Using Google Vision to Tag Greeting Cards

In February 2018, RJB used Google Vision to tag products with search keywords for an existing e-commerce customer. What makes this project different? Well, most e-commerce systems display images of their products (mostly without text) combined with HTML product descriptions that are easily index-able by search engines. This case was quite different – to find out why read on.

The Client

The client is an e-commerce vendor that sells greeting cards from independent artists with your personalized message in your own handwriting. The process is quite simple. Simply choose a card, write your personal message on a blank piece of paper, snap a photo with your mobile phone – and voila! The message appears on the card. Proceed to checkout and supply the name and address of the recipient – and the process is complete. The card will arrive at the recipients address at the specified time with your personal message inside in your own handwriting.

The Use Case

Greeting cards can present a bit of a problem for search indexing. Most cards contain meaningful graphic images. Many cards contain inscriptions, either on the front cover or inside or both. These inscriptions are images, not documents. Most inscriptions are in an elegant or stylized script that may not use a recognizable font. Greeting cards often express a particular relationship between the sender and the recipient. These are important factors when designing a search indexing solution.

For example, a birthday card may have an inscription with an age and relationship e.g. Happy 84th Birthday, Grandpa! Other cards might be humorous, including humor containing mature content. Some cards might be more appropriate to give to a man or a woman e.g. a Happy Retirement! card with an image of a man on a golf course. Since there are thousands of cards available, we want to be able to assist the buyer in finding just the right card for the occasion, and in some cases, a card that reflects the relationship of the sender and recipient (see The Greeting Card as Emotional Commodity).

Just to throw a little twist into the indexing, it is important to consider bias and subjectivity. While one might expect that *most* human beings could come up with some keywords to tag *most* cards, not all human beings will pick the same keywords. We know this from practical experience in tagging cards. Tests of card tagging conducted with popular crowdsourcing platforms such as Amazon's Mechanical Turk have shown that gaining plurality (i.e. consensus) on appropriate keywords is more difficult than one might imagine. Simple tagging situations, like the aforementioned birthday card are easy. However, cards that are less specific can pose a problem. How would one tag these cards?

Use of human labor to tag inexpensive cards can be problematic. Factored in with the Long Tail, where some cards might be best sellers and some rarely sell, it is important to consider cost as well as quality and consistency of card tagging. Offloading tagging cost to the independent artists that upload the cards often does not achieve the desired quality and consistency.

To summarize, we want to make use of images, graphics, and stylized script in each card to make our best determination of whether the card is for a holiday, special occasion, or personal message. In some cases, we want to know if there is an explicit or implicit relationship between sender and receiver. Finally, we want to know whether specific objects are shown graphically, or indicated by text e.g. dog, balloon etc.

From an operational standpoint, we want the card tagging to be near real-time, relatively cost effective, accurate and unbiased.

The Solution

While search engines have improved in indexing images, they fall short of addressing the requirements we have laid out. For that reason we looked to artificial intelligence (AI) solution(s) specifically targeted at image recognition to see what they had to offer. We selected Google Vision and IBM Watson Visual Recognition for a proof of concept.

Both offered a drag and drop web interface where we could test various greeting cards to see what information was returned. Both also offered an API for programmatic access. Due to the volume of cards we ultimately processed, early on we switched to the API. We used a Private Beta version of Watson Visual Recognition that included a Text Model. Since our participation in the Private Beta necessitated that we sign a non-disclosure agreement (NDA) we will not be publishing those results in this post (we will likely publish in an upcoming post when the Text Model is Generally Available (GA)).

We randomly selected greeting cards, presented those same cards to each vision system, and took note of the JSON results returned.

Our summarized results for using Google Vision with the ***DOCUMENT_TEXT_DETECTION*** Feature were as follows:

- We processed ~2500 cards, each with a cover image and optional inside image. Both images can optionally have text
- We detected an occasion (e.g. Christmas, Birthday) on ~1700 cards – some cards legitimately are not for a particular occasion
- Where detected we had an almost 100% accuracy rate with unambiguous occasions like Birthday, Christmas, Valentines

We did find some cases that baffled the Google Vision pattern recognition. The following 3 cards differ only by color. What Google Vision saw is printed below.

