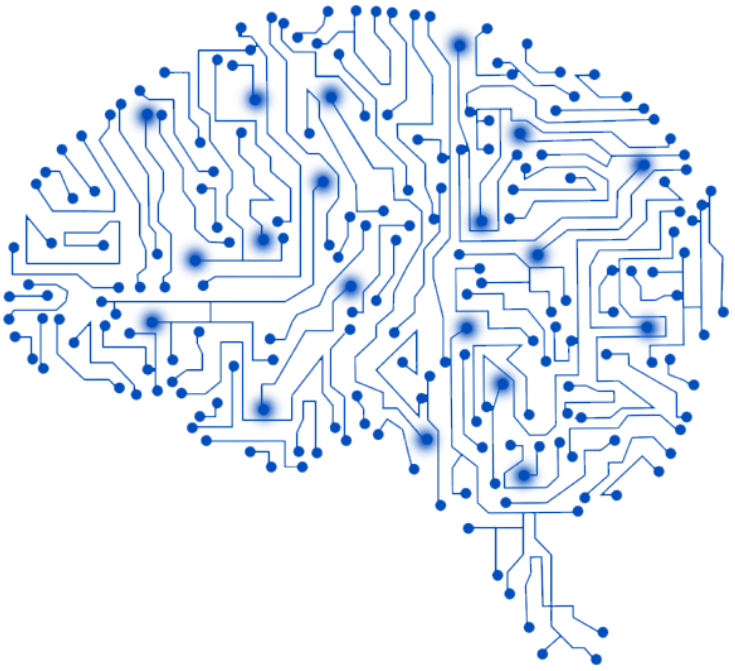
Recurrent Cortical Networks

Modelos de Computación

Valentín Pedrosa Campoy



Índice

Introducción

En la actualidad nos encontramos en un momento muy interesante para las ciencias de la computación. En el campo de la inteligencia artificial se están dando numerosos avances en los últimos años los cuales están suponiendo una revolución sin precedentes en el mundo de la informática.

En el campo del aprendizaje automático y sistemas inteligentes hemos visto como, en escasos años, hemos partido desde unos sistemas expertos a unos sistemas que son capaces de aprender por sí mismos. Estos sistemas han evolucionado desde sistemas iniciales basados en funciones de refuerzo, hasta sistemas más complejos dónde se utilizan humanos para generar un conjunto limitado de ejemplos de los que el propio programa va aprendiendo.

El verdadero problema

Nos hemos encontrado con un problema a la hora de enseñar a los sistemas inteligentes basados en aprendizaje automático. Necesitamos un conjunto de ejemplos excesivamente grande. Por ejemplo, para enseñar a un programa a reconocer una cara en una imagen, necesitamos millones de imágenes con caras previamente clasificadas. A día de hoy se han conseguido grandes avances e implementaciones prácticas muy útiles para la sociedad gracias a estos sistemas.
Sin embargo, debido al volumen de datos que es necesario tener de partida, se hace complicado y costoso el desarrollo de nuevas implementaciones. Además de eso, en algunos casos el porcentaje de acierto en el aprendizaje para sistemas que ya han aprendido, suele ser muy bajo en el momento en el que han de resolver un problema ligeramente diferente de los encontrados en el conjunto de datos de prueba.

Si fuéramos capaces de mejorar la velocidad de aprendizaje y de reducir el número de ejemplos necesarios para que un sistema inteligente fuera capaz de resolver problemas con un porcentaje de éxito lo suficientemente elevado, podríamos desarrollar aplicaciones y sistemas que aprendieran en tiempo real de los mismos problemas que están resolviendo. Si consiguiéramos que, además, esos sistemas fueran capaces de generalizar los resultados obtenidos y resolver problemas para los que no estaban preparados, estaríamos ante sistemas capaces de mostrar inteligencia artificial completa.

Las soluciones actuales

El uso de redes neuronales y redes convolucionales para la gestión de problemas de aprendizaje automático obliga a los desarrolladores e investigadores a utilizar un conjunto de datos muy grande y unas funciones de refuerzo muy depuradas.

Hasta hace muy poco tiempo no existía una solución eficiente y tecnológicamente posible que permitiera el reconocimiento de textos en imágenes con una precisión fiable en un tiempo considerablemente pequeño y con un conjunto de datos del orden de los miles de imágenes.

Vicarious AI

Sin embargo, el 8 de diciembre de 2017, unos investigadores del grupo de investigación Vicarious AI, en Union City, California, publicaron un artículo en la revista Science llamado “A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs”. En este paper se describe un modelo basado en redes corticales recursivas capaces de aprender con un conjunto de ejemplos muy limitado, en un tiempo mucho menor que las redes convolucionales usuales, y con un porcentaje de acierto mucho mayor que sus competidoras.

Figura 1. Logotipo de vicarious AI.

Además presentan un salto en cuanto a la generalización con respecto a las redes neuronales y convolucionales actuales. Son capaces de generar un modelo jerárquico que reconoce la forma y la separa de la apariencia, con lo que consiguen resolver problemas y discernir ejemplos que se diferencian bastante del conjunto de datos de prueba que se le dio al sistema en un inicio.

El problema

Los ordenadores y el general las máquinas, siempre han servido al ser humano para la automatización de procesos que anteriormente eran desempeñados por cadenas

de manufacturación. Asimismo en el transcurso de nuestras vidas encontramos tareas relacionadas con la visión que requieren de la supervisión de un humano y que de forma normal no son fácilmente automatizables.

Este tipo de tareas como por ejemplo la vigilancia de un centro comercial o el diagnóstico de un paciente en función de una radiografía requieren, por una parte, de un conocimiento de la situación y el contexto, y por otra parte del análisis y comprensión de imágenes. Además de ser un problema en términos de tiempo y coste el hecho de que una persona tenga que realizar estas tareas, en algunos casos las tareas se requieren de una velocidad o complejidad que las vuelven imposibles e inalcanzables.

Este tipo de trabajos tienen un impacto directo en la calidad de vida de la humanidad. Por ejemplo si fuéramos capaces de transportar determinadas mercancías sin necesidad de la intervención humana, el precio de los productos que consumimos día a día se vería reducido drásticamente. De la misma manera, si fuéramos capaces de introducir el diagnóstico proporcionado por un médico especialista, en un smartphone o similar, podríamos, por ejemplo, aumentar la frecuencia de las consultas médicas de previsión que se realizan a los pacientes que aún no saben que tienen cáncer, aumentando en un gran número las detecciones de tumores malignos en estadíos curables u operables.

En este trabajo me centro sobre el reconocimiento de textos en imágenes, concretamente aquellos textos que aparecen deformados, incompletos, o de forma incoherente en una imagen. Estas aplicaciones y muchas otras son las que intenta hacer posible una rama de la informática, en el campo de la inteligencia artificial, llamada visión por computador o visión artificial.

Visión artificial

La visión por computador es un campo interdisciplinar cuya misión es obtener una solución que permita construir sistemas que tengan un nivel de comprensión muy alto de vídeos e imágenes digitales, sin la supervisión de un ser humano.

Desde la perspectiva de la ingeniería, busca automatizar tareas que los sistemas visuales humanos pueden hacer, como por ejemplo la vigilancia o la lectura.

