EN VIVO

FEMT CONF 2020

Organiza:

[LAS] DE SISTEMAS

¿Dónde está Wally?: desarrollando modelos de predicción de la visión al buscar objetos

Melanie Sclar

Su último premio en competencias fue en la Google Codejam to I/O for Women 2019, donde ganó un viaje todo pago a Google I/O. No tengan miedo de anotarse!!!

Melanie Sclar

Lic. en Cs. de la Computación de la Universidad de Buenos Aires, con fuerte foco en la formación matemática. Realizó su tesis en modelos de búsqueda visual.

Multipremiada en olimpíadas de matemática e informática a nivel internacional: campeona nacional en la olimpiada matemática argentina, medalla de bronce en la olimpiada iberoamericana de matemática, campeona latinoamericana en la competencia mundial de programación para universitarios (ICPC), entre otros. Actual miembro del jurado de la Olimpíada Informática Argentina y organizadora del Torneo Argentino de Programación.

Melanie trabajó en Facebook, obteniendo dos patentes por su trabajo. También fue Lead Data Scientist en BrightSector y actualmente es Lead Machine Learning Engineer en ASAPP.

Búsqueda visual: qué es y por qué es importante?

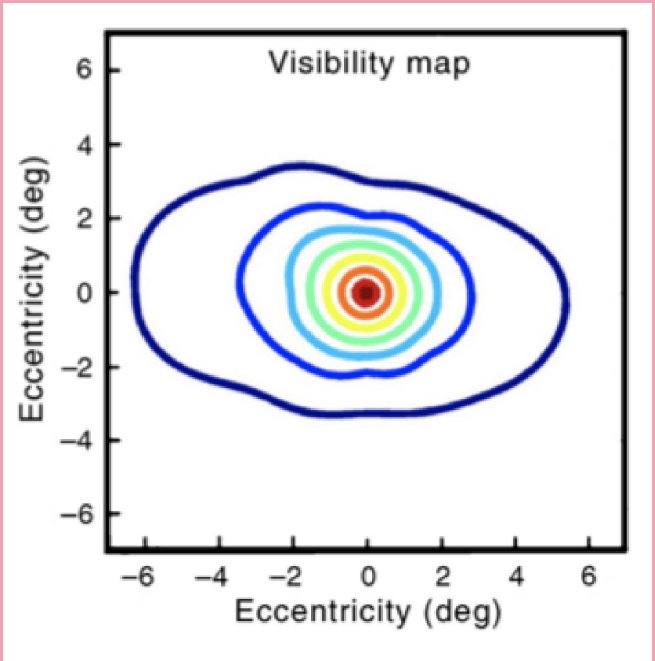
Los humanos buscan objetos intuitiva y eficientemente. Qué estrategias usan?

- Uno de los desafíos centrales de la ciencia cognitiva y de la neurociencia de la visión es entender cómo percibimos una escena visual (y por ende, el mundo que nos rodea)
- Hace más de medio siglo que se sabe que los lugares a los que miramos no son aleatorios, sin embargo todavía no existe una teoría formal de la interacción entre la visión y los movimientos oculares
- **Queremos diseñar algoritmos que puedan simular los movimientos oculares en una búsqueda visual, como modelos del comportamiento**

Cómo funciona el ojo humano?

- Solamente podemos ver en detalle una pequeña porción del campo visual, aquella que cae en un área de la retina llamada fóvea.

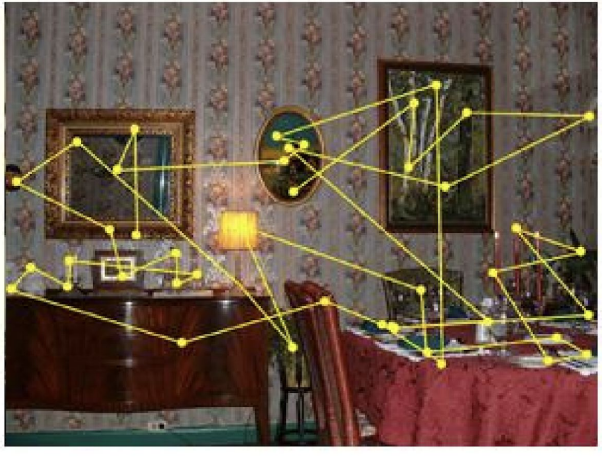
- **Debido a las limitaciones de agudeza en la retina, los movimientos oculares son necesarios para procesar detalles.**



