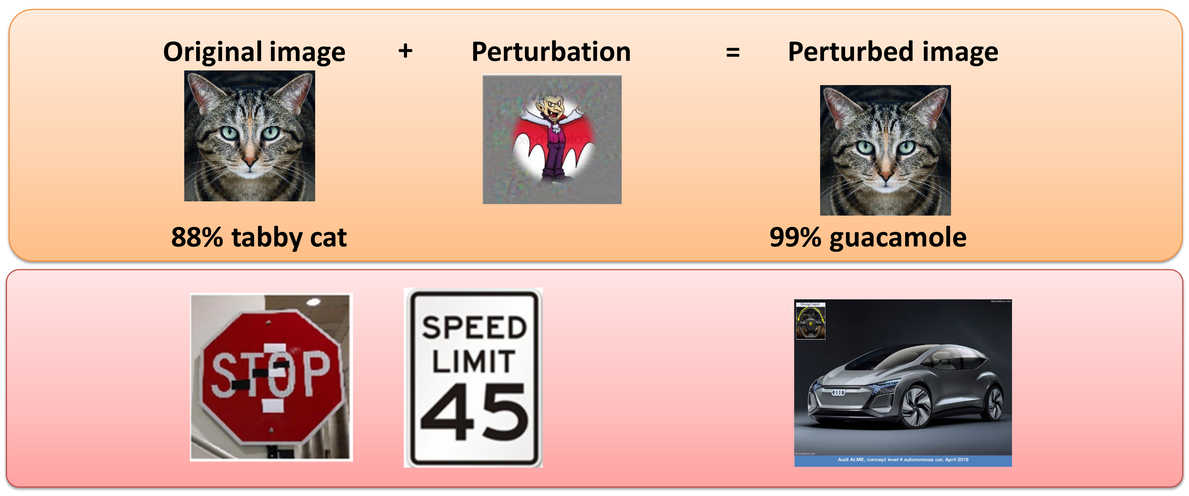
On Relating Explanations and Adversarial Examples

Alexey Ignatiev, Nina Narodytska, and Joao Marques Silva

Adversarial examples and explanations in neural networks

[Szegedy et al., Goodfellow et al., Ribeiro et al.]

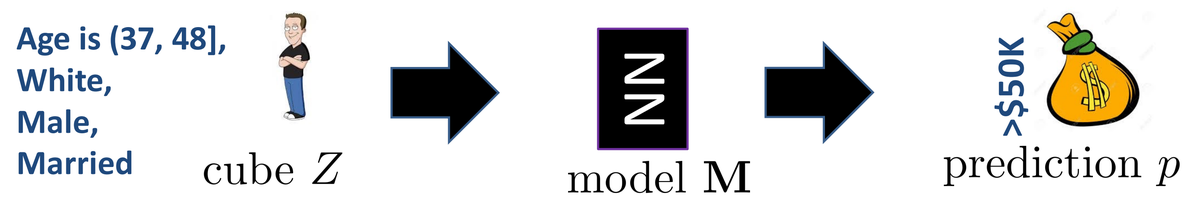


Is there a connection between adversarial examples and explanations?

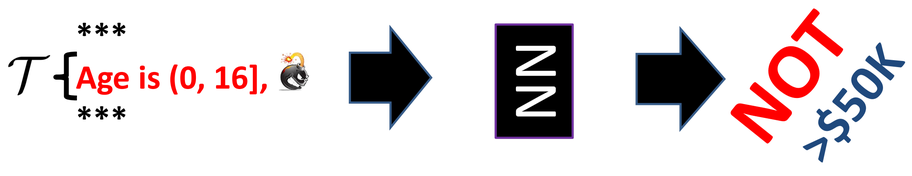
Logic-based approach to attacks and explanations

[Ignatiev et al., Shih et al.]

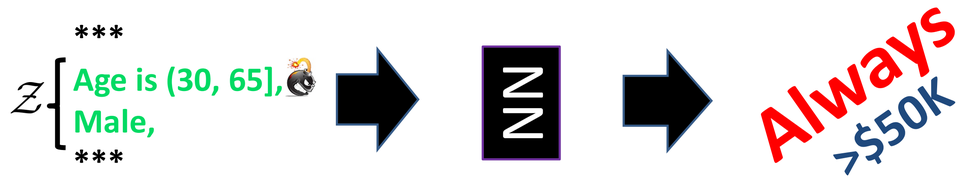
Given a classifier $\mathbf{M}$ (in a logic encoding) and a prediction p,