Las tareas de la visión por computador incluyen métodos para adquirir, procesar, analizar y comprender imágenes digitales, y la extracción de datos multidimensionales del mundo real, con el objetivo de producir información

numérica o simbólica. Ubicada en este contexto, significa la transformación de imágenes visuales, por ejemplo lo que ve la retina en un ojo humano, en descripciones del mundo que puedan ser consultadas e intervenidas a través de otros procesos. Esta comprensión de imágenes puede ser vista como la descomposición de la información simbólica de los datos de las imágenes usando modelos construidos con la ayuda de la geometría, la física, las estadísticas, y la teoría del aprendizaje.

Historia de la Visión por Computador

La visión por computador comienza en las universidades que eran pioneras en inteligencia artificial al final de los años 60. Surge como la necesidad de reproducir el sistema visual humano, como un paso clave y necesario para llegar al desarrollo de robots con inteligencia y comportamientos inteligentes. En 1966, se pensaba que podía ser conseguido de forma sencilla añadiendo una cámara a un robot y simplemente “que la cámara describiera lo que iba viendo”.

Lo que distingue la visión por computador del prevalente campo del procesamiento de imágenes digitales en este momento es el deseo de extraer estructuras tridimensionales de imágenes con el objetivo de conseguir la comprensión completa de una escena. Fueron los estudios de los años 70 los que fundaron y construyeron la mayoría de los algoritmos de visión por computador que existen hoy día, incluyendo la extracción de ejes de imágenes, el marcado de líneas, el modelado poliédrico y no poliédrico, flujo óptico y estimación del movimiento, entre otros.

La siguiente década vió estudios basados en análisis matemáticos más rigurosos y aspectos cuantitativos de la visión por computador. Estos incluyeron el concepto de la escala espacial, la inferencia de las formas a partir de varias señales como el sombreado, la textura y el foco, y los modelos de contorno conocidos como “snakes”.

Los investigadores también se dieron cuenta de que muchos de estos conceptos matemáticos podían ser tratados con los mismos marcos de optimización que la regularización y los campos aleatorios de Markov. Para los 90, la investigación cambió su curso normal y se dirigió hacia temas que hasta el momento parecían muertos. La investigación en reconstrucciones proyectivas en tres dimensiones llevo a una comprensión mucho más avanzada de la calibración de la cámara. Con la llegada de métodos de optimización para la calibración de la cámara, la comunidad científica se dio cuenta de que un montón de las ideas se habían explorado anteriormente en la teoría de ajustes de haces del campo de la fotogrametría. Esto llevó a métodos de reconstrucción de escenas tridimensionales desde múltiples

imágenes. Al mismo tiempo, se utilizaron variaciones de las técnicas de corte de gráficos para resolver los problemas de segmentación en imágenes. Esta década también marcó el inicio de las primeras técnicas de aprendizaje estadístico usadas para reconocimiento práctico de caras en imágenes. Al final de los años 90, se dio un cambio muy significativo con el incremento de la interacción entre los campos de computación gráfica y visión por computador.

En la actualidad las investigaciones que se están llevando a cabo están viendo el resurgir de los métodos basados en características, usados en conjunción con técnicas de machine learning y marcos de trabajo de optimización complejos.

Aplicaciones en la actualidad

Las aplicaciones de la visión por computador son muy variadas. Van desde sistemas de visión industrial que, por ejemplo, son capaces de clasificar diferentes objetos en función de forma y color en una cadena de producción, hasta investigaciones en los campos de inteligencia artificial y computación o robótica con el objetivo de crear sistemas que comprendan el mundo que les rodea. Los campos de visión de máquinas o machine vision y visión por computador se entrelazan entre ellos. La visión por computador cubre la tecnología principal del análisis automático de imágenes, que es usado en muchos campos, mientras que la visión de máquinas usualmente se refiere a procesos que combinan el análisis automático de imágenes con otros métodos y tecnologías para proveer de inspección y guía automática en aplicaciones industriales.
En muchos sistemas y aplicaciones de visión por computador, los ordenadores están pre-programados para resolver una tarea en particular, aunque los métodos basados en el aprendizaje automático son cada vez más frecuentes.
Los ejemplos de aplicaciones de visión por computador incluyen sistemas para:

- Inspección automática, por ejemplo, en aplicaciones de manufacturación.
- Asistencia a humanos en tareas de identificación, por ejemplo, sistemas de identificación de especies para biólogos.
- Control de procesos, por ejemplo, un robot industrial.
- Detección de eventos, por ejemplo, supervisión y vigilancia industrial conteo de personas.
- Interacción, por ejemplo, como sistema de entrada a un dispositivo para interacción entre humanos y computadores.
- Modelado de objetos o entornos, por ejemplo, análisis de imágenes médicas o modelado topográfico.
- Navegación, por ejemplo, para vehículos autónomos o robots móviles.

- Organización de la información, por ejemplo, para indexar bases de datos de imágenes y de secuencias de imágenes.

Uno de los campos de aplicación más prominentes es la visión por computador en la medicina o procesamiento de imágenes médicas. Este área está caracterizada por la extracción de información de las imágenes con el propósito de realizar un diagnóstico médico a un paciente.

Un segundo área de aplicación en visión artificial está en la industria, muchas veces llamada visión de máquinas o machine vision, de la que hemos hablado anteriormente, donde la información es extraída con el propósito de ayudar a un proceso de manufacturación. Un ejemplo es el control de calidad donde los productos son inspeccionados de forma automática para encontrar defectos.

Sobre la exactitud y los porcentajes

Al estudiar el campo de la visión por computador o visión artificial hoy en día nos encontramos con soluciones y sistemas capaces de resolver problemas siempre con un porcentaje de éxito. Este porcentaje de éxito depende de muchos factores, pero es muy importante que lo situamos en contexto para entender lo que realmente significa.

Por ejemplo, en un estudio de 2010 de la Universidad de Stanford, donde se presentaban imágenes con textos deformados o incompletos a diferentes grupos de personas, se demuestra que dependiendo de la prueba que se le realizaba a los humanos, el porcentaje de acierto de estos era significativamente diferente.

En este estudio se muestra que independientemente del objetivo del test, cuando se presentaban las imágenes a 3 personas diferentes, solamente estaban de acuerdo en el 71% de los casos de media. Este contraste es muy significativo en relación, por ejemplo, a un artículo en 2013 donde Google anunciaba que había desarrollado un algoritmo que era capaz de leer sus propios sistemas de elusión de bots en un 99,8% de los casos.

Esto quiere decir que aunque los algoritmos de visión por computador puedan parecer inexactos, hemos de recordar siempre que estamos intentando imitar la visión humana, que en muchos casos puede ser inexacta.

La solución actual

Actualmente el ser humano ha decidido atajar el problema diseñando sistemas parecidos al cerebro humano o animal.

Tras una época en la que los sistemas expertos demostraron sus limitaciones, la inteligencia artificial y las matemáticas han tomado el camino de modelos probabilísticos de análisis inteligente.