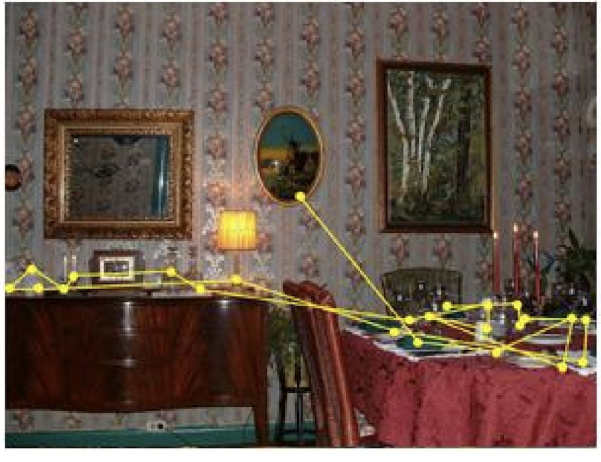
Búsqueda visual: diferencias con observación libre

Los recorridos de la mirada varían mucho según el objetivo que tengamos en mente

Memorization Task

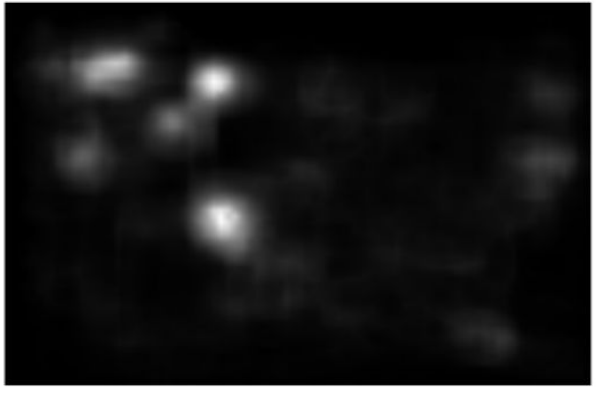


Visual Search Task



Modelos de saliencia

- Como no podemos procesar toda la información de una imagen a la vez, restringimos nuestra atención a un área reducida.
- **¿Cómo decidimos qué región seleccionar para captar nuestra atención primero?**



Ver más en https://saliency.tuebingen.ai/

Búsqueda visual: bottom-up (saliencia) + top-down (contexto)

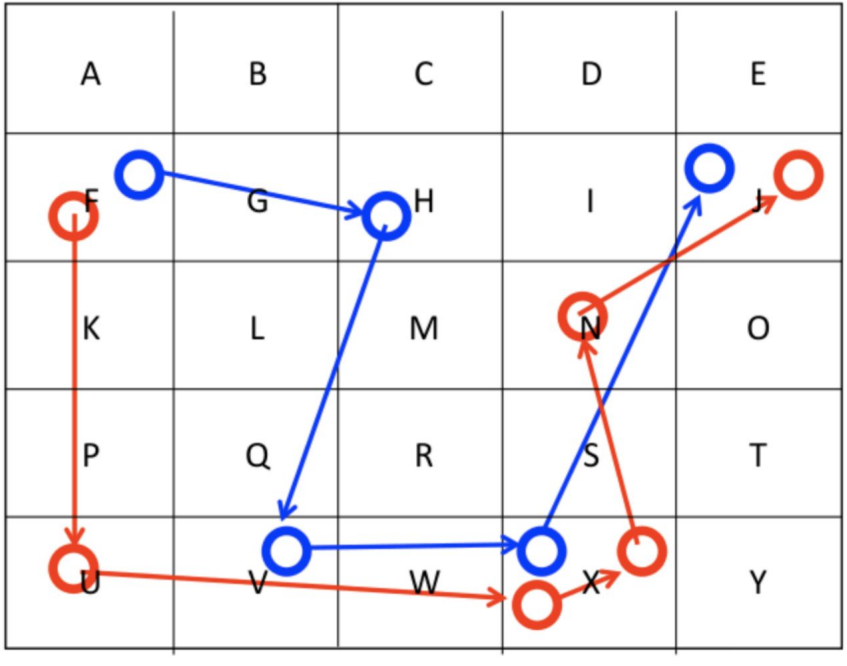
- En una búsqueda visual a veces se ignoran regiones muy salientes porque son poco relevantes para la tarea.
- La saliencia suele funcionar mejor en escenas artificiales que en naturales.
- En contextos naturales hay que combinar saliencia con elementos que describan el target.

