Evaluación Automática de la Calidad de Páginas Web Mediante Deep Learning

Christian Mejía, Miguel Cazorla y Ester Martínez-Martin

Universitat d'Alacant
Universidad de Alicante

Objetivos

General

Desarrollar un sistema de Deep Learning para la evaluación y categorización de la calidad de páginas Web.

Específicos

- Crear un dataset extenso y confiable de imágenes y etiquetas de calidad de páginas Web.
- Diseñar e implementar un modelo de Deep Learning basado en una red neuronal convolucional para la evaluación y clasificación de páginas Web.
- Generar recomendaciones automáticas para mejorar la calidad de una página Web.

Introducción

El *World Wide Web* es uno de los medios de comunicación más importantes en la actualidad. Un sitio Web (colección organizada de páginas Web sobre un tema específico y vinculadas entre sí) [1], en muchas ocasiones es la opción preferida por las personas para conocer una organización con fines comerciales, académicos o laborales, además del ahorro de tiempo y dinero que significa [2]. Un sitio Web de calidad atrae y mantiene la atención del usuario, incrementa descargas, registros y suscripciones, transmite confiabilidad y credibilidad, provocando una mejor impresión de la organización [3]. Por tanto, **es de gran relevancia analizar la calidad de un sitio Web**, en especial de su página principal (*homepage*), por ser la más representativa de todo el sitio. La calidad incluye el diseño estético, la estructura y funcionalidad, el contenido, la seguridad, las herramientas tecnológicas, la popularidad, y muy posiblemente otros componentes más [4]. **No existe una lista de criterios de calidad** constante a nivel mundial, varios países han definido sus propias directrices pero han demostrado no coincidir, dependen de la cultura y asignan diferente importancia a un mismo criterio [5]. La tarea de evaluar la calidad de un sitio Web se vuelva compleja y subjetiva para el ser humano, sobre todo la parte estética, ya que las preferencias varían sustancialmente de un individuo a otro [3]. Utilizar una computadora con Inteligencia Artificial y **entrenar un sistema de Machine Learning** con los datos adecuados, captaría de manera automática patrones de calidad, resultando un enfoque más prometedor [4].

Propuesta

Planteamos el diseño e implementación de un algoritmo clasificador que:

1. Reciba como entrada un *screenshot* de la página principal del sitio Web. No se extrae código HTML, comúnmente utilizado en trabajos similares. La página Web se analizará tal como aparece ante el usuario, por lo que será independiente del lenguaje y la tecnología de implementación.
2. Evaluará la imagen buscando las características de calidad identificadas en el proceso de entrenamiento. La técnica de *Deep Learning* denominada ***Red Neuronal Convolucional*** (*CNN* o *ConvNet*) evita la definición manual de características al extraer automáticamente los patrones más relevantes en los datos suministrados.
3. Clasificará entre varias categorías, el nivel de calidad de la página Web, misma que no fue observada antes por el sistema.
4. Emitirá recomendaciones para la mejora de calidad de la página Web.

Hipótesis

El problema de categorización de la calidad de un sitio Web puede ser tratado con un modelo de *Deep Learning* basado en una *CNN*, ya que permite la extracción automática de características, las cuales manualmente son difíciles de establecer.

Arquitectura del Sistema

Las *CNN* constituyen el *estado del arte* dentro del campo de la visión por computadora. Serán entrenadas para categorizar páginas Web en niveles de calidad. Con *Aprendizaje supervisado* se detectará y extraerá automáticamente las características de calidad desde un dataset inicial confiable. Lo aprendido será transferido a una red neuronal que recibirá los *screenshots* de un dataset extenso y los clasificará por categorías de calidad.