Cover

Inside

```
"HAPPY Birthday
"
```

```
"der eine run flappiestwa me
"
```

Cover

Inside

```
"HAPPY Birthday
"
```

```
"GO
der ne van flappiestwamem
"
```

Cover

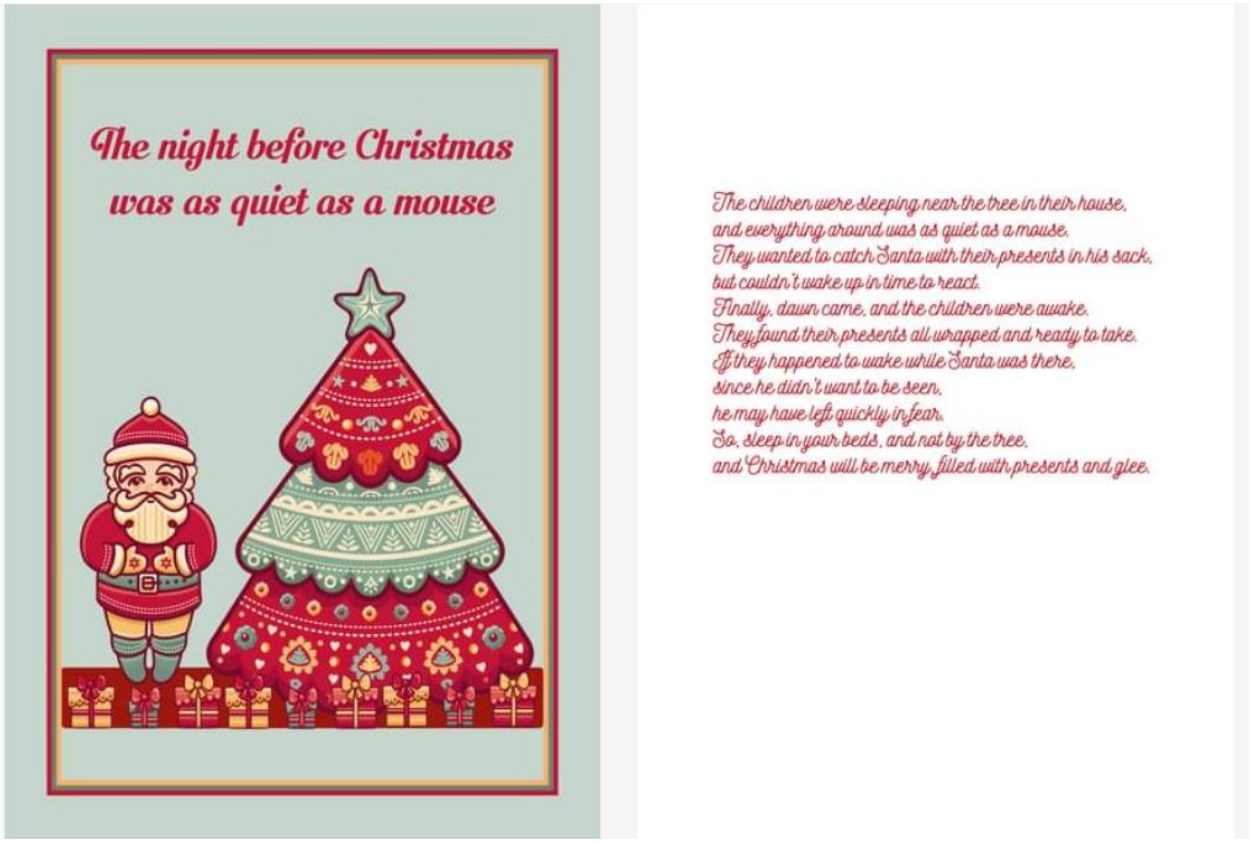
Inside

```
"HAPPY Birthday
"
```

```
"MAY IT BE THE
may in en Happicctoare
ONE YET!
"
```

Google Vision generally has some trouble with very exotic fonts. For all 3 cards, Google Vision was able to see "HAPPY Birthday". In the last card Google Vision saw "may it be the" and "one yet" on the Inside – but in the 1st two cards it saw a mangled form of German.