Métricas de comparación de scanpaths (recorrido de la mirada)

- No hay consenso sobre qué métrica utilizar para comparar scanpaths, ya que hay muchos factores a tener en cuenta. Así, algunas de las más usadas son:

    1. Cantidad de fijaciones hasta encontrar el target
    2. Ángulos entre fijaciones
    3. String edit distance

Métricas de comparación de scanpaths: edit distance



F U X X N J

F H V X J

Creación del corpus de imágenes y datos humanos

Cómo podemos juntar datos de búsqueda visual?

Todo comenzó con Yarbus en 1967...

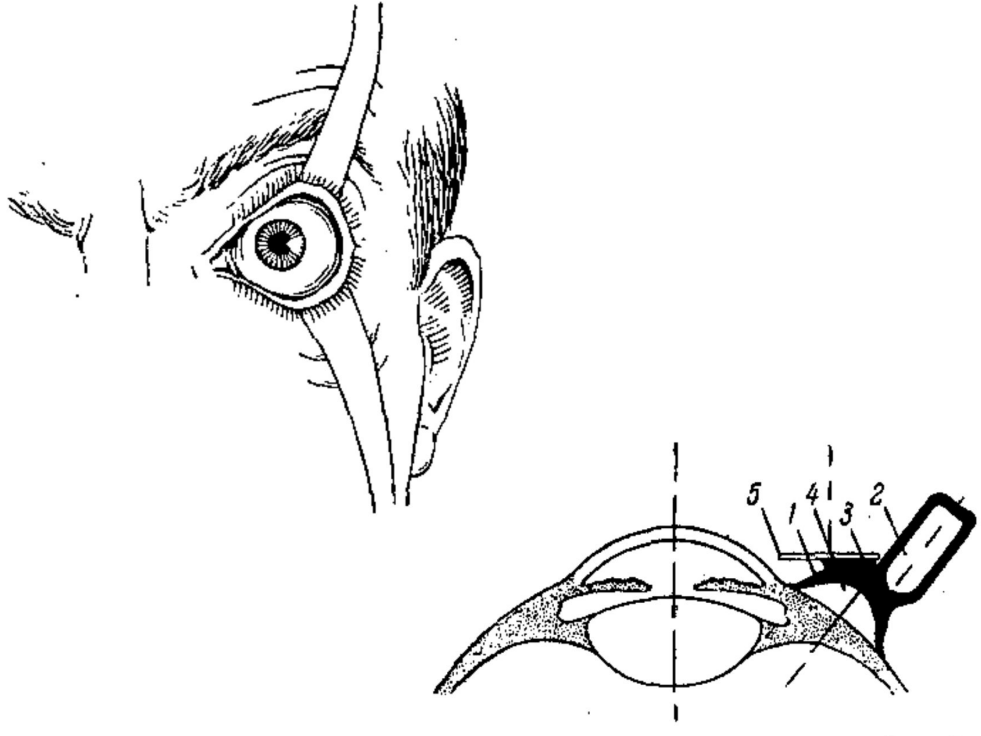


Fig. 13. The P_1 suction device or "cap."