Red Neuronal

Las redes neuronales artificiales son sistemas de computación inspirados por las redes neuronales biológicas, que constituyen los cerebros de los animales. Estos sistemas aprenden tareas, observando ejemplos, generalmente sin la necesidad de programación específica para la resolución de ellas.

Una red neuronal artificial está basada en una colección de unidades conectadas o nodos, llamadas neuronas artificiales como analogía a las neuronas biológicas en los cerebros animales. Cada conexión entre las neuronas (o sinapsis) puede transmitir una señal de una a otra. La neurona receptora puede procesar la señal y reenviarla a las neuronas conectadas a ella.

En las implementaciones comunes de las redes neuronales, la señal de sinapsis es un número real y la salida de cada neurona se calcula mediante una función no lineal de la suma de sus entradas. Las neuronas y las sinapsis tienen de forma usual un peso que ajusta los procesos de aprendizaje. este peso se ve incrementado o decrementado con la fuerza de la señal que envía a través de la sinapsis. Las neuronas tienen una función de límite que es capaz de bloquear aquellas sinapsis débiles que intentan cruzar el sistema sin la suficiente fuerza.

Típicamente, las neuronas se organizan en capas. Dependiendo del número y tipología de las capas, obtendremos unos resultados diferentes en función de los mismos valores de entrada. Las señales viajan desde la primera a la última capa, con la posibilidad de atravesar una capa más de una vez.

El objetivo original de las redes neuronales artificiales era resolver problemas de la misma manera que un cerebro humano los resolvería. A lo largo del tiempo, la atención se centró en imitar determinadas habilidades mentales, desviando la investigación y el desarrollo de las redes neuronales de la biología.

Red Neuronal Convolucional

Una red neuronal convolucional, en Machine Learning, es una clase de red neuronal artificial profunda de alimentación en una dirección, que se ha aplicado de forma exitosa al análisis de imágenes visuales. Las redes neuronales convolucionales utilizan alguna variante de las neuronas de las redes neuronales artificiales, con el objetivo de reducir al mínimo el preprocesamiento. También se conocen como redes neuronales artificiales invariantes en el espacio, basándose en la arquitectura de capas compartidas y sus características de invarianza de traslación.

Las redes neuronales convolucionales se inspiran en procesos biológicos en los cuales existen patrones de conectividad entre neuronas vistos en el cortex visual animal. Las neuronas corticales responden a estímulos solamente en una zona restringida del campo visual conocida como el campo receptivo. Los campos receptivos de las diferentes neuronas se superponen de tal forma que cubren el campo visual completo.

Las redes neuronales convolucionales utilizan muy poco preprocesamiento comparado con otros algoritmos de clasificación de imágenes. Esto significa que la red aprende los filtros que los en algoritmos tradicionales han de diseñar e implementar de forma manual los investigadores y desarrolladores. Está independencia de conocimiento previo y de esfuerzo humano es una de las principales ventajas de este tipo de red neuronal.

Sus principales aplicaciones se encuentran en el reconocimiento de imágenes y vídeo, sistemas de recomendación y procesamiento del lenguaje natural.

El perceptron multicapa

Un perceptron multicapa es un tipo de red neuronal artificial de alimentación en un único sentido. Un perceptron multicapa consiste en al menos tres capas de nodos. Excepto por los nodos de entrada, cada nodo es una neurona que usa una función de activación no lineal. Un perceptron multicapa utiliza una técnica de aprendizaje supervisado llamada aprendizaje por propagación hacia atrás. Sus múltiples capas y su función de activación no lineal distinguen el perceptron multicapa del perceptron lineal. Puede distinguir datos que no son linealmente separables.
En geometría, que un dato sea linealmente separable es una propiedad de un par de conjuntos puntos en el espacio. Dos conjuntos puntos en el espacio son linealmente separables si existe al menos una línea en el plano que deja a todos los puntos de un conjunto a un lado y a los del otro en el otro.

Características que distinguen a las Redes Convolucionales

Para comprender mejor una de las principales desventajas de los perceptrones multicapa hemos de entender la “maldición de la dimensionalidad”.

La maldición de la dimensionalidad se refiere a varios fenómenos que se dan cuando se analiza y organiza data en espacios de multi-dimensionales (usualmente con cientos o miles de dimensiones), que no ocurren en espacios con pocas dimensiones como el espacio físico de tres dimensiones de nuestra experiencia cotidiana. Esta expresión fue acuñada por Richard Bellman cuando trabajó con problemas de optimización dinámica.

Hay múltiples fenómenos referidos por este nombre en dominio es como el análisis numérico, muestreo, combinatorias, machine learning, data mining, y bases de datos. El tema común de estos problemas es que cuando la dimensionalidad se ve incrementada, el volumen del espacio se incrementa tan rápido que los datos disponibles son escasos. Esta escasez es muy problemática para cualquier método que requiera significancia estadística. De acuerdo a la obtención de un resultado estadístico fiable, la cantidad de datos necesaria para dar lugar a un resultado normalmente crece de forma exponencial con la dimensionalidad. De la misma manera, organizar y buscar en datos normalmente depende de la detección de áreas donde los objetos forman grupos con propiedades similares. Sin embargo en datos multidimensionales todos los objetos parecen ser escasos y poco similares entre sí, lo que hace que las estrategias de organización de datos comunes sean ineficientes.

Los modelos tradicionales basados en perceptron multicapa se utilizan de forma exitosa para reconocimiento de imágenes, pero dada la conectividad completa entre los nodos sufren de la maldición de la dimensionalidad anteriormente mencionada, y por ello no son capaces de escalar bien imágenes de alta resolución.

Para ponernos en contexto, utilizando las imágenes de CIFAR-10, un conjunto de datos de prueba y entrenamiento estándar que utiliza imágenes de 32 x 32 píxeles, nos encontramos con neuronas completamente conectadas que presentan hasta 3072 pesos. Las mismas imágenes, en un tamaño de 200 x 200 píxeles, darían lugar a neuronas con hasta 120.000 pesos, lo que hace el sistema de inviable para alta resolución.

De la misma manera ese tipo de arquitecturas de red neuronal no tienen cuenta la estructura espacial de los datos, tratando a los píxeles de entrada que están muy lejos entre sí, de la misma manera que los píxeles que están muy cerca. Por ello la

conectividad completa de las neuronas supone un mal uso de los recursos para propósitos como el reconocimiento de imágenes que está dominado por patrones espacialmente diferenciados.

Las redes neuronales convolucionales son variantes de los perceptrones multicapa inspiradas biológicamente, diseñadas para simular el comportamiento del cortex visual. Utilizan volúmenes neuronales tridimensionales, conectividad local, y pesos compartidos, características que permiten mitigar o eliminar los problemas que presentaban los perceptrones multicapa de cara la escalabilidad de los sistemas.

Red Neuronal Convolucional Recurrente

Las redes neuronales recursivas usuales presentan el problema de que la captura de la información sobre las frases se realizaba en árboles, pero su construcción era ineficiente. Las redes convolucionales pueden aprender las frases más importantes pero tienen problemas para el procesamiento de textos, puesto que encontrar el núcleo de la frase óptimo es muy complicado.

