

Visualización

Interpretación de la red

Verónica E. Arriola-Rios

Facultad de Ciencias, UNAM

26 de marzo de 2023



Oclusión de regiones

- 1 Oclusión de regiones
- 2 Optimización de la entrada
- 3 Optimización con ejemplares

¿Qué aprendió la red?

- Un problema frecuente con las redes es que las respuestas aparentemente correctas podrían corresponder al aprendizaje de conceptos equivocados.

Caso:

- *En una ocasión se entrenó una red para distinguir marcas de coches. La red obtuvo un desempeño perfecto.*
- *Posteriormente se descubrió que todos los coches de una marca estaban en fondo negro y los otros en fondo blanco.*

Atribución

- “La *Atribución* estudia qué parte de un ejemplar es responsable de que la red se active de una forma en particular.” Olah, Mordvintsev y Schubert 2017

Técnica de evaluación

- En una red para clasificación de imágenes entrenada, introducir las fotografías con un parche: una región en ceros.
- Graficar la probabilidad de la clase correcta vs la posición del parche.
- Si la probabilidad es mínima cuando el parche cubre al objeto a clasificar es indicación de que la red está observando el objeto correcto.

<https://cs231n.github.io/understanding-cnn/>

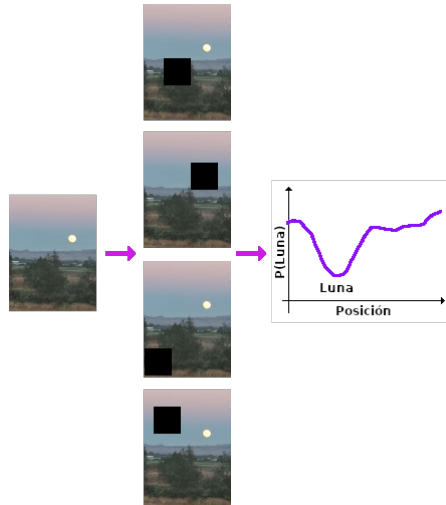


Figura: La probabilidad de la clase correcta debe ser menor cuando esta no aparece en la imagen.

Optimización de la entrada

- 1 Oclusión de regiones
- 2 Optimización de la entrada
- 3 Optimización con ejemplares

Visualización de características

- “La *Visualización de características* responde a la pregunta sobre lo que busca una red (o partes de una red) mediante la generación de ejemplos” Olah, Mordvintsev y Schubert 2017

Optimización

- La técnica de visualización por optimización busca generar imágenes que maximizan la activación de algún elemento de la red, de modo que se identifique a qué patrones reacciona ese elemento.
- Se puede optimizar la activación de:
 - Una neurona
 - Un canal
 - Una capa completa (*DeepDream*)
 - Las *logits* (valor de activación antes de *softmax*).
 - La probabilidad de la clase.
- La solución no es única, como en toda optimización numérica, pueden existir varios óptimos en diferentes regiones del espacio. Una forma de explorarlo es tomar como puntos iniciales diferentes imágenes del conjunto de datos.

De avión a camión

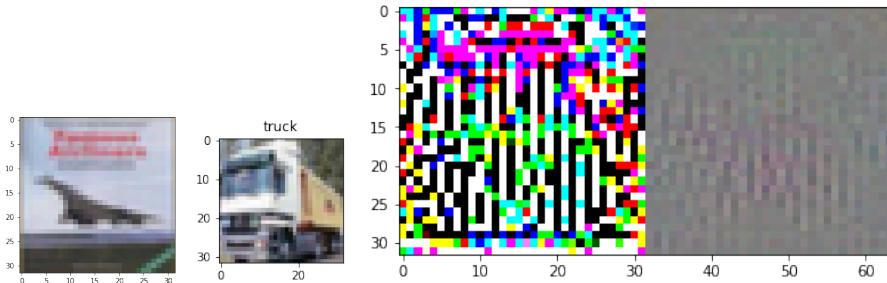


Figura: Ejemplos de avión y camión en CIFAR-10. Entrada optimizada para activar a la neurona que reacciona con imágenes de camiones, a partir de un ejemplar de avión, el gradiente en la última iteración se muestra a la derecha.

Neurona pájaro

Optimizar la activación inocentemente producirá imágenes con poco significativo semántico, creando *ilusiones neurales*.