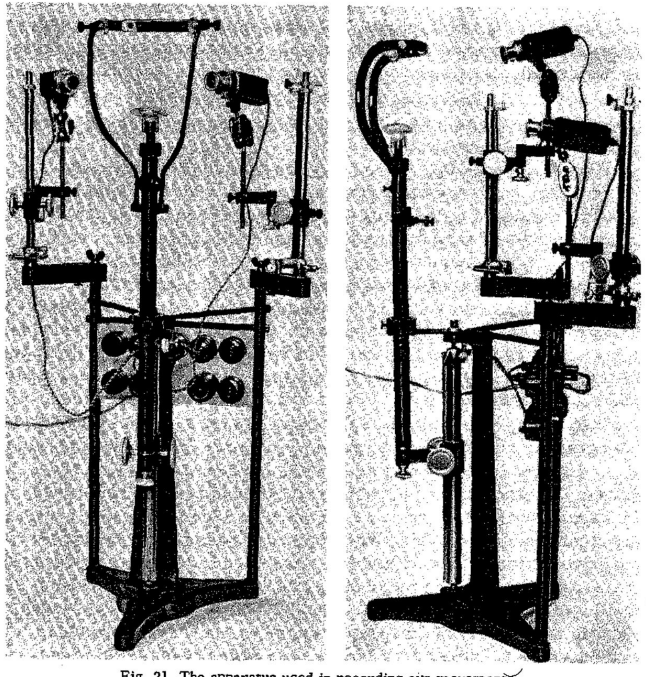
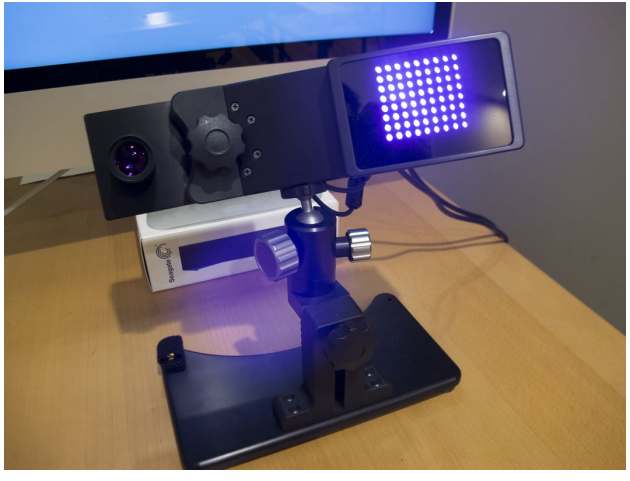


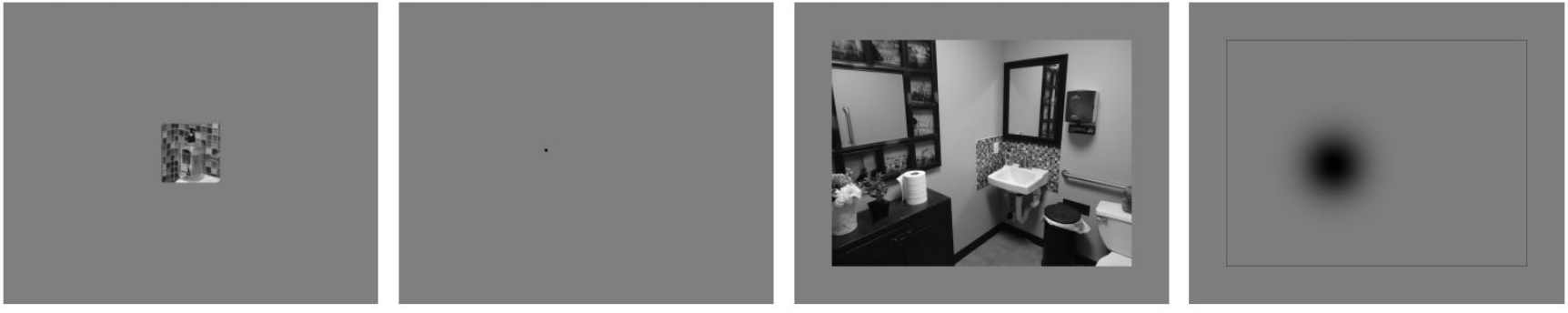
Fig. 21. The apparatus used in recording eye movements.

Cómo podemos juntar datos de búsqueda visual?

... por suerte ahora tenemos el EyeLink (y más!)



Experimento y toma de datos



(a) Muestra de target (b) Fijación forzada (c) Imagen completa (d) Respuesta subjetiva

Se fuerza a los sujetos a mirar en un punto negro a 300 píxeles del target. Este lugar varía por imagen pero es el mismo para todos los sujetos. Al final les preguntamos dónde estaba el objeto y qué nivel de seguridad tienen (no vamos a hablar de eso hoy).