Here is a sampling of a few more cards along with what Google Vision “saw” in each. To reiterate, we used the ***DOCUMENT_TEXT_DETECTION*** feature for these tests. This will give some idea of the capability of the tool to accurately interpret stylized text lifted from graphical images.




```
"The night before Christmas
was as quiet as a mouse
See
"
```

```
"The children were sleeping near the tree in their house,
and everything around was as quiet as a mouse,
They wanted to catch Santa with their presents in his sack
but couldn't wake up in timeta react.
Finally, dawn came, and the children were awake.
They found their presents all wrapped and ready ta take,
It they happened to wake while Santa was there,
since he didn'tuant to be deen.
he may have left quickly in fear
So deep in your beds, and not by the tree,
and erristmas will be merry filled with presents and glee.
"
```

No corrections have been made to the results. Notice that the captured text also includes line breaks matching the original. Highlights in yellow have been added (by RJB) where the text deviates from the original.

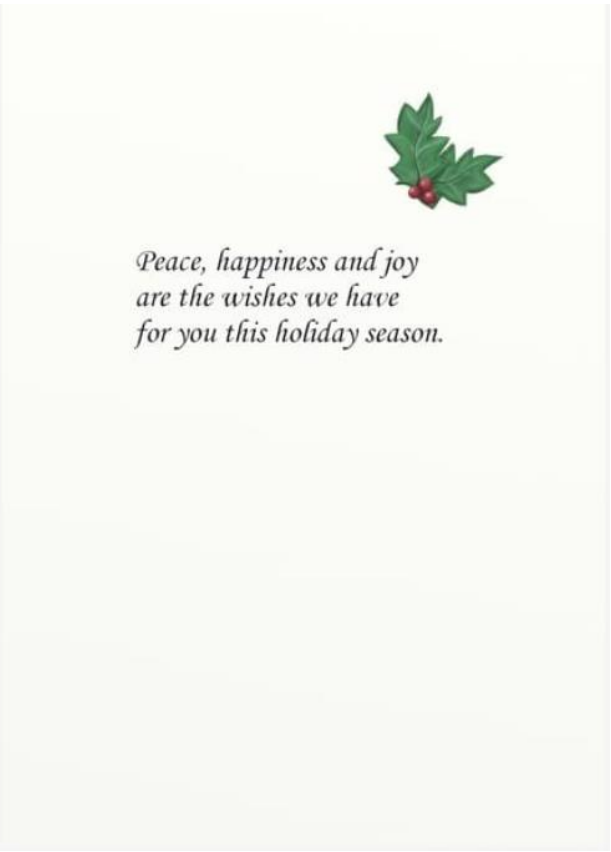
Here is another card where Google Vision didn’t do quite so well:


```
"som gerry
Christmas
- OO
"
```

```
"To my faverde
Panin the Neck
"
```

The level of stylization in the “M” of Merry definitely threw off the algorithm, as did the same style on the inside cover.

On the other hand,




```
"Merry
Christmas
"
```

```
"Peace, happiness and joy
are the wishes we have
for you this holiday season.
"
```

worked out perfectly. Even though the lettering is stylized, the level of stylizing did not exceed the capability of the pattern recognition.

Let's review what we are trying to achieve.

1. We want to make use of images, graphics, and stylized script in each card to make our best determination of whether the card is for a holiday, special occasion, or personal message. (Convert to italicized quote or block quote)

Google Vision seems to have significant capability to extract text from images, even when that text is stylized – very heavily stylized text may pose a problem for the pattern recognition.

Once the text is parsed, given that the text contains some reference to the holiday, special occasion, or personal message, we can do pattern matching to get that information from the text. While Google Vision itself is not doing the matching, it is providing the capability to make that the 2nd step of the process.

The level of accuracy of text extraction has the added potential benefit of providing a form of product description from the greeting card itself. Typical e-commerce offerings consist of a product image, and a separate product description often created by the seller. External search engines such as Google, Yahoo, Bing etc. typically index on product names and descriptions. With the use of Google Vision we open the door to high quality indexing on greeting card images by simply incorporating the extracted text into the product detail page.