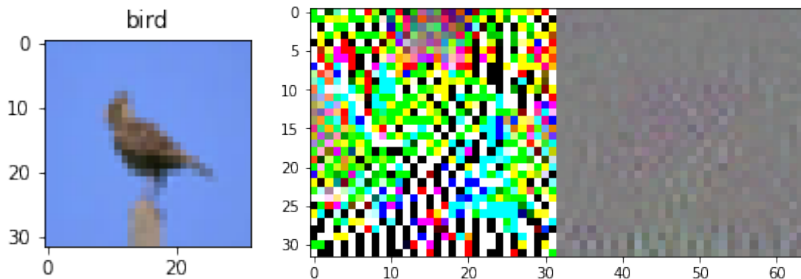


Figura: Ejemplo de pájaro en CIFAR-10. Entrada optimizada para activar a la neurona que reacciona con imágenes de pájaros a partir de ruido, el gradiente en la última iteración se muestra a la derecha.

Técnicas de regularización en el gradiente I

Penalización de la frecuencia Penaliza el ruido de alta frecuencia aplicando operaciones que lo reduzcan. Por ejemplo: desenfocando la imagen después de cada paso de optimización.

Robustez ante transformaciones Se buscan ejemplos que activen al objetivo aún después de haber sido transformados. La imagen se rota, escala o sus píxeles se alteran aleatoriamente antes del paso de optimización.

Aprendizaje de distribuciones a priori Se busca optimizar la activación únicamente sobre ejemplares tomados de la distribución de probabilidad que describe a las imágenes del conjunto de datos. Es necesario aprender a modelar esta distribución.

Técnicas de regularización en el gradiente II

Precondicionamiento Realizar la optimización con los parámetros en otro espacio. Los puntos críticos se conservan, pero el paisaje se modifica alterando la dirección y velocidad del proceso de optimización. Por ejemplo: se puede optimizar sobre la transformada de Fourier de la imagen (espectro de frecuencias).

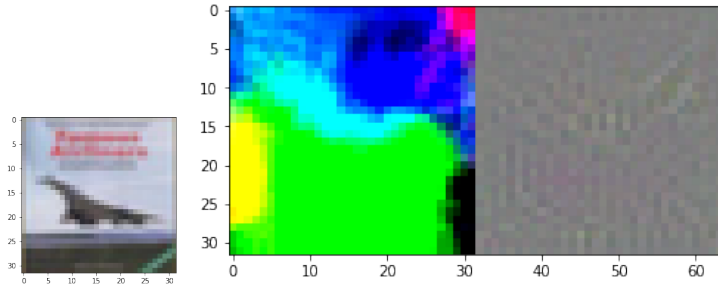


Figura: Ejemplo de avión en CIFAR-10. Entrada optimizada para activar a la neurona que reacciona con imágenes de aviones a partir de ruido, utilizando desenfoque (*blur*) entre optimizaciones.

Optimización con ejemplares

- 1 Oclusión de regiones
- 2 Optimización de la entrada
- 3 Optimización con ejemplares

La neurona de Jennifer Aniston



Figura: La Neurona de Jennifer Aniston fue descubierta en 2005 por el neurocientífico argentino Rodrigo Quian Quiroga, de la Universidad de Leicester (Reino Unido)^[1]



[1] <https://www.muyinteresante.es/curiosidades/20506.html>

Estrategia

- “Se puede interpretar que la redes neuronales convolucionales transforman a las imágenes gradualmente hacia una representación en la que las clases son separables mediante un filtro lineal.” [2]
- Es posible evaluar la red sobre un gran conjunto de imágenes, obteniendo el vector de valores de activación de la última capa (antes de softmax en caso de haberla)
- Luego se utiliza algún algoritmo de reducción de dimensionalidad que mapee ese vector al espacio 2D.
- Finalmente se coloca la imagen que produjo al vector en la posición que le fue asignada por el algoritmo.
- Dos ejemplos de algoritmos serían el mapeo autoorganizado o t-SNE. Éste último se usa mucho en el área.
- Sólo tiene la desventaja de estar completamente sesgado a fotografías de cosas existentes.

[2] <https://cs231n.github.io/understanding-cnn/>

Referencias I

-  Olah, Chris, Alexander Mordvintsev y Ludwig Schubert (nov. de 2017). *Feature Visualization* How neural networks build up their understanding of images. English. Google. URL: <https://distill.pub/2017/feature-visualization/>.
-  Stanford University (2022). *Convolutional Neural Networks for Visual Recognition*. English. Stanford University. URL: <https://cs231n.github.io/convolutional-networks/>.

Licencia

Creative Commons
Atribución-No Comercial-Compartir Igual

