Global Tech Mining Conference 2018, Leiden

Neural Network-Based Paper-Matching with Relevant Products through Patents

Seonho Hwang, Juneseuk Shin

Contents

1. Motivation
2. Methodology
3. Illustrative Case
4. Conclusion

Product vs. Publications (1/2)

- Planning future R&D leading to successful products in corporations
  - → Need to understand research landscape from the product perspective

Products

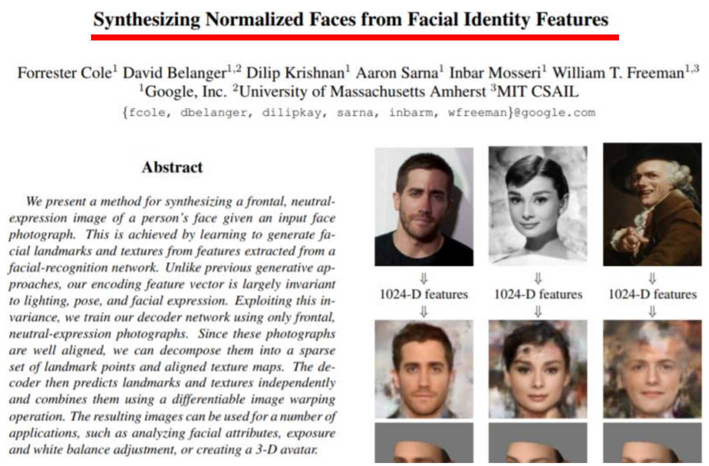
Publications

"Artifact Level"

"Knowledge Level"

But, it has not been feasible to automate the link between the two data

Product vs. Publications (2/2)

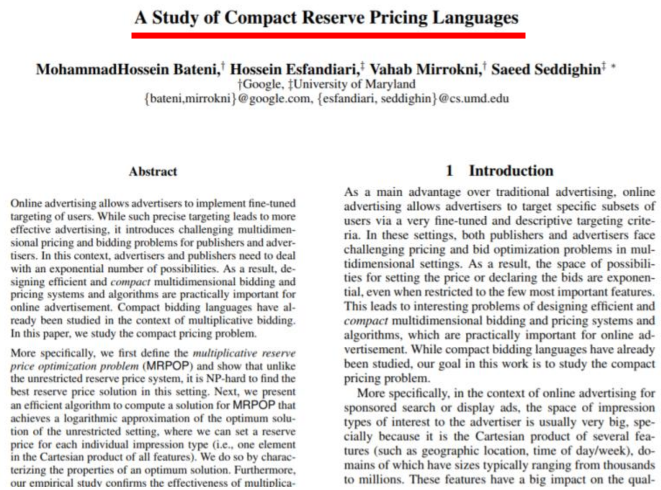


Synthesizing Normalized Faces from Facial Identity Features

Forrester Cole[1] David Belanger[1,2] Dilip Krishnan[1] Aaron Sarna[1] Inbar Mosseri[1] William T. Freeman[1,3]
[1]Google, Inc. [2]University of Massachusetts Amherst [3]MIT CSAIL
{fcole, dbelanger, dilipkay, sarna, inbarm, wfreeman}@google.com

Abstract

We present a method for synthesizing a frontal, neutral-expression image of a person's face given an input face photograph. This is achieved by learning to generate facial landmarks and textures from features extracted from a facial-recognition network. Unlike previous generative approaches, our encoding feature vector is largely invariant to lighting, pose, and facial expression. Exploiting this invariance, we train our decoder network using only frontal, neutral-expression photographs. Since these photographs are well aligned, we can decompose them into a sparse set of landmark points and aligned texture maps. The decoder then predicts landmarks and textures independently and combines them using a differentiable image warping operation. The resulting images can be used for a number of applications, such as analyzing facial attributes, exposure and white balance adjustment, or creating a 3-D avatar.