Experimento y toma de datos

- Tomamos datos de 57 sujetos, en 134 imágenes llenas de objetos para buscar
- Las 134 fueron seleccionadas a propósito para que tengan muchos objetos, y dentro de ellas se recortó un objeto de 72 x 72 píxeles que era el elemento a buscar
- En cada imagen se permitía un número fijo de lugares adonde mirar (“fijaciones”) que las personas no sabían de antemano

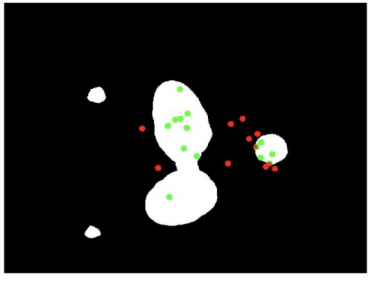
Predecir la mirada solo usando mapas de saliencia

Medidas de performance (TPR)

- Tomamos los mapas de saliencia como un clasificador binario de los píxeles. Un porcentaje dado de los píxeles de una imagen se clasifican como positivos y el resto negativos.

- Se grafica para varios porcentajes el True Positive Rate (TPR), que indica qué porción de las fijaciones caen en la región saliente.

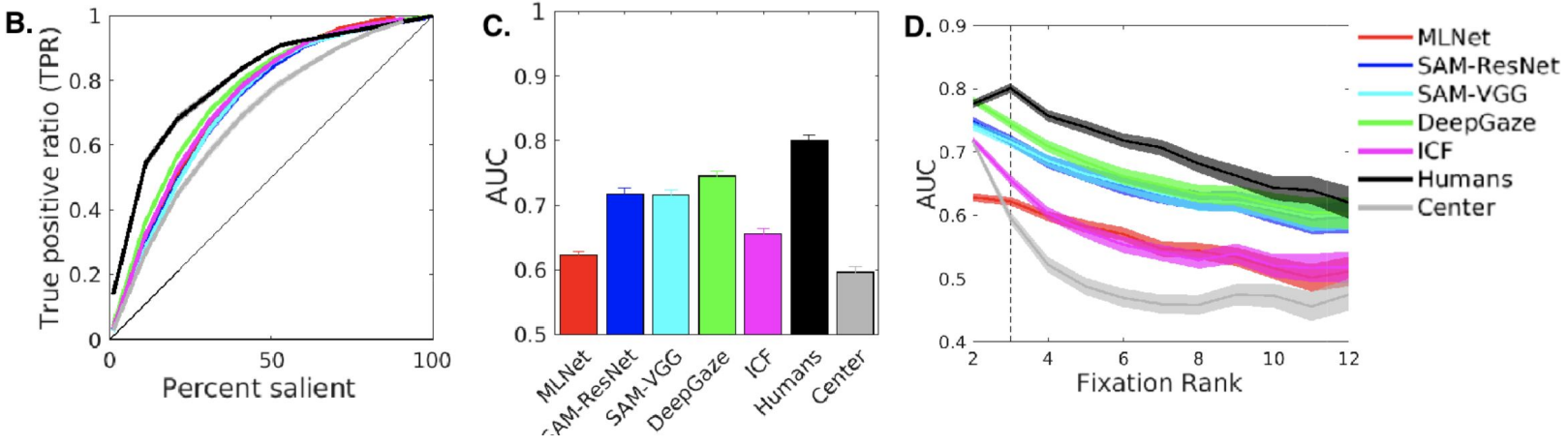
(a) Imagen original



(b) 8.14 % más saliente de la imagen según SAM

(c) 27.73 % más saliente de la imagen según SAM

True Positive Rate (TPR) como métrica



Todos los colores son diferentes modelos de saliencia. Predicen muy bien las primeras fijaciones (ver B, C que muestran la fijación número 3). Sin embargo, a medida que miramos a más lugares, la saliencia es cada vez menos útil.