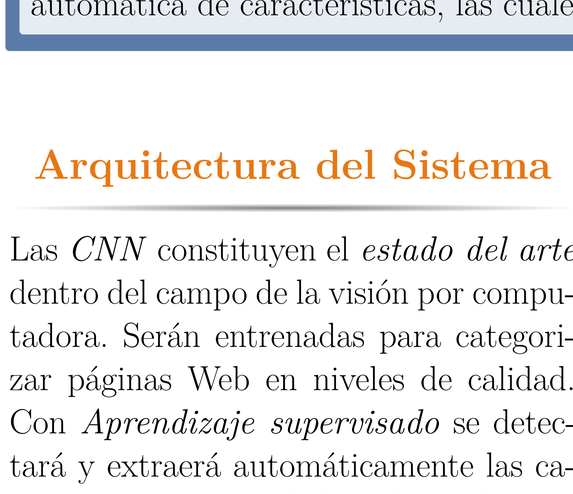

Figura 2:Arquitectura del sistema

Problemáticas

El éxito del proyecto depende de un conjunto de datos (*dataset*) extenso y confiable, que tenga asociadas sus correspondientes salidas o respuestas correctas (etiquetas) para todas las imágenes. Tales datasets no siempre están disponibles y recolectarlos consume mucho tiempo y esfuerzo. No existe un dataset de *screenshots* de páginas Web suficiente y correcto para nuestro propósito. El primer reto es generar los datos de entrenamiento (recopilación y etiquetado) sabiendo que hacerlo a mano es costoso y uno de los cuellos de botella en *Machine Learning*. Además, no son imágenes del mundo real, como fotografías que incluyen personas, animales, plantas, u objetos, donde sus características son bien definidas, sino que son *screenshots* de páginas Web, cuyos elementos son más difíciles de detectar pues mezclan texto, gráficos y multimedia.

Metodología

Las actividades por cumplir durante la investigación son:

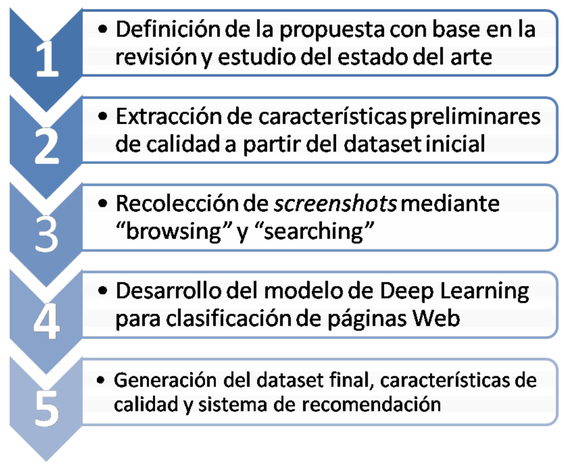

Figura 1:Flujo de Actividades

Los plazos y resultados asociados a cada una de las actividades planteadas:

No.	Año	Resultado
1	1	Plan de Investigación
2	1	Métricas iniciales
3	2	Dataset sin etiquetar
4	3	Modelo de *CNN*
5	3	Clasificador y recomendador

Tabla 1:Cronograma

Contribuciones

El sistema clasificador y recomendador podría tener numerosas aplicaciones:

- Soporte para el trabajo de diseño o rediseño, tanto para principiantes como expertos.
- Preselección de las mejores páginas.
- Sugerir guías de diseño y métricas más importantes que pueden ser tomadas por los diseñadores como prioritarias al momento de crear un sitio Web, ahorrando tiempo.
- Mejorar el desempeño de motores de búsqueda como *Google* y *Bing* que necesitan optimizar la calidad de los resultados para sus usuarios.
- Ser incorporado en sistemas de recomendación, indexadores de páginas Web, y bases de conocimiento.
- Apoyo para concursos donde se premian a las mejores páginas y sitios Web.

Por otra parte, el dataset estaría disponible para ser reutilizado en otros proyectos de investigación científica y empresarial.

Referencias

[1] Ebubekir Buber and Banu Diri.
Web Page Classification Using RNN.
Procedia Computer Science, 154:62–72, 2018.

[2] Radek Burget and Ivana Rudolfová.
Web page element classification based on visual features.
Proceedings - 2009 1st Asian Conference on Intelligent Information and Database Systems, ACIIDS 2009, pages 67–72, 2009.

[3] Masoud Ganj Khani, Mohammad Reza Mazinani, Mohsen Fayyaz, and Mojtaba Hoseini.
A novel approach for website aesthetic evaluation based on convolutional neural networks.
2016 2nd International Conference on Web Research, ICWR 2016, pages 48–53, 2016.

[4] Minoi Jacey-Lynn, Alvin W. yeo, and Abdelnour-Nocera Jose.
Designing For Global Markets 10.
(May 2015), 2011.

[5] M. Y. Ivory, R. R. Sinha, and M. A. Hearst.
Empirically validated web page design metrics.
Conference on Human Factors in Computing Systems - Proceedings, pages 53–60, 2001.

Más información

- Website: https://osf.io/7ghd2
- Email: cme26@alu.ua.es