A Study of Compact Reserve Pricing Languages

MohammadHossein Bateni,† Hossein Esfandiari,‡ Vahab Mirrokni,† Saeed Seddighin‡ *
†Google, ‡University of Maryland
{bateni,mirrokni}@google.com, {esfandiari, seddighin}@cs.umd.edu

Abstract

Online advertising allows advertisers to implement fine-tuned targeting of users. While such precise targeting leads to more effective advertising, it introduces challenging multidimensional pricing and bidding problems for publishers and advertisers. In this context, advertisers and publishers need to deal with an exponential number of possibilities. As a result, designing efficient and *compact* multidimensional bidding and pricing systems and algorithms are practically important for online advertisement. Compact bidding languages have already been studied in the context of multiplicative bidding. In this paper, we study the compact pricing problem.

More specifically, we first define the *multiplicative reserve price optimization problem* (MRPOP) and show that unlike the unrestricted reserve price system, it is NP-hard to find the best reserve price solution in this setting. Next, we present an efficient algorithm to compute a solution for MRPOP that achieves a logarithmic approximation of the optimum solution of the unrestricted setting, where we can set a reserve price for each individual impression type (i.e., one element in the Cartesian product of all features). We do so by characterizing the properties of an optimum solution. Furthermore, our empirical study confirms the effectiveness of multiplica-

1 Introduction

As a main advantage over traditional advertising, online advertising allows advertisers to target specific subsets of users via a very fine-tuned and descriptive targeting criteria. In these settings, both publishers and advertisers face challenging pricing and bid optimization problems in multidimensional settings. As a result, the space of possibilities for setting the price or declaring the bids are exponential, even when restricted to the few most important features. This leads to interesting problems of designing efficient and *compact* multidimensional bidding and pricing systems and algorithms, which are practically important for online advertisement. While compact bidding languages have already been studied, our goal in this work is to study the compact pricing problem.

More specifically, in the context of online advertising for sponsored search or display ads, the space of impression types of interest to the advertiser is usually very big, specially because it is the Cartesian product of several features (such as geographic location, time of day/week), domains of which have sizes typically ranging from thousands to millions. These features have a big impact on the qual-

↕ "Image" ↕ "Advertisement"

(1) Manual matching can be possible but takes too much resource

(2) Patents are of both artifact and knowledge levels!!!

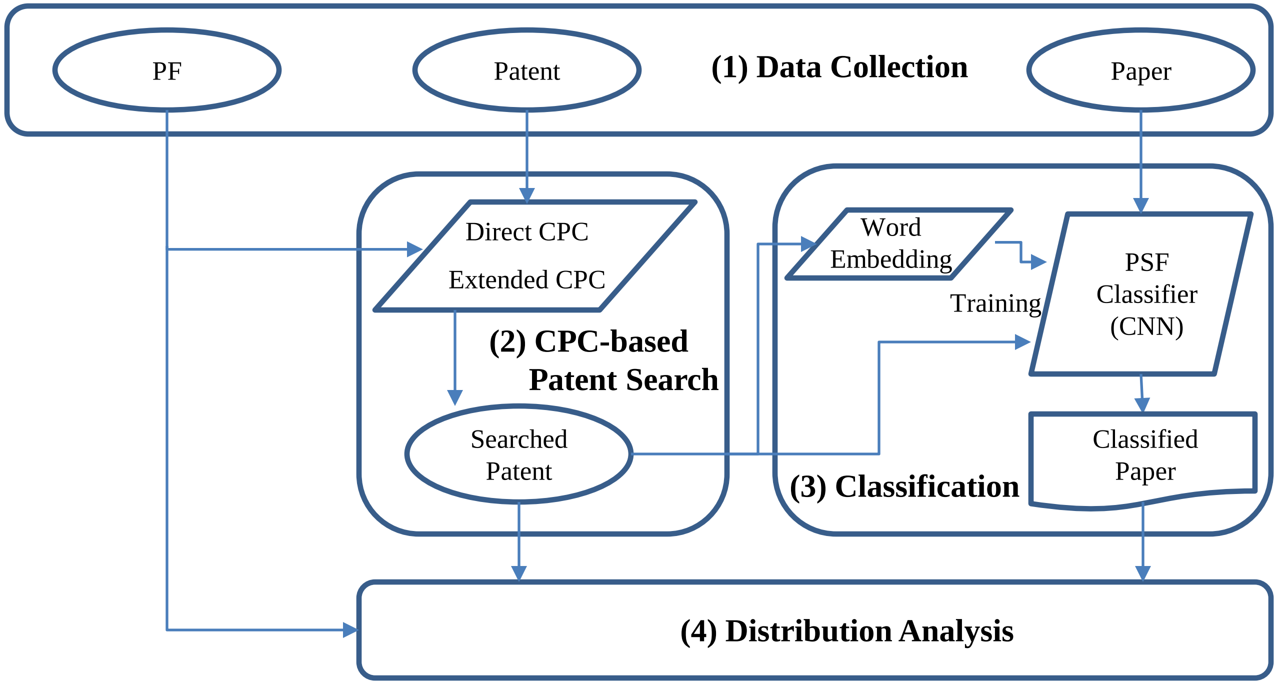
→ Patents could bridge the level difference between products and papers