En el paper “Recurrent Convolutional Neural Networks for Text Classification” de 2015, unos investigadores del Laboratorio Nacional de reconocimiento de patrones de la Academia China de Ciencias, describe las limitaciones que han encontrado a la hora de utilizar redes neuronales recursivas y redes neuronales convolucionales para el procesamiento del lenguaje natural y describe las redes convolucionales recurrentes como una solución a la limitación de los modelos anteriores, que se limitaban a la información convencional basada en el marco de las redes neuronales, dando lugar a un modelo que es capaz de representar la semántica de los textos de forma más precisa para su clasificación.

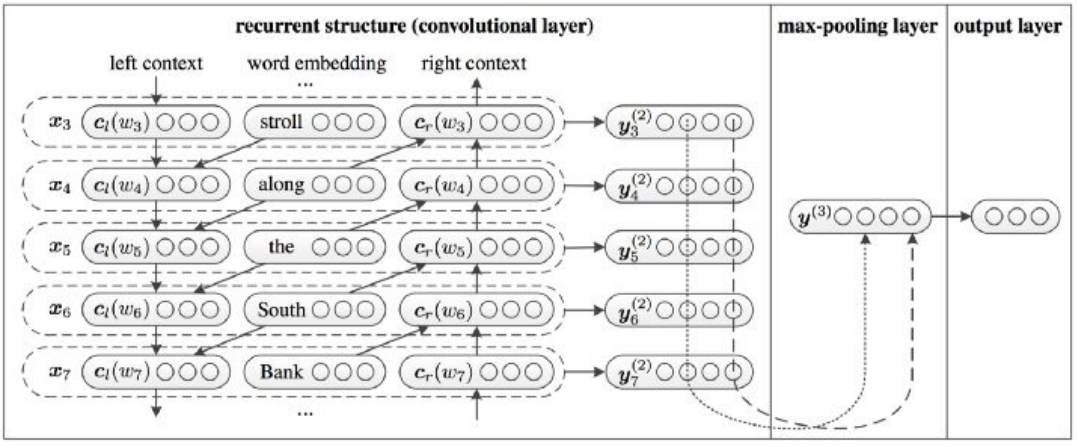


Figura 2. La estructura de una red neuronal convolucional recurrente para procesamiento de lenguaje natural. Esta imagen es un ejemplo parcial de la frase “A sunset stroll along the South Bank affords an array of stunning vantage points”, y la estructura que tendría en la red convolucional recurrente.

El modelo que propusieron fue una estructura bidireccional recurrente, que introduciría menos ruido comparada con la estructura basada en ventana de la red neuronal tradicional, capturando la información contextual de la extensión más grande posible mientras aprendía la representación de las palabras. De hecho, el modelo podía recordar un rango bastante grande de la ordenación de las palabras cuando estaba aprendiendo las representaciones de los textos.

Además, emplearon una capa de agrupación máxima o “max-pooling”, que juzgaba de forma automática qué características jugaban roles importantes en la clasificación de los textos, para capturar el componente clave de los mismos. Gracias a la combinación de la estructura recurrente y la capa de agrupación máxima, su modelo utilizaba las ventajas tanto de las redes neuronales recurrentes como de las redes convolucionales. De hecho su modelo exhibía una complejidad en el tiempo de O(n), que se relacionaba linealmente con la longitud del texto.

Durante las pruebas que realizaron, compararon su modelo con otros modelos diferentes de vanguardia utilizando cuatro tipos de tareas diferentes en inglés y chino. La taxonomía de la clasificación contenía clasificaciones por tema, sentimientos y estilo de escritura. Los experimentos demuestran que su modelo era mejor que el resto en tres de los cuatro bancos de datos comúnmente usados.

Red Cortical Recursiva

Y con tan increíble precedente llegamos a las redes corticales recursivas. Una red cortical recursiva es un término biológico que se refiere a cómo se comportan determinadas neuronas de la parte el cortex visual en el cerebro humano. En una publicación en Science del 8 de diciembre, un equipo de investigadores de California describe lo que ellos llaman una red cortical recursiva y la definen como un modelo generativo que satisface una serie de requisitos funcionales que actualmente los modelos computacionales existentes eran incapaces de satisfacer. Además consigue un rendimiento óptimo y una eficiencia de datos muy alta en un conjunto diverso de tareas de visión por computador o visión artificial.

La red cortical recursiva representa una salida de patrón de trabajo de Deep Learning y aprendizaje automático que premia el aprendizaje desde el inicio, sin ninguna base preestablecida. La red cortical recursiva que describen los investigadores de Vicarius AI comienza y se basa sobre un andamiaje o estructura que permite y facilita la construcción de modelos.

Por ejemplo, mientras la mayoría de redes convolucionales contemplan modelos de imagen completos que asumen muy poco sobre los objetos y las imágenes que analizan, la red cortical recursiva es un modelo basado en objetos, que asume la factorización de contornos y superficies, y objetos y fondo. La red cortical recursiva también representa explícitamente las formas, y la presencia de conexiones laterales que permiten agrupar grandes transformaciones sin perder especificidad, con lo que incrementa su invarianza. La composicionalidad permite a las redes corticales recursivas representar escenas con múltiples objetos habiendo estado expuestas solamente a entrenamiento específico para objetos individuales. Todas estas características de las redes recursivas corticales derivan de la suposición de que la evolución ha dotado al neocortex animal con estas estructuras que le permiten aprender de forma sencilla las representaciones de nuestro mundo, comparándolo con la dificultad de empezar todo desde un papel completamente en blanco.

El sentido común y sus problemas

Las investigaciones recientes en machine learning e inteligencia artificial son usualmente reduccionistas. Los investigadores identifican un aspecto de la inteligencia, lo aislan y definen sus características, y crean una serie de pruebas para evaluar los progresos en ese problema en concreto, mientras controlan las variables externas todo lo que pueden. El problema del sentido común es que es resistente a este tipo de reducciones, debido a que envuelve muchos aspectos diferentes de la inteligencia desde el mismo modelo. En el caso de la visión, después de que un modelo de sentido común se ha construido, tiene que ser capaz de permitir el reconocimiento de objetos, segmentación, imputación, generación, y un número factorial de consultas que ligan las variables representadas de diferentes formas, sin requerir reentrenamiento para cada uno de los tipos de estas consultas.

La investigación en modelos generativos usualmente se centra en soluciones concretas que pueden resolver consultas específicas, pero no ofrecen una forma simple para elevar el conocimiento del modelo completamente a través de consultas probabilísticas arbitrarias. Por ejemplo, en los autoencoders variacionales (VAEs), un derivado del entrenamiento es una red de inferencia rápida. Sin embargo, si al modelo se le pregunta utilizando una variable diferente de las observadas en el conjunto cada vez, necesitamos mantener una red diferente por cada consulta, convirtiendo el modelo en algo inútil. Encima de esto, el abrumador énfasis sobre la optimización de la “lower bound evidence” (ELBO) en modelos de caja negra, le quita el énfasis a la importancia de obtener variables con mucho sentido que se encuentren latentes. El uso de una estructura generativa adecuada (parcialmente

inductiva) es beneficiosa tanto desde el punto de vista de la interpretabilidad como para permitir integraciones ricas en sistemas más complejos, incluso si el precio a pagar es una “lower bound evidence” un poco más baja. Una de las fortalezas, pero al mismo tiempo limitaciones, de las redes generativas adversarias, es que no prescriben ningún mecanismo de inferencia así que incluso después de haber entrenado satisfactoriamente un modelo generativo, nos tenemos que ir a una técnica diferente para resolver consultas probabilísticas. Incluso los modelos más manejables, se definen siguiendo un orden que hace manejable solamente algunas consultas haciendo otras completamente inmanejables.

