A. Additional results.

A.1. Too many or too few experts.

Too many experts When there are too many experts, for most tasks only one wins all the examples, as shown in Figure 9 where the model has 16 experts for 10 tasks. In this case the remaining experts do not specialize at all and therefore can be removed from the architecture. Had several experts specialized on the same task, they could be combined after determining that they perform the same task. Since the accuracy on the transformed data tested on the pretrained classifier reaches again the upperbound of the untransformed data, and since the progress is very similar to that illustrated in Figure 6, we omit this plot.

Too few experts For a committee of 6 experts, the networks do not reconstruct properly most of the digits, which is reflected by an overall low objective function value on the data. Also, the accuracy achieved by the pretrained MNIST classifier does not exceed 72%. A few experts are inevitably assigned to multiple tasks, and by looking at Figure 9 it is interesting to see that the clustering result is still meaningful (e.g. expert 5 is assigned to left, down-left, and up-left translation).

B. Details of neural networks

In Table 1 we report the configuration of the neural networks used in these experiments.

For the approximate identity initialization we train each network for a maximum of 500 iterations, or until the mean squared error of the reconstructed images is below 0.002.

C. Transformations

In our experiments we use the following transformations

- Translations: the image is shifted by 4 pixels in one of the eight directions up, down, left, right and the four diagonals.
- Contrast (or color) inversion: the value of each pixel — originally in the range $[0, 1]$ — is recomputed as $1-$ the original value.
- Noise addition: random Gaussian noise with zero mean and variance 0.25 is added to the original image, which is then clamped again to the $[0, 1]$ interval.

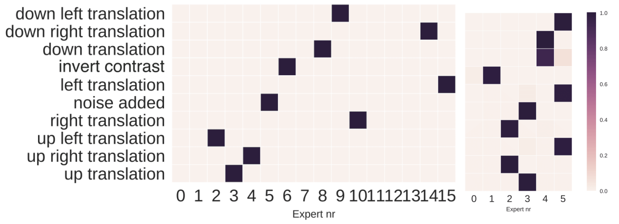

Figure 9. The proportion of data won by each expert for each transformation on the digits from the test set, for the case of 10 mechanisms and more experts (16 on left) or too few (6 on the right). Note how on the left experts 0, 1, 7, 11, 12, 13, do not win any data points, and can therefore be discarded.

Table 1. Architectures of the neural networks used in the experiment section. BN stands for Batch normalization, FC for fully connected. All convolutions are preceded by a 1 pixel zero padding.

Expert
Layers
$3 \times 3, 32$, BN, ELU
$3 \times 3, 32$, BN, ELU
$3 \times 3, 32$, BN, ELU
$3 \times 3, 32$, BN, ELU
$3 \times 3, 1$, sigmoid

Discriminator
Layers
$3 \times 3, 16$, ELU
$3 \times 3, 16$, ELU
$3 \times 3, 16$, ELU
2×2, avg pooling
$3 \times 3, 32$, ELU
$3 \times 3, 32$, ELU
2×2, avg pooling
$3 \times 3, 64$, ELU
$3 \times 3, 64$, ELU
2×2, avg pooling
1024, FC, ELU
1, FC, sigmoid

D. Notes on the Formalization of Independence of Mechanisms

In this section we briefly discuss the notion of independence of mechanisms as in Janzing & Schölkopf (2010), where the independence principle is formalized in terms of algorithmic complexity (also known as Kolmogorov complexity). We summarize the main points needed in the present context. We parametrize each *mechanism* by a bit string x. The Kolmogorov complexity $K(x)$ of x is the length of the shortest program generating x on an a priori chosen universal Turing machine. The **algorithmic mutual information** can be defined as $I(x : y) := K(x) + K(y) - K(x, y)$, and it can be shown to equal

$$I(x : y) = K(y) - K(y|x^*), \tag{4}$$

where for technical reasons we need to work with x^*, the shortest description of x (which is in general uncomputable). Here, the conditional Kolmogorov complexity $K(y|x)$ is defined as the length of the shortest program that generates y from x. The algorithmic mutual information measures the algorithmic information two objects have in common. We define two mechanisms to be **(algorithmically) independent** whenever the length of the shortest description of the two bit strings together is not shorter than the sum of the shortest individual descriptions (note it cannot be longer), i.e., if their algorithmic mutual information vanishes.[6] In view of (4), this means that

$$K(y) = K(y|x^*). \tag{5}$$

We will say that two mechanisms x and y are independent whenever the complexity of the conditional mechanism $y|x$ is comparable to the complexity of the unconditional one y. If, in contrast, the two mechanisms were closely related, then we would expect that we can mimic one of the mechanisms by applying the other one followed by a low complexity conditional mechanism.

[6] All statements are valid up to additive constants, linked to the choice of a Turing machine which produces the object (bit string) when given its compression as an input. For details, see Janzing & Schölkopf (2010).