But, how?

Contents

1. Motivation
2. Methodology
3. Illustrative Case
4. Conclusion

Overall Process



(1) Data Collection : PF (Product Field), Patent Data, Papers

(2) CPC-based Patent Search : Direct CPC (DCPC), Extended CPC (ECPC)

(3) Classification : Training word2vec and CNN, and then classifying papers

(4) Distribution Analysis : Analyzing the distributions of the classified papers

CPC-based Patent Search

- Objective : To search patents corresponding to each PF

- Primary vs Secondary CPCs corresponding to a patent

Patent Number	Title	Primary CPC (pCPC)	Secondary CPC (sCPC)
US20170115853	Determining Image Captions	G06F3/04842	G06F17/30247, G06K9/00456, G06K 9/344, G06T7/0081, G06F3/0482, G06 K2209/01, G06T2207/10016
US9460348	Associating location history with photos	G06K9/00677	None

- Direct CPC (DCPC) corresponding to a PF

PF	DCPC	CPC Description
Image	G06T1/00	General purpose **image** data processing
	G06F17/30247	using image data, e.g. **images**, **photos**, **pictures** taken by a user
	⋮	⋮

- PPAT (Primary PATent)

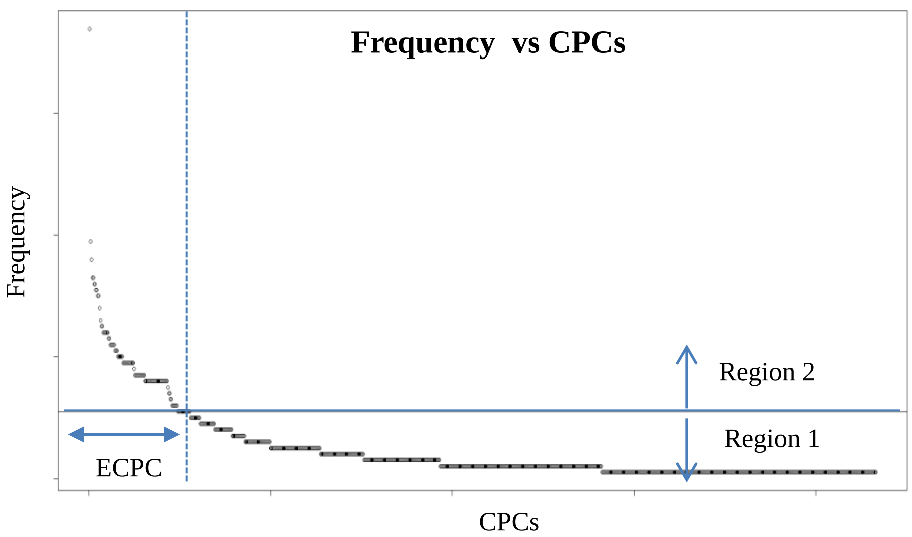
  is defined to be the **patents whose pCPCs are DCPC of the PF**

