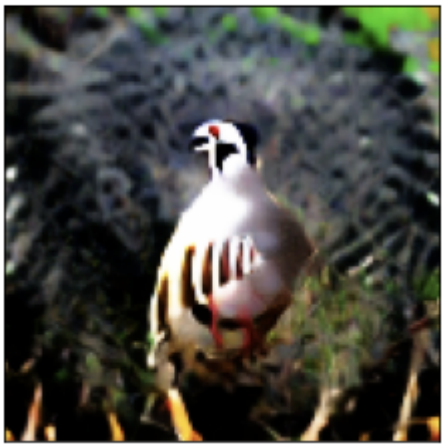
First, have a look at these 11 universal patch adversaries.

They are made in the same way as those in the top row of fig. 4. We trained 50 universal adversarial patches with random source/target classes and selected the 11 of them that had average fooling confidences higher than 0.4.

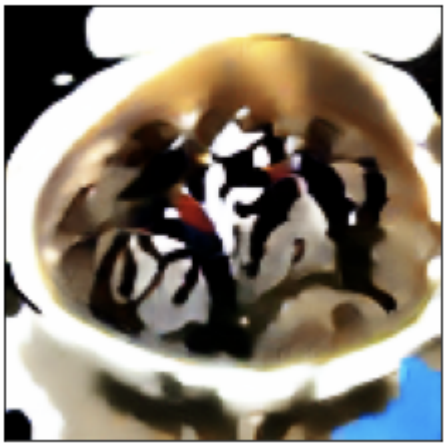
Without knowing them in advance, what do you think the **target** and **disguise** classes might be? Grab a scratch paper and list down some possibilities for each of the 11.

1. 

2. 

3. 

4. 

5. 

6. 

7. 

8. 

9. 

10. 

11. 

# ANSWERS ON NEXT PAGE

# How often could you predict the target class versus the disguise class?

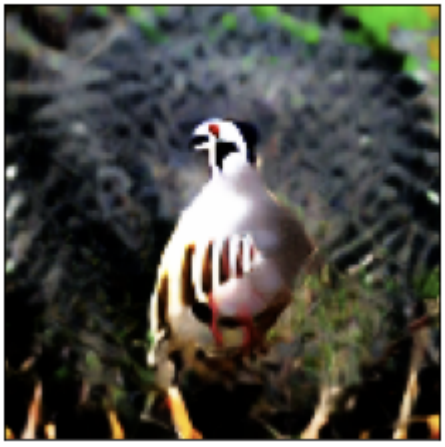
“As patch” = target class

“As image” = disguise class

1. as patch: neck brace, conf: 0.8431
as img: Japanese spaniel, conf: 0.8109

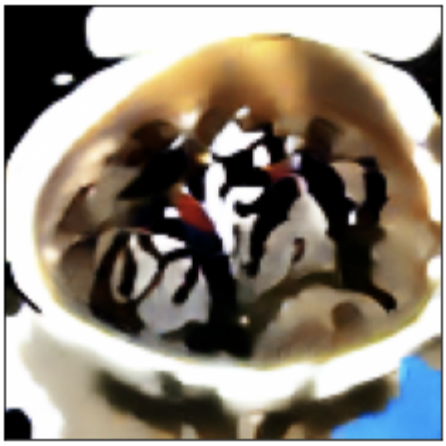
2. as patch: briard, conf: 0.4933
as img: dalmatian, conf: 0.9872

3. as patch: sloth bear, conf: 0.4404
as img: partridge, conf: 0.7247



4. as patch: ant, conf: 0.4273
as img: fur coat, conf: 0.377

5. as patch: Pekinese, conf: 0.6999
as img: chocolate sauce, conf: 0.9543



6. as patch: cleaver, conf: 0.5282
as img: bulbul, conf: 0.9998

7. as patch: cocker spaniel, conf: 0.5715
as img: jacamar, conf: 0.9981

8. as patch: centipede, conf: 0.6523
as img: holster, conf: 0.9654

9. as patch: snow leopard, conf: 0.4404
as img: agama, conf: 0.9995

10. as patch: Airedale, conf: 0.6344
as img: hot pot, conf: 0.6479

11. as patch: Brittany spaniel, conf: 0.4546
as img: football helmet, conf: 0.9999