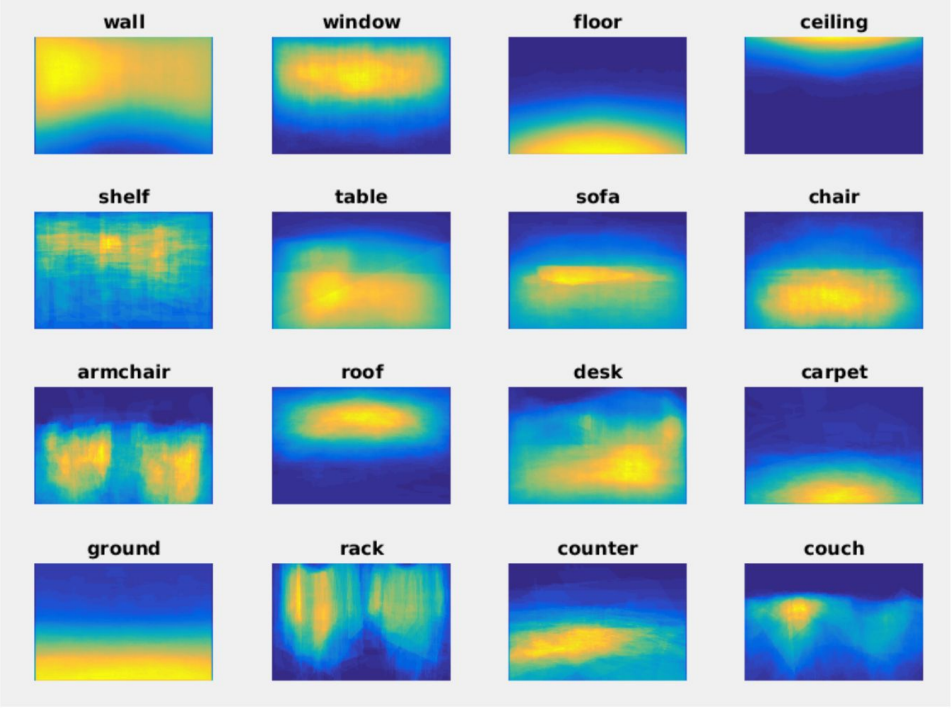
Entrenar con features específicos de búsqueda visual mejora las predicciones

Queremos crear un mapa teniendo en cuenta la semántica de la imagen mostrada.

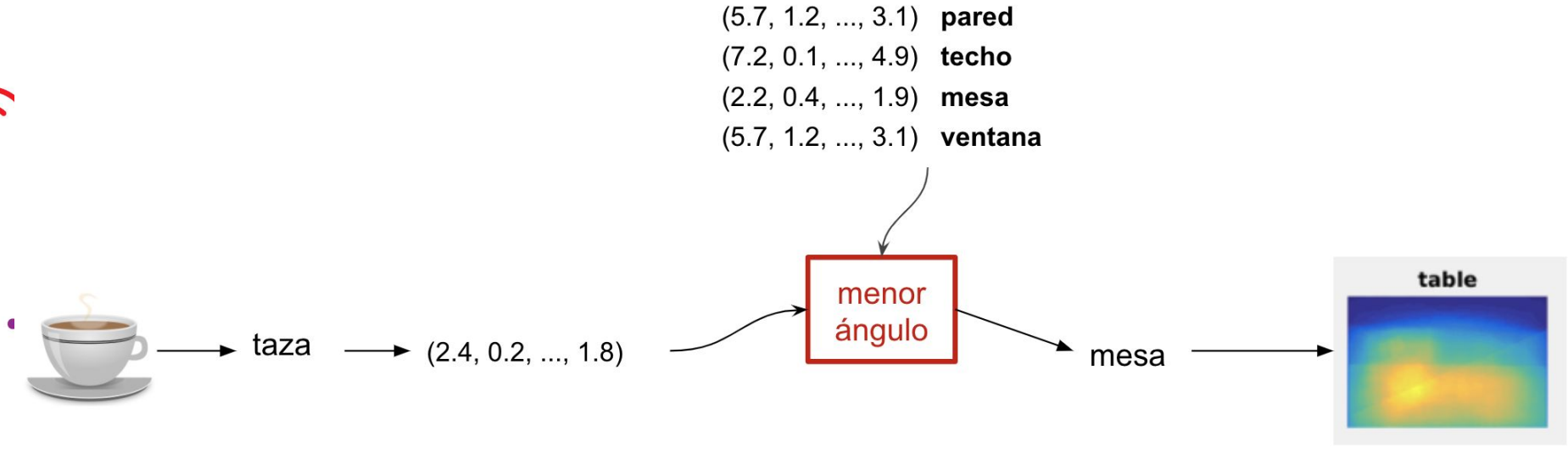
Recordemos que sabemos a qué clase de objeto pertenece cada target por nuestro experimento online.

Además, encontramos un dataset llamado LabelMe que posee 8000 imágenes segmentadas y parcialmente anotadas por la comunidad.

