



Matching images is hard!

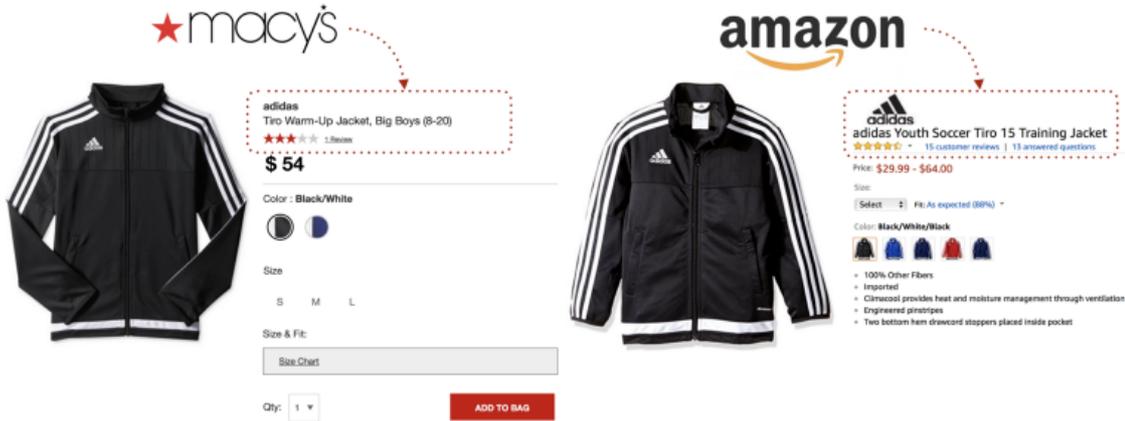
Images, intrinsically, are complex forms of information, with varying backgrounds, orientations, and noise. Developing a reliable system that achieves human-like accuracy in identifying, interpreting, and comparing images, without investing in expensive resources, is no mean task.

For DataWeave, however, the ability to accurately match images is fundamental to the value we provide to retailers and consumer brands.

Why Match Images?

Our customers rely on us for timely and actionable insights on their competitors' pricing, assortment, promotions, etc. compared to their own. To enable this, we need to identify and match products across multiple websites, at very large scale.

One might hope to easily match products using just the product titles and descriptions on websites. However, therein lies the rub. Text-based fields are typically unstructured, and lack consistency or standardization across websites (especially for fashion products). In the following example, the same Adidas jacket is listed as "Tiro Warm-Up Jacket, Big Boys (8-20)" on Macy's and "Youth Soccer Tiro 15 Training Jacket" on Amazon.



Hence, instead of using text-based information, we considered using deep-learning techniques to match the images of products listed on e-commerce websites. This, though, requires massive GPU resources and training data fed into the deep-learning model—an expensive proposition.

The solution we arrived upon, was to complement our image-matching system with the text-based information available in product titles and descriptions. Analyzing this combination of both text- and image-based information enabled us to efficiently match products at greater than 90% accuracy.

How We Did It

A couple of weeks ago, I gave a talk at Fifth Elephant, one of India's renowned data science conferences. In the talk, I demonstrated DataWeave's innovation of augmenting the NLP capabilities of Solr (a popular text search engine) with deep-learning features to match images with high accuracy.

Check out the video of the presentation for a detailed account of the system we built:



Human-Aided Machine Intelligence

All products matched with the seed product are tagged with a corresponding confidence score. When this score crosses a certain threshold, it's presumed to be a direct match. The ones that are part of a lower range of confidence scores are quickly examined manually for possible direct matches.

The outcome, therefore, is that our technology narrows down the consideration set of possible product matches from a theoretical upper limit of millions of products, to only a few tens of products, which are then manually checked. This unique approach has two distinct advantages:

- The human-in-the-loop enables us to achieve greater than 90% accuracy in matching millions of products—a key differentiator.
- Information on all manually matched products is continually fed to the deep-learning model, which is used as training data, further enhancing the accuracy of the product matching mechanism. As a result, both our accuracy and delivery time keep improving with time.

As the world of online commerce continues to evolve and becomes more competitive, retailers and consumer brands need the ability to make quick proactive and reactive decisions, if they are to stay competitive. By building an automated self-improving system that matches products quickly and accurately, DataWeave enables just that.

Find out more about how retailers and consumer brands use DataWeave to better understand their competitive environment, optimize customer experience, and drive profitable growth.

- **DataWeave Marketing**

9th Aug, 2017

ARTIFICIAL INTELLIGENCE

E COMMERCE

