Exercise 11

Combining Logic and Deep Learning

Reliable and Interpretable Artificial Intelligence
ETH Zurich

Problem 1 (Interpreting Queries)**.** Describe (in words) the objective of the following queries. Here, `ref` is an image from the test set, and `N`, $\texttt{N}_1$, $\texttt{N}_2$ are neural networks with two output classes 1 and 2 such that `class(N(ref)) = 1`.

1.
```
find i[10,10]
where i in [0,1],
      ‖i - ref‖∞ < 0.1,
      ‖i - ref‖∞ > 0.05,
      class(N(i)) = 2
```

2.
```
find i[10,10]
where i in [0,1],
      ‖i‖∞ < 0.2,
      class(N(i)) = 1
```

3.
```
find i[10,10]
where i in [0,1],
      ‖i - ref‖2 < 2,
      class(N1(i)) = 1,
      class(N2(i)) = 2
```

Problem 2 (Translating Negations)**.** In this question, we will inspect how to support negation ($\neg$) within constraints.

1. Translate the following constraint φ to a loss function using the rules discussed in the lecture. Here, `i` is a 2 by 2 pixel query image and `ref` is a given 2 by 2 pixel image from the test set.

   φ := `(i[0,0] = ref[0,0] ∧ i[1,0] ≠ ref[1,0]) ∨`
   `(i[0,0] ≤ ref[0,0] ∧ i[1,0] < ref[1,0])`

2. Describe a way to transform a constraint involving negation ($\neg$) to a constraint that lies in the fragment discussed in the lecture (e.g., a constraint that only uses the operations described on lecture slide 20).

3. Transform the constraint $\neg\varphi$ to a constraint not involving negation.

Problem 3 (Alternative Translation)**.** In the lecture, we studied one particular way to translate constraints to nonnegative loss functions. Consider the following alternative translation T, which also produces nonnegative loss functions:

ω	$T(\omega)$
$t_1 = t_2$	$(t_1 - t_2)^2$
$t_1 \leq t_2$	$\max(\text{sgn}(t_1 - t_2) \cdot (t_1 - t_2)^2,\ 0)$
$\phi \vee \psi$	$T(\phi) \cdot T(\psi)$
$\phi \wedge \psi$	$T(\phi) + T(\psi)$

Further, consider the formula

$$\psi := (\text{ReLU}(x_1 + 2x_2) = x_3 \ \wedge \ x_3 \leq 4) \ \vee \ (x_3 \leq 0 \ \wedge \ x_1 + x_2 \geq 0),$$

which has free variables x_1, x_2, x_3. We denote the set of free variables as $\mathbf{x}$, and the assignment to these variables $x_1 \leftarrow y_1, \ldots, x_3 \leftarrow y_3$ as $\mathbf{y}$. The translation of ψ according to T is denoted $T(\psi)$ and the numerical value of the translation evaluated for assignment $\mathbf{y}$ is indicated by $T(\psi)(\mathbf{x} \leftarrow \mathbf{y})$.

1. Derive the translation $T(\psi)$ of formula ψ.

2. Prove that for any assignment $\mathbf{y}$, $T(\psi)(\mathbf{x} \leftarrow \mathbf{y}) = 0$ implies that $\mathbf{y}$ is a satisfying assignment of ψ.