To continue our review of what we are trying to achieve, recall that:

2. We want to know whether specific objects are shown graphically, or indicated by text e.g. dog, balloon etc. (Convert to italicized quote or block quote)

Since we have already discussed extraction of text, let's turn to extraction of information related to images. For this test we'll use the ***LABEL_DETECTION*** feature of Google Vision. Take this greeting card for example:

Google sees quite a bit in the image, but there is one small problem. Here is the result – omitting some results where Google had a low score (confidence) level:

Description	Score
dog like mammal	0.9549466
mammal	0.9264370
dog	0.9255791
dalmatian	0.9198946
dog breed	0.9133532
vertebrate	0.9037037
christmas	0.8415493

Now we know that all of Google Vision's high scoring answers are correct in this instance, but as a human being in this context is it useful to know that we are seeing vertebrates? (Google expressed 90% confidence that we are).

Let's suppose we were planning on using this result for the purpose of search indexing – we would definitely want "dog", and "Dalmatian" as keywords. "Christmas" might be another keyword of interest. The rest are debatable. Is there any harm in keeping all of the keywords? Perhaps not for indexing an *internal site search engine*, although it is a bit of a stretch to envision our customers looking for greeting cards with images of "dog like mammals". We probably don't want to make these other keywords visible to external search engines like Google, Yahoo, Bing, and the rest as they may bias the site SEO rankings. So on balance, we don't need them and should find a way to remove them. We could use a heuristic to remove unwanted keywords. For example, keywords like mammal or vertebrate show up frequently for images of animals, birds etc. – we could introduce a *negative list* of keywords we don't want included and add those to the list.

Another point regarding internal site search is that we can use the Score to boost search results. Since Google Vision has reported that it is 91.99% confident that the dog in the image is a Dalmatian, we can pass that on to our site search engine as a *boost* factor to rank it high on a search for "Dalmatian". Unfortunately, we can't boost external search engines like Google, Yahoo, Bing, etc. For those search engines, we need to make sure that the keywords discovered by both ***LABEL_DETECTION*** and ***DOCUMENT_TEXT_DETECTION*** are included in the product detail page content for crawlers to find.

Continuing our solution requirements review we have:

3. In some cases, we want to know if there is an explicit or implicit relationship between sender and receiver. (Convert to italicized quote or block quote)

In many cases this is simpler than it sounds. Given that there are a finite number of relationships to deal with, if there are relationship indicators in the text portion of the greeting card they can be compared against a relationship list to make a determination. To go back to our earlier example, the Happy 84th Birthday, Grandpa! clearly implies a grandchild giving a card to a grandfather. RJB implemented this as a follow on step after the visual text extraction. Cards that have no text inscription are more difficult to process, and may also have no implied relationship.

And our final requirement:

4. From an operational standpoint, we want the card tagging to be near real-time, relatively cost effective, accurate and unbiased. (Convert to italicized quote or block quote)

Google Vision is a Cloud based service, accessible through an Application Programming Interface (API). As such, a single greeting card can be processed in

seconds yielding the results we have described. From the Google pricing page one can see that it costs $1.50 USD to use one Vision Feature on 1,000 images (the 1st 1,000 images are free). Since a greeting card has a cover image and an optional inside image, and we are using 2 features, that's up to 4 API calls per card. Therefore, our cost to process 250 greeting cards will not exceed $1.50.

Can AI vision produce biased results? To the extent that Machine Learning (ML) is used to train the system it is technically possible because the training data itself could be biased. We are not currently using a custom model for Google Vision – we are relying on the general model that comes with the product. Bias in AI systems is a big concern in some cases. Imagine AI systems that may in future make decisions that affect human beings e.g. human resource, or legal systems – we must make sure that those decisions are not biased. Companies like IBM and others are creating systems that observe AI systems at work, monitoring them to see that they do not produce biased results (see Mitigating Bias in AI Models).

An important final point to make is that AI vision systems in general are not perfect. A needed element for most operations is human augmentation of the AI results, particularly if absolute accuracy is a requirement. In our case, we built a simple editor screen for product detail information that allows a human to page through greeting cards that have been recently tagged by the AI vision system. This allows the human touch to ensure we have the results we need, while allowing a significant amount of the work to be automated.

Google Vision is a useful tool for extracting information from images contained in greeting cards for search indexing. That information can in turn enhance search engine optimization (SEO) for popular external search engines. At a fraction of a cent per card, Google Vision is a cost effective first step towards card indexing. Human activities can augment the process by providing any additional quality assurance deemed necessary.