Estos modelos generativos individuales son potentes en los confines de su régimen de entrenamiento, pero no son capaces de alzarse con conocimiento coherente del mundo que nosotros somos capaces de reconocer como sentido común. En la búsqueda de los principios más allá de estos sucesos concretos, nos encontramos con la única implementación exitosa del sentido común: el cerebro humano.

El papel de la neurociencia

La neurociencia es un campo del conocimiento científico que estudia la estructura, función, el desarrollo de la bioquímica, la farmacología y la patología del sistema nervioso y de como sus diferentes elementos interactúan, dando lugar a las bases biológicas de la conducta.

Debido a que todas las redes neuronales descritas en este trabajo son una imitación o un intento de imitación del cerebro humano, la neurociencia está completamente ligada a todos los campos de la inteligencia artificial que involucran redes neuronales.

El estudio biológico del cerebro es un área multidisciplinar que abarca muchos niveles de estudio, desde el puramente molecular hasta el específicamente conductual y cognitivo. Es en este nivel donde los investigadores de Inteligencia Artificial pueden obtener la ayuda de la neurociencia para el diseño de redes neuronales cada día más eficientes y más parecidas a un cerebro humano.

En el nivel más alto las neurociencias se combinan con la psicología para crear la neurociencia cognitiva, una disciplina que al principio fue dominada totalmente por psicólogos cognitivos. Hoy en día la neurociencia cognitiva proporciona una nueva manera de entender el cerebro y la conciencia, pues se basa en un estudio científico que une disciplinas tales como la neurobiología, la psicobiología, o la propia psicología cognitiva, un hecho que con seguridad cambiar a la concepción actual

que existe acerca de los procesos mentales implicados en el comportamiento y sus bases biológicas.

A la par que la neurociencia y la psicología avanza sobre el estudio del cerebro humano, las disciplinas de la inteligencia artificial encargadas del aprendizaje automático de las máquinas progresan.

CAPTCHA

Los captcha son comúnmente conocidos como las pruebas requeridas por determinados sistemas para evitar el envío automatizado de información por parte de robots en formularios de páginas webs.

Sin embargo los captcha son las siglas de “Prueba de Turing completamente automática y pública para diferenciar a ordenadores de humanos”, en inglés. Lo que estamos acostumbrados a ver en los diferentes sitios webs es una de las múltiples pruebas captcha que existen. Este test es controlado por una máquina, en lugar de por un humano como en la prueba de Turing, por lo que se le llama una prueba de Turing inversa.

A día de hoy existen algunas aproximaciones de cómo se puede romper el sistema captcha usando humanos como mano de obra barata en voluntaria para reconocerlos, explotando pus en la implementación y finalmente mejorando el software de reconocimiento óptico de caracteres. Sin embargo las técnicas de machine learning y de visión por computador cómo las redes corticales recursivas suponen un antes y un después en lo que a los captcha de texto se refiere.

En concreto nos centraremos en el captcha creado por Google, llamado reCAPTCHA. Este captcha concretamente protege a los usuarios de Internet del spam y el abuso de los servicios allá donde vayan. Desde la propia web de Google nos lo anuncian como un servicio de seguridad avanzada, facilidad de uso y que crea valor. Sin embargo recientemente los esfuerzos por utilizar reCAPTCHA como un sistema de digitalización de textos se están viendo frustrados por sistemas de visión por computador que son capaces de igualar con precisión la lectura de un ser humano.

Figura 3. Cuadro de diálogo con opciones disponibles para seleccionar y generar un reCAPTCHA en Google. Como puede observarse no está disponible la opción de reCAPTCHA de texto.

En las últimas implementaciones de reCAPTCHA, Google ha optado por poner un botón que dice “I’m not a robot”, dónde ahora el usuario simplemente ha de hacer clic y esperar unos segundos para poder enviar el formulario o acceder a la web. Está medida se ha tomado ya que los sistemas de visión por computador son lo suficientemente potentes como para suponer una amenaza a los reCAPTCHA tradicionales de texto.

El problema de la letra ‘A’

El equipo de Vicarious AI lleva desde 2013 trabajando sobre redes corticales recursivas. En aquel entonces anunciaron como un éxito inicial de las capacidades del sistema basado en redes corticales recursivas. Era capaz de conseguir un ratio de precisión del 66% en reCAPTCHA. Cuándo utilizamos el sistema para un estilo en concreto podían llegar hasta una precisión del 90%.

Figura 4. A la izquierda conjunto de ejemplos sobre los que Vicarious AI realizó las primeras pruebas en el año 2013. a la derecha ejemplos de las diferentes generalizaciones de la letra A, que suponen

uno de los principales problemas para la visión por computador debido a la necesidad de generalización de la letra A basándose en los conjuntos de prueba.

El problema de captcha para la inteligencia artificial es el número combinatorio de maneras en las cuál se puede mostrar una letra y se reconoció por los humanos, sin ser explícitamente entrenados en este tipo de variaciones. Ninguna de las APIs públicas de reconocimiento de caracteres ópticos (OCR) evaluado por el equipo de investigación fueron capaces de capturar esta diversidad, porque requiere que el motor de reconocimiento generalice a distribuciones que no están presentes en el conjunto de prueba y entrenamiento. Los métodos que utilizan estos motores están basados en el reconocimiento de patrones por fuerza bruta. No tienen nociones de composicionalidad, y por lo tanto ningún mecanismo para separar una letra de su fondo. De hecho, no tienen ningún conocimiento ni comprensión sobre los objetos, y por lo tanto ninguna forma de razonar sobre la forma o la apariencia de una letra concreta en soledad.

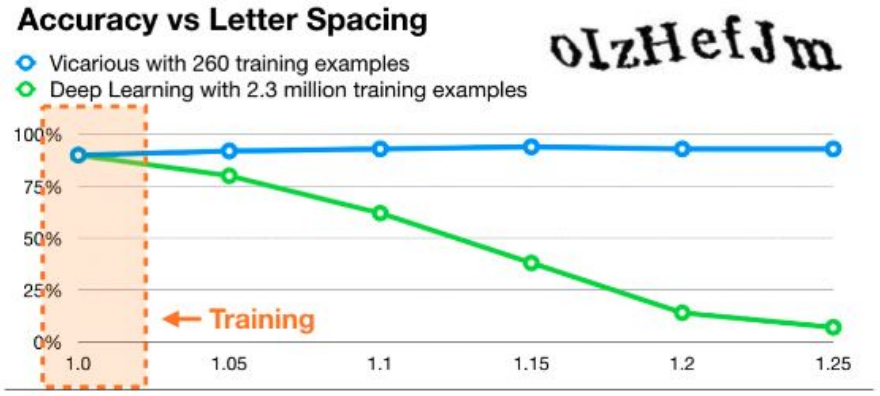


Figura 5. Muestra el porcentaje de precisión de una red neuronal convolucional entrenada con dos millones de ejemplos (en verde), y una red cortical recurrente entrenada con 260 ejemplos. En el eje de abscisas tenemos el espaciado de las letras. Se muestra el porcentaje de acierto correspondiente en función del entrenamiento en el el marco naranja.

