal partition as it is. The third model induces causal and observational partitions that are incompatible – that is, neither is a coarsening of the other. This is also a measure-zero case. We provide a Tetrad (`http://www.phil.cmu.edu/tetrad/`) file that contains these three models at `http://vision.caltech.edu/~kchalupk/code.html`. It can be used to verify our observational and causal partition computations.

C PROOF OF THE COMPLETE MACRO-VARIABLE DESCRIPTION THEOREM

Theorem (Complete Macro-variable Description) *The following two statements hold for C and S as defined in the main text:*

1. $P(T \mid I) = P(T \mid C, S)$.
2. *Any other variable X such that $P(T \mid I) = P(T \mid X)$ has Shannon entropy $H(X) \geq H(C, S)$.*

Proof. The first part follows by construction of S. For the second part, note that by the CCT there is a bijective correspondence between the pairs of values (c, s) and the observational probabilities $P(T \mid I)$. Call this correspondence f, that is $f(c, s) = P(T \mid c, s)$ and $f^{-1}(p) = (c, s \text{ s.t. } P(T|c, s) = p)$. Further, define g as the function on X, with $g \colon x \mapsto P(T \mid x)$. But since $P(T \mid X) = P(T \mid I)$, we have $(c, s) = f^{-1}(g(x))$. That is, the value of C and S is a function of the value of X, and thus the entropy of C and S is smaller than the entropy of X. □

D PREDICTIVE NON-CAUSAL INFORMATION IN CAUSAL VARIABLE C

In some cases C retains predictive information that is not causal. Consider the following example: We have a causal graph consisting of three variables $\{I, T, H\}$ where the causal relations are $I \to T$ and $I \leftarrow H \to T$. All three variables are binary and we have a positive distribution over the variables. In the general case, distributions over this graph satisfy

1. $P(T|do(I = 1)) \neq P(T|do(I = 0))$
2. $P(T|I = 1) \neq P(T|I = 0)$, and importantly
3. $P(T|I) \neq P(T|do(I))$.

If we view I as an image (which can either be all black or all white), T as the target behavior and H as a hidden confounder, analogous to the set-up in the main article, then the observational partition Π_o has just two classes, namely $\{1, 0\}$. But in this case the observational partition *is the same* as the causal partition: $\Pi_o = \Pi_c$. So

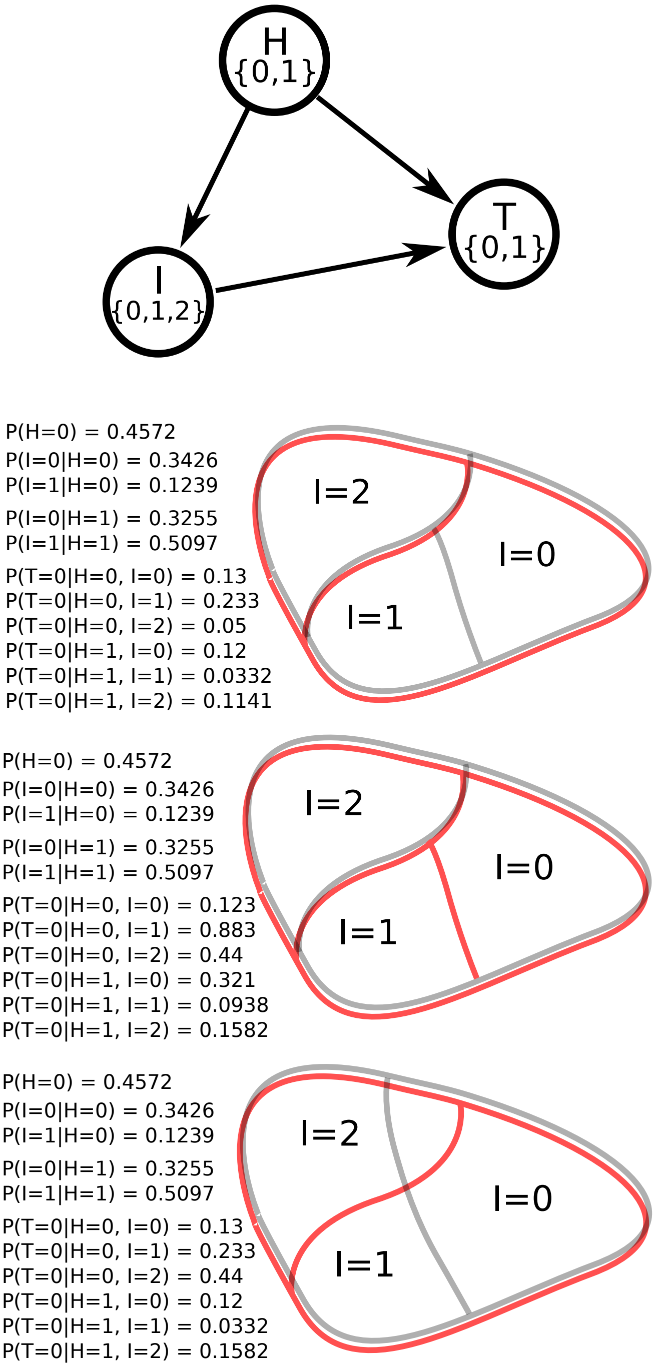


Figure 1: A graphical causal model and three faithful probability tables. The first (from the top) table induces a causal partition (red) that is a coarsening of the observational partition (gray) – specifically, as the figure shows, $P(T|I = 0) \neq P(T|I = 1)$ but $P(T|man(I = 0)) = P(T|man(I = 1))$. The second table induces an observational partition that is a coarsening of the causal partition. The last table induces a causal and an observational partition such that neither is a coarsening of the other.