Entrenar con features específicos de búsqueda visual mejora las predicciones



Entrenar con features específicos de búsqueda visual mejora las predicciones



Modelar el proceso de búsqueda como un modelo bayesiano

Qué propiedades tiene que tener nuestro modelo?

- Las redes neuronales profundas son muy poco interpretables con las herramientas de hoy en día.
- Si nuestro objetivo es entender el comportamiento humano, usar redes neuronales no nos ayuda
- **Los modelos bayesianos son altamente interpretables, y ya se utilizaron para modelar el comportamiento humano en varias tareas. Modelan el update de información en cada paso.**

Modelo IBS (basado en Najemnik & Geisler 2005)

Quiero calcular la probabilidad de que el objeto esté en el lugar **L**, dado que ya miré a **N** lugares anteriormente (y recuerdo cuáles son!)

Modelo IBS (basado en Najemnik & Geisler 2005)

Quiero calcular la probabilidad de que el objeto esté en el lugar **L**, dado que ya miré a N lugares anteriormente (y recuerdo cuáles son!)

$$visibilidad(L,\ lugarvisto_t) \times similitud(L,\ lugarvisto_t)$$

Puede ser negativo!

Modelo IBS (basado en Najemnik & Geisler 2005)

Quiero calcular la probabilidad de que el objeto esté en el lugar **L**, dado que ya miré a N lugares anteriormente (y recuerdo cuáles son!)

$$\prod_{t=1}^{N} visibilidad(L,\ lugarvisto_t) \times similitud(L,\ lugarvisto_t)$$

Puede ser negativo!

Modelo IBS (basado en Najemnik & Geisler 2005)

Quiero calcular la probabilidad de que el objeto esté en el lugar **L**, dado que ya miré a N lugares anteriormente (y recuerdo cuáles son!)

$$\prod_{t=1}^{N} e^{visibilidad(L, lugarvisto_t) \times similitud(L, lugarvisto_t)}$$

Si era un número positivo grande, ahora es mucho más grande. Si era un negativo grande, ahora es casi casi cero.
Es el mismo truco que en softmax!

Modelo IBS (basado en Najemnik & Geisler 2005)

Quiero calcular la probabilidad de que el objeto esté en el lugar **L**, dado que ya miré a N lugares anteriormente (y recuerdo cuáles son!)

$$saliencia(L) \times \prod_{t=1}^{N} e^{visibilidad(L, lugarvisto_t) \times similitud(L, lugarvisto_t)}$$

Si era un número positivo grande, ahora es mucho más grande. Si era un negativo grande, ahora es casi casi cero.
Es el mismo truco que en softmax!

Modelo IBS (basado en Najemnik & Geisler 2005)

Quiero calcular la probabilidad de que el objeto esté en el lugar **L**, dado que ya miré a N lugares anteriormente (y recuerdo cuáles son!)

$$saliencia(L) \times \prod_{t=1}^{N} e^{visibilidad(L, lugarvisto_t) \times similitud(L, lugarvisto_t)}$$

Luego, lugarvisto_{N+1} será la posición de la imagen que maximice la probabilidad de identificar bien la ubicación del objeto después de que la fijación N + 1 sea realizada.

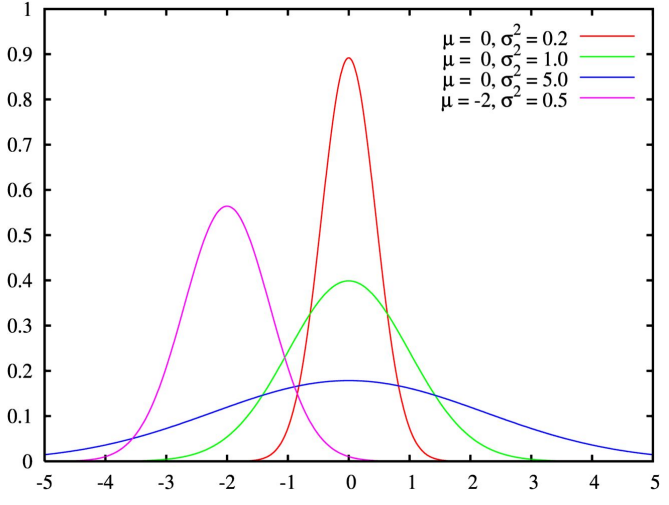
Definición de similitud al target (objetivo de búsqueda)