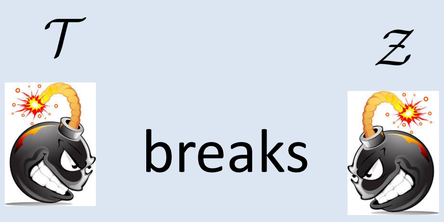
a counterexample to p is
a subset-minimal $\mathcal{T}$, s.t. $\mathcal{T} \models \vee_{t,t\neq p}(\mathbf{M} \rightarrow t)$



an explanation of p is
a subset-minimal $\mathcal{Z}$ s.t. $\mathcal{Z} \models (\mathbf{M} \rightarrow p)$



Duality between counterexamples and explanations



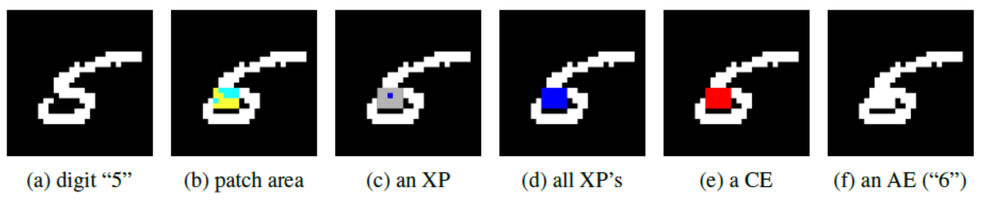
- every explanation $\mathcal{Z}$ of p *breaks every* counterexample of p, and
- every counterexample $\mathcal{T}$ of p *breaks every* explanation of p.

Duality-based computation of all explanations (or all counterexamples).

Input: formula $\mathbf{M}$ and prediction p
Output: set $\mathbb{E}$ of all absolute explanations of prediction p

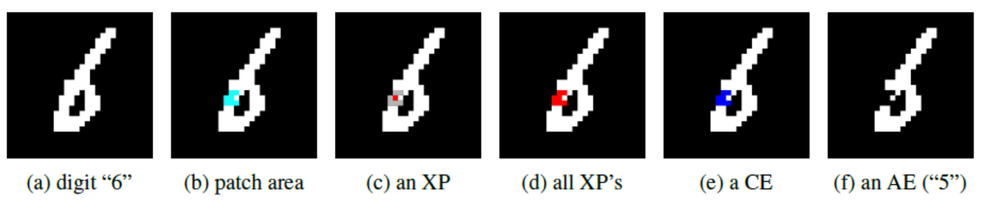
```
1  (𝕋, 𝔼, 𝒵) ← (∅, ∅, ∅)
2  do:
3      if 𝒵 ⊨ (M → p) :
4          𝔼 ← 𝔼 ∪ {𝒵}                # 𝒵 is an explanation; save it
5      else:
6          (𝒯, t) ← ExtractInstance()   # get an instance 𝒯 with a
             prediction t, t ≠ p
7          for l ∈ 𝒯 :
8              if (𝒯 \ {l}) ⊨ (M → t) :
9                  𝒯 ← 𝒯 \ {l}
10         𝕋 ← 𝕋 ∪ {𝒯}              # update 𝕋 with a new counterexample 𝒯
11     ℰ ← MinimumHS(𝕋)             # get a new hitting set of 𝕋
12 while 𝒵 ≠ ∅
13 return 𝔼
```

Algorithm 1: Duality-based computation of all absolute explanations



(a) digit "5" (b) patch area (c) an XP (d) all XP's (e) a CE (f) an AE ("6")

Figure 1: An example of digit five.



(a) digit "6" (b) patch area (c) an XP (d) all XP's (e) a CE (f) an AE ("5")

Figure 2: An example of digit six.

[Goodfellow et al.] *Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In ICLR, 2015.*

[Szegedy et al.] *Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In ICLR, 2014.*

[Ribeiro et al.] *Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. Why Should I Trust You?" Explaining the Predictions of Any Classifier. In KDD, 2016.*

[Ignatiev et al.] *Alexey Ignatiev, Nina Narodytska and Joao Marques-Silva. Abduction-Based Explanations for Machine Learning Models. In AAAI, 2019.*

[Shih et al.] *Andy Shih and Arthur Choi and Adnan Darwiche . Abduction-Based Explanations for Machine Learning Models. In AAAI, 2019.*

[Narodytska et al.] *Nina Narodytska, Shiva Prasad Kasiviswanathan, Leonid Ryzhyk, Mooly Sagiv, Toby Walsh. Verifying Properties of Binarized Deep Neural Networks. In AAAI, 2018.*