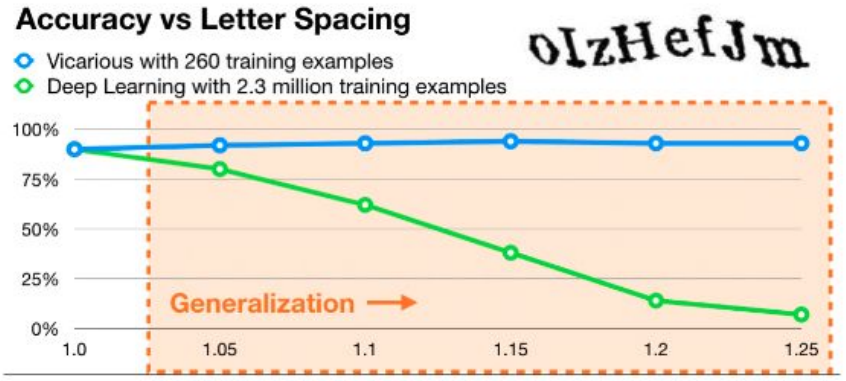


Figura 6. Mismo gráfico que el anterior. Se muestra el porcentaje de acierto correspondiente a la generalización realizada por el algoritmo.

Como se muestra en las figuras 5 y 6 los métodos como redes convolucionales entrenadas con captchas generalizan de forma muy mala cuando se varía ligeramente el espaciado de las letras individuales. Por contraposición, las redes corticales recurrentes permanecen de forma robusta aunque las letras se separen.

Douglas Hofstadter, un filósofo e influyente investigador en inteligencia artificial, bromeó con que el problema central de la inteligencia artificial es comprender la letra “A”. Vicarious AI coincide con que cualquier programa que sea capaz de gestionar las formas de las letras con la flexibilidad que un humano puede, podría poseer inteligencia artificial completa. Aunque en la actualidad las implementaciones que podemos ver en el mercado presentan precisiones superhumanas de clasificación de imágenes, y da la sensación de que el problema de la percepción está resuelto, problemas que pueden parecer simples como el reconocimiento de una letra, pueden proveer una profundidad enorme de cara al desarrollo de una inteligencia similar a la humana.

El trabajo de Vicarious AI plasmado en el paper, objeto de estudio de este trabajo, representa un pequeño paso en el camino de los computadores la comprensión de la forma de las letras, con la flexibilidad y fluidez de la percepción humana. Incluso con sus avances, seguimos estando lejos de haber resuelto el siempre problema que nos propone Hofstadter detectando las “A” con la misma fluidez y dinamismo que los humanos.

Modelo generativo de reconocimiento jerárquico de formas

La habilidad humana para reconocer objetos es invariante a los cambios de apariencia drásticos. Si viéramos, por ejemplo, un árbol completamente azul por primera vez en nuestra vida, seríamos capaces de reconocer correctamente el objeto como un árbol identificar su color como azul, aunque nunca antes hubiéramos visto un arbol azul, me supiéramos lo que es. Dejando de lado la sorpresa inicial, no estaríamos confundidos inclinados a pensar que es un arándano gigante. Esto sugiere fuertemente que somos capaces de percibir la forma de los objetos independientemente de su apariencia y que nuestra categorización de los objetos recaen más fuertemente en pistas y señales de la forma que en pistas y señales de la apariencia.

De forma similar, incluso si nunca hemos visto un árbol azul completamente, somos capaces de imaginarlo componiendo nuestro modelo de árbol, que contribuye con su forma, con nuestra idea de azul, que contribuye con la apariencia. Nuestra

representación interna de los objetos es capaz de factorizar la forma y la apariencia y componer la de cara a formar objetos.

Basándonos en las observaciones anteriores, parece razonable esperar que un modelo de una imagen con capacidades de reconocimiento del nivel de un ser humano, tenga una representación factorizada de la forma y la apariencia. Muy pocos trabajos e investigaciones han buscado la factorización de la forma y la apariencia para el reconocimiento de imágenes en la actualidad.

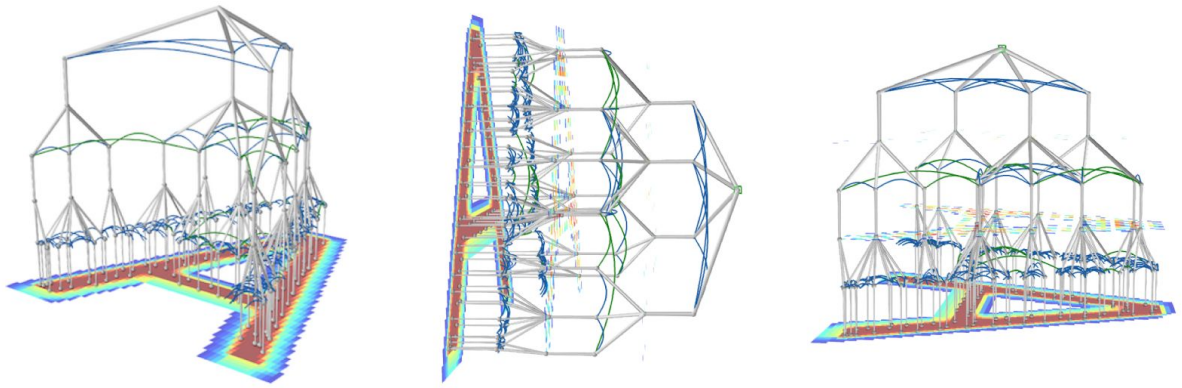


Figura 7. Modelo jerárquico de la forma y límites interiores y exteriores de la letra A representado en tres dimensiones.

Esta factorización permite a un modelo generalizar desde menos ejemplos, ya que los datos de entrenamiento que necesita solo deben contener imágenes con la suficiente diversidad de formas y apariencias (y no todas las combinaciones de ellas), para que el modelo sea capaz de generar una situación que es capaz de manejar el espacio completo del producto cruzado. Si nos fijamos en las redes neuronales convolucionales, incluso con el éxito que tienen, mezclan la forma y la apariencia. Estos modelos son incapaces de reconocer en tiempo de prueba objetos con una apariencia que cambie significativamente de otras que hayan visto para eso objeto en particular en el conjunto de prueba, es decir, probablemente fallen en la prueba del árbol azul. La solución para estos modelos es aumentar el conjunto de entrenamiento para incluir imágenes con más combinaciones de formas y apariencias. Aunque usar árboles de cada color posible durante el entrenamiento de la red neuronal puede resolver el problema del árbol azul, usar esta solución de forma generalizada representa una complejidad que vuelve el problema irresoluble utilizando esa solución.