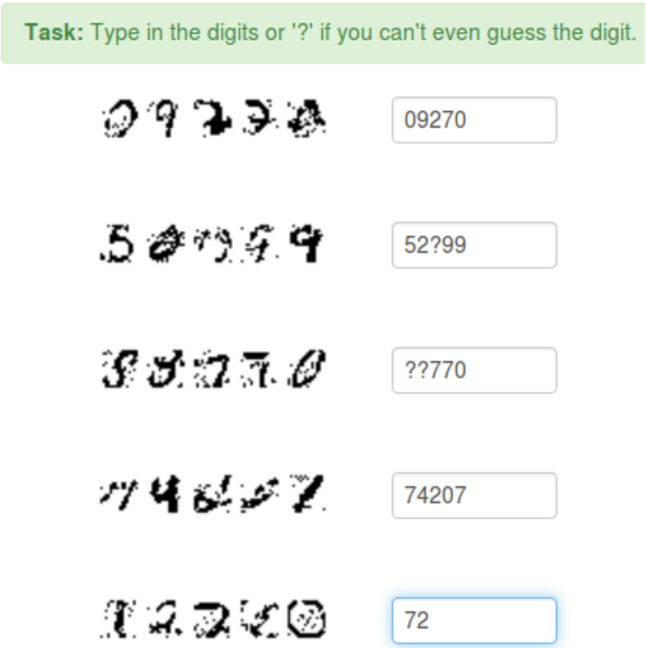
by our definition of a spurious correlate, S is a constant, since there are no further distinctions to be made within any of the causal classes. S would be omitted from any standard causal model. Nevertheless, we have in our model still that $P(T|C) \neq P(T|do(C))$, i.e. the causal variable C still contains predictive information that is not causal. Given that there is by construction no other than the causal and the trivial partition in this example, it must be the case that C retains predictive non-causal information. It follows that in our definitions of C and S, it is not the case that the predictive non-causal components of an image can always be completely separated from the causal features.

E THE MNIST ON MTURK EXPERIMENT

For this experiment, we started off by training ten one-vs-all neural nets. We used cross-validation to choose among the following architectures: 100 hidden units (h.u.), 300 h.u. (one layer), 100-100 h.u (two layers), 300-300 h.u. (two layers). We used maxout [Goodfellow and Warde-Farley, 2013] activations (each of which computed the max of 5 linear functions). For training we used stochastic gradient descent in batches of 50 with 50% dropout [Hinton and Srivastava, 2012] on the hidden units, momentum adjustment from 0.5 to 0.99 at iteration 100, learning rate decaying from 0.1 to 0.0001 with exponential coefficient of 1/0.9998, no weight decay, and we enforced the maximum norm of a column of hidden units to 5. The training stopped after 1000 iterations and the iteration with best validation error was chosen. We used the Pylearn2 package [Goodfellow and Warde-Farley, 2013] to train the networks.

This initial training was done on 5000 training points and 1250 validation points (both of which come from the MNIST dataset) for each machine. The training points were chosen at random to include 2500 images of a specific digit class (that is, 2500 zeros for the first machine, 2500 ones for the second machine and so on), and 2500 images of random other digits for each machine. The validation sets were composed similarly. Each machine then used Algorithm 2 to transform 1000 images of digits *from its training set* into maximally similar images of the opposing class.

We thus started off with ten manipulated datasets of 1000 images each. The first dataset contained images of zeros manipulated to be non-zeros, and all the other digits manipulated to be zeros. The tenth dataset contained images of nines manipulated to be non-nines and the other digits manipulated to be nines. We then used Amazon Mechanical Turk to present all those images to human annotators, using the interface shown in Fig. 2. The images created by all the manipulator networks were mixed at random together, so that each single annotator (annotating 250 images in one task) would see some images created by each machine. Finally, each of the 10000 images was shown to five annotators; we used 5×40=200 annotators total on each iteration. The annotators labeled the images as either one of the ten digits, or the question mark '?' if there was no recognizable digit in an image. The final label ("target digit" or "not target digit") was chosen using majority of the annotators' votes.


Task: Type in the digits or '?' if you can't even guess the digit.

09270

52?99

??770

74207

72

Figure 2: The Amazon Mechanical Turk interface we used to query online annotators. An annotator is shown five rows of five manipulated digit images, and is requested to type the digit labels (or '?') into the input boxes. Each annotator goes through ten similar screens, annotating a total of 250 digits.

The annotated manipulated digits were then added to the datasets which their respective original images belonged to. We then proceeded to train the next iteration of neural network manipulators on the updated datasets, and so on until completion of the manipulator training.

References

I. J. Goodfellow and D. Warde-Farley. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013.

G. E. Hinton and N. Srivastava. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

C. Meek. Strong completeness and faithfulness in Bayesian networks. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 411–418, 1995.

M. Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *The Annals of Statistics*, 1(4):763–765, 1973.