- **similitud(L, lugarvisto_t)** representa cuán parecida es la posición **L** al target para el observador que se encuentra fijando su vista en **lugarvisto_t**.

$$similitud\_target_{ik(f)} \in \mathcal{N}(\mu_{ik(f)}, \sigma^2_{ik(f)})$$

$$\mu_{ik(f)} = \begin{cases} 0{,}5 & si\ i = \ ubicación\ del\ target \\ -0{,}5 & caso\ contrario \end{cases}$$

$$\sigma^2_{ik(f)} = \frac{1}{visibilidad^2_{ik(f)}}$$



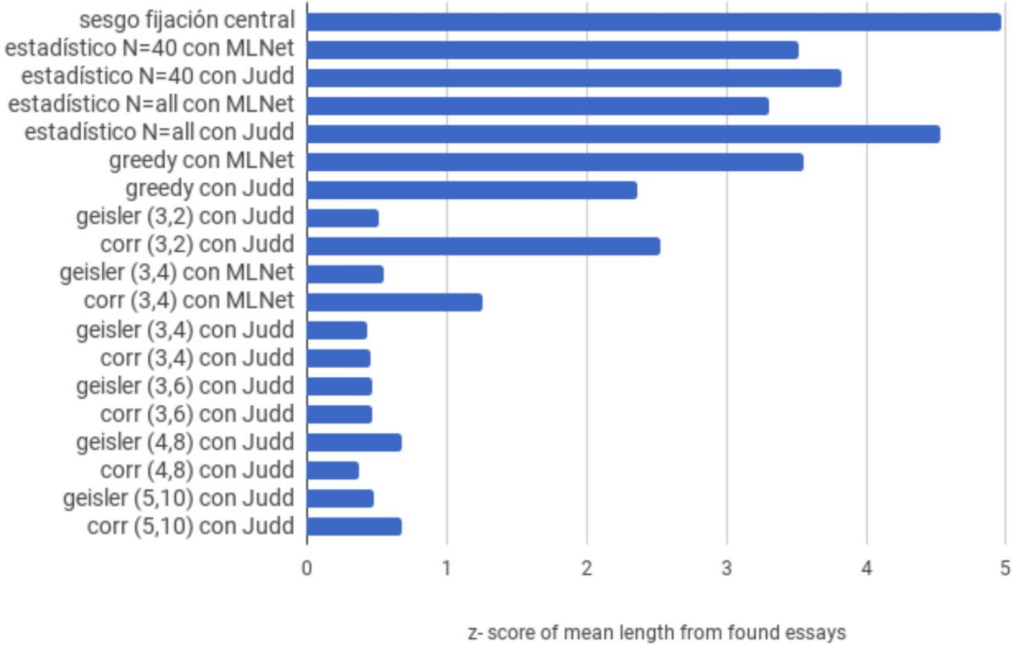
Métricas de comparación con humanos

Cómo medimos performance?

- Métrica de número de fijaciones esperadas para encontrar el target, solamente para ensayos exitosos.
- Métrica de performance sobre todos los ensayos: calcula la proporción de ensayos exitosos promedio entre todos los ensayos con c sacadas permitidas c = {2, 4, 8, 12}.
- Métrica de string edit distance.

Resultados respecto de la primera métrica

Métrica de número de fijaciones esperadas para encontrar el target, solo para ensayos exitosos (z-score de diferencia contra humanos). **Más chico es mejor, significa que se diferencia menos de los humanos.**



Aplicaciones de búsqueda visual a otras disciplinas

- **Interfaces Web**: Entender cómo los usuarios buscan en UIs y mejorarlas
- **Marketing:** si sabemos el recorrido que hace la visión al buscar algo (o al mirar libremente) se puede decidir qué objetos o elementos poner en el camino que recorre la mirada
- Hoy en día se hacen experimentos de este tipo, pero sí o sí se necesita un Eye Tracker. Las cámaras de los celulares o laptops no pueden distinguir el área de la mirada

Imagen: https://www.usability.gov/how-to-and-tools/methods/eye-tracking.html

¡GRACIAS!

Melanie Sclar

@melaniesclar

@melaniesclar