El modelo propuesto por Vicarious AI asume que la forma y la apariencia han de ser factorizadas, y genera imágenes combinando estos dos elementos. Primero se

genera la forma del objeto, que define los límites internos y externos del mismo. Luego es “coloreado” rellenando las regiones del interior del objeto con la apariencia (colores o texturas).

Inferencia

El modelo general imágenes correspondientes a diferentes objetos en solitario. Una vez que los parámetros del modelo jerárquico han sido definidos utilizando una colección de entrenamiento de objetos aislados, el reconocimiento se suma a la inferencia en el modelo probabilístico. En particular, reconocer un objeto aislado se suma a inferir el valor de la agrupación que nos informa de la categoría y ubicación del objeto.

Cuándo existen múltiples objetos en la misma imagen, podemos seguir usando el procedimiento anterior para encontrar los objetos que más destacan en la imagen, enmascarándolos de forma secuencial y volviendo a buscar por los objetos más destacados.

Figura 8. Muestra como la máscara elimina de la imagen el objeto más destacado permitiendo encontrar en la siguiente iteracion nuevos objetos destacados delante o detrás de este.

La inferencia exacta en un modelo tan complejo y recursivo es, por supuesto, intratable. Para obtener una inferencia válida el modelo ofrece una solución aproximada utilizando diferentes técnicas. Utilizando un único forward-pass de abajo hacia arriba es suficiente para encontrar una buena aproximación de la ubicación y categorías de los objetos, y el subsecuente backward-pass de arriba a abajo será suficiente para encontrar las máscaras correspondientes a cada uno de los objetos reconocidos.

Aprendizaje

El algoritmo es capaz de aprender un modelo jerarquico de los datos. Existen dos componentes del modelo que el algoritmo ha de aprender: características y conexiones laterales. Las características se aprenden de un conjunto de

entrenamiento con imágenes procesadas, que son capaces de extraer los contornos, utilizando aprendizaje de diccionario sin supervisión y código escaso.

Las conexiones laterales para las capas de agregación se aprenden de los contornos de la entrada, y sus perturbaciones son generadas paramétricamente utilizando un modelo de perturbación proporcionado. Dado que cada una de las capas se aprende de forma independiente, se pueden reemplazar de forma individual sin que afecte al conjunto.

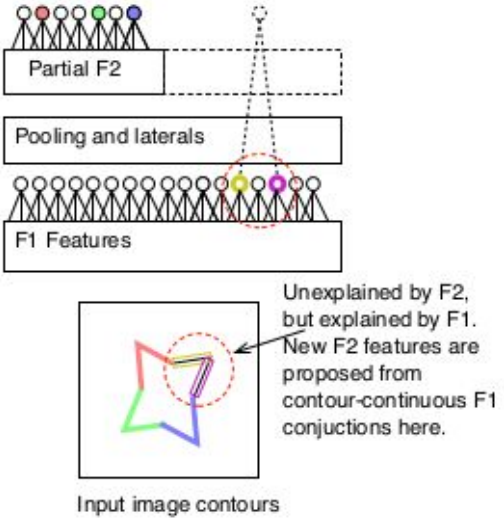


Figura 9. Aprendiendo características en un segundo nivel de característica. los círculos coloreados representan activaciones de características. el círculo discontinuo es una característica propuesta por el propio modelo que ha sido generalizada.

El aprendizaje recae de forma principal en información sin clasificar. La clasificación y etiquetado de la información es necesaria sola y exclusivamente para características de nivel superior, y este paso solamente necesita una única imagen etiquetada por característica clasificadora o clase.

Los algoritmos que presenta Vicarius AI también hacen una asunción más sobre la jerarquía. Asumen que hay un mapeo directivo de cada característica en un nivel dado, a una agregación en el nivel superior que se dice que está centrada en esa característica.

Experimentos

Se realizaron con la red cortical recursiva múltiples experimentos para comprobar su eficiencia. cubriré algunos de ellos y explicaré cómo se hicieron.

reCAPTCHA

El equipo de investigación se descargó 5500 imágenes de reCAPTCHA de Google. Utilizó 500 de esas imágenes para componer los conjuntos de pruebas. Una vez el sistema estaba entrenado fue capaz de resolver con un 84,2% de precisión cualquier reCAPTCHA generado.

En conjunto, el 66,6% de todos los reCAPTCHA que se presentaron un sistema fueron identificados correctamente con todas sus letras. El reconocimiento de caracteres permitió que en el 94,3% de los casos se reconocieran los caracteres correctamente.

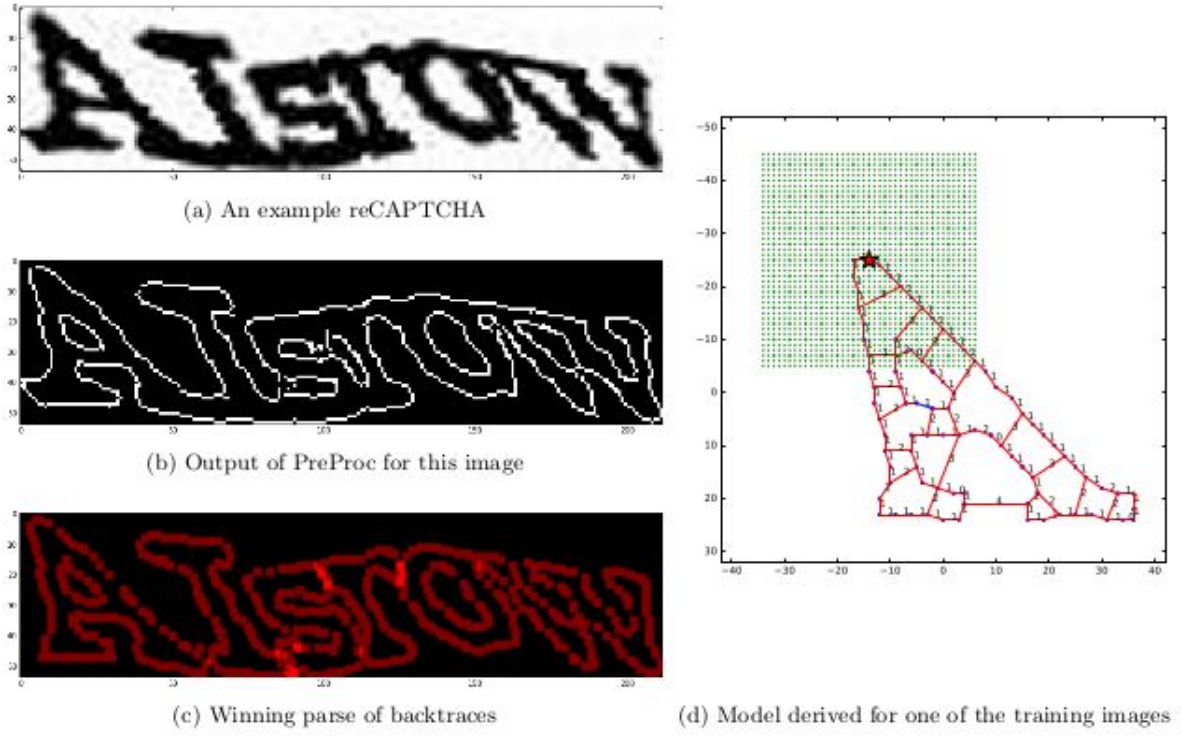


Figura 10. En esta figura podemos observar un recaptcha de ejemplo (a), la salida del preprocesamiento de la imagen (b), la fase de la RCN que genera texto válido para el captcha (c), y el modelo derivado de uno de los ejemplos de entrenamiento, con los miembros agregados para una característica concreta expandidos como si fueran puntos (d).

La red se entrenó en una única instancia de Amazon Web Services. Concretamente se realizó en un único núcleo de una instancia m4.xlarge. el entrenamiento de la seda se realizó en tan solo 67 segundos. Las 5000 imágenes fueron analizadas en 94 segundos de media cada por núcleo. Teniendo en cuenta la capacidad de paralelización que tiene el proceso de inferencia, Nico reCAPTCHA podría ser analizado por un MacBook Pro retina del año 2012 en 31 segundos.

La precisión de el ser humano en reCAPTCHA fue estimada por los empleados de Amazon Mechanical Turk. Obtuvieron una precisión del 87,4%, y si uso el mismo banco de pruebas que ha usado Vicarious para el entrenamiento de su red cortical recursiva.

Aplicación sobre objetos en tres dimensiones

Durante el experimento y la investigación que realizaron los investigadores, entrenar un también una red cortical recursiva para el reconocimiento de imágenes en tres dimensiones. En este experimento se ve como este tipo de red cortical recursiva es capaz de resolver problemas en un contexto más general.

Figura 11. Imágenes de ejemplo de conjunto de entrenamiento. Cada una representa una categoría de las que se le mostró a la red cortical recursiva.

Para este experimento concreto se le dieron 10 categorías diferentes de objetos en tres dimensiones, cada una con 5 objetos diferentes desde 12 perspectivas diferentes, que en total supusieron 600 ejemplos de entrenamiento. El conjunto de datos de las pruebas contenía escenas complejas con objetos que no habían sido vistos durante el entrenamiento. Las escenas eran conjuntos ordenados de diferentes objetos, con algunas parcialmente ocultados, sobre fondos aleatorios.

Los modelos generados por la red cortical recursiva fueron capaces de reconocer objetos tridimensionales en escenas con fondos aleatorios que no habían sido vistos nunca con un gran porcentaje de éxito.

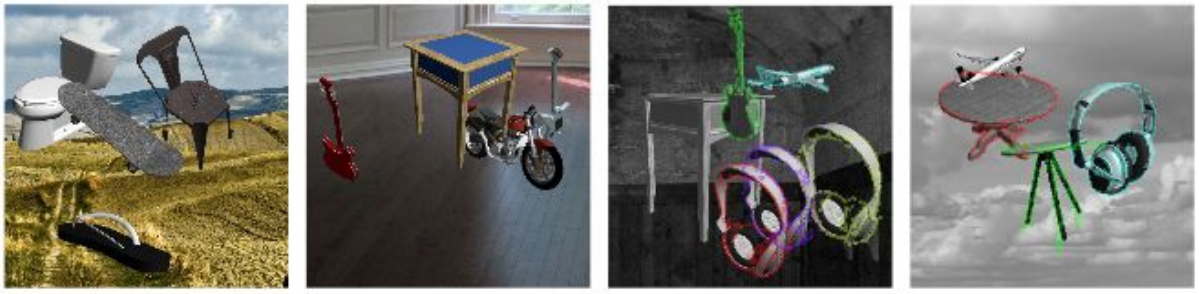


Figura 12. A la izquierda podemos ver imágenes de ejemplo del conjunto de pruebas, se usaban escenas mezcladas dónde se veían múltiples objetos en fondos aleatorios. A la derecha se pueden ver como el algoritmo era capaz de inferir los objetos y bordearlos.

Sin embargo durante los test también se observaron que a medida que se iba aumentando el número de objetos en las imágenes y el ruido, se encontraban falsos positivos y había ocasiones en las que no se detectan objetos, sobre todo en aquellos en los que los colores de los bordes eran muy parecidos con los del fondo. En algunos casos tampoco era capaz de detectar determinados ángulos para objetos como la silla. La red neuronal capaz de obtener rotaciones de hasta 20 grados sobre los modelos que no había visto nunca, pero tenía que almacenar un gran número de instancias inferidas en el nivel más alto de la red.

Bibliografía

Bursztein, Elie, et al. “How Good Are Humans at Solving CAPTCHAs? A Large Scale Evaluation.” *2010 IEEE Symposium on Security and Privacy*, 2010, doi:10.1109/sp.2010.31.

George, Dileep, et al. “A Generative Vision Model That Trains with High Data Efficiency and Breaks Text-Based CAPTCHAs.” *Science*, vol. 358, no. 6368, 2017, doi:10.1126/science.aag2612.

Krizhevsky, Alex, et al. “NIPS Proceedingsβ.” *ImageNet Classification with Deep Convolutional Neural Networks*, 1 Jan. 1970, papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.

Liang, Ming, and Xiaolin Hu. “Recurrent Convolutional Neural Network for Object Recognition.” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, doi:10.1109/cvpr.2015.7298958.

Liao, Qianli, and Tomaso Poggio. “Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex.” *[1604.03640] Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex*, 13 Apr. 2016, arxiv.org/abs/1604.03640.

“Maldición De La Dimensión.” *Wikipedia*, Wikimedia Foundation, 8 Dec. 2017, es.wikipedia.org/wiki/Maldición_de_la_dimensión.

Mei, Jieru, et al. “Scene Text Script Identification with Convolutional Recurrent Neural Networks.” *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016, doi:10.1109/icpr.2016.7900268.

Pinheiro, Pedro, and Ronan Collobert. “Recurrent Convolutional Neural Networks for Scene Labeling.” *PMLR*, 26 Jan. 2014, proceedings.mlr.press/v32/pinheiro14.html.

SinaHonari. “SinaHonari/RCN.” *GitHub*, 4 Apr. 2017, github.com/SinaHonari/RCN.

Spoerer, Courtney J., et al. “Recurrent Convolutional Neural Networks: A Better Model of Biological Object Recognition.” *Frontiers in Psychology*, vol. 8, Dec. 2017, doi:10.3389/fpsyg.2017.01551.

“Visión Artificial.” *Wikipedia*, Wikimedia Foundation, 8 Dec. 2017, es.wikipedia.org/wiki/Visión_artificial.

Voges, N., and L. Perrinet. “Complex Dynamics in Recurrent Cortical Networks Based on Spatially Realistic Connectivities.” *Frontiers in Computational Neuroscience*, vol. 6, 2012, doi:10.3389/fncom.2012.00041.

Wu, Jian, et al. “Delving Deeper into Convolutional Neural Networks for Camera Relocalization.” *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, doi:10.1109/icra.2017.7989663.

Yamins, Daniel L K, and James J Dicarlo. “Using Goal-Driven Deep Learning Models to Understand Sensory Cortex.” *Nature Neuroscience*, vol. 19, no. 3, 2016, pp. 356–365., doi:10.1038/nn.4244.