CPC-based Patent Search

➢ Needs for Extension

Patent Number	Title	Primary CPC	Secondary CPC
US20150169186	METHOD AND APPARATUS FOR SURFACING CONTENT DURING IMAGE SHARING	H04L67/10	G06F17/30, **G06F17/30247**, G06K9/00 677, G06Q10/10, G06Q50/01, H04L67/14, H04N5/2251 H04W4/206

➢ Extended CPC (ECPC) corresponding to a PF



Algorithm

1. Search patents such that DCPCs in pCPC
2. Extract all CPCs from the patents
3. Take only sCPCs from the CPCs
4. Plot Frequency vs CPCs
5. Cluster the frequency into two regions using *k*-means

   Where, Region 2 : higher frequency

   Region 1 : lower frequency
6. Select CPCs whose frequency is in Region 2 → ECPC found

CPC-based Patent Search

➢ Patent Search

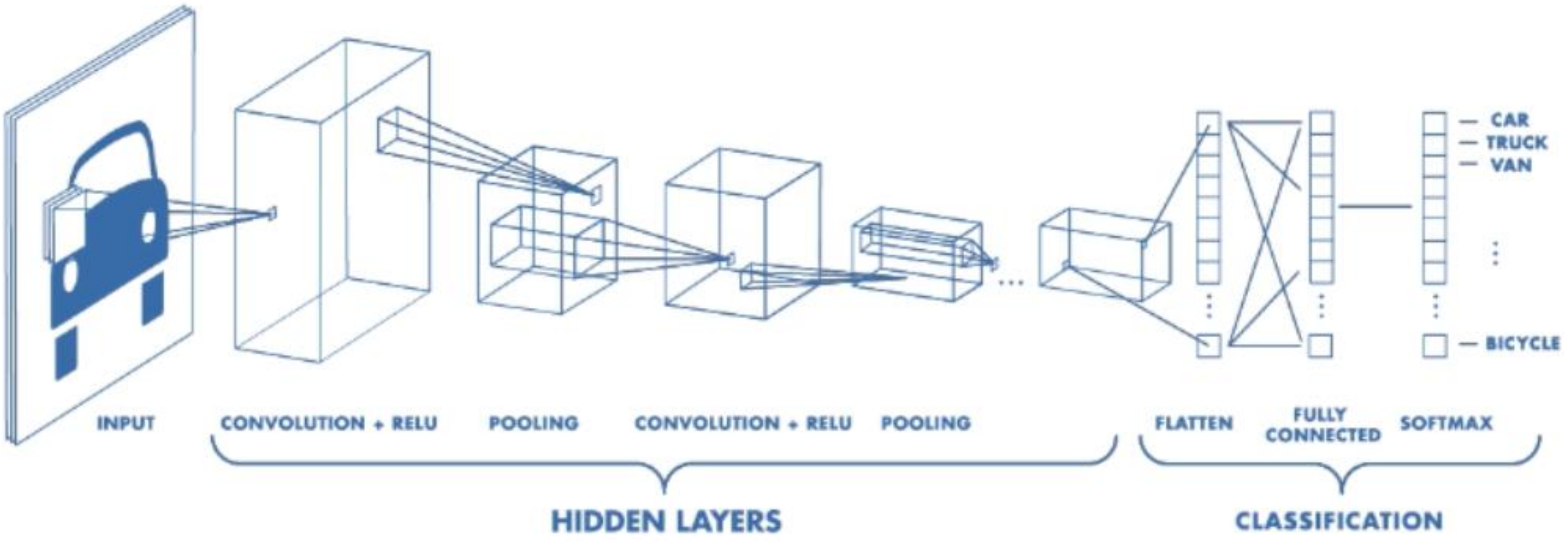
PF	**PF1**		**PF2**	
pCPC	DCPC1	ECPC1	DCPC2	ECPC2
sCPC	None	DCPC1	None	DCPC2
Searched Sets	PPAT1	EPAT1	PPAT2	EPAT2
Searched Patents	**Union (PPAT1, EPAT1)**		**Union (PPAT2, EPAT2)**	

* EPAT : Extended PATent

Classification

- Objective : To classify papers according to PFs

- Convolutional Neural Network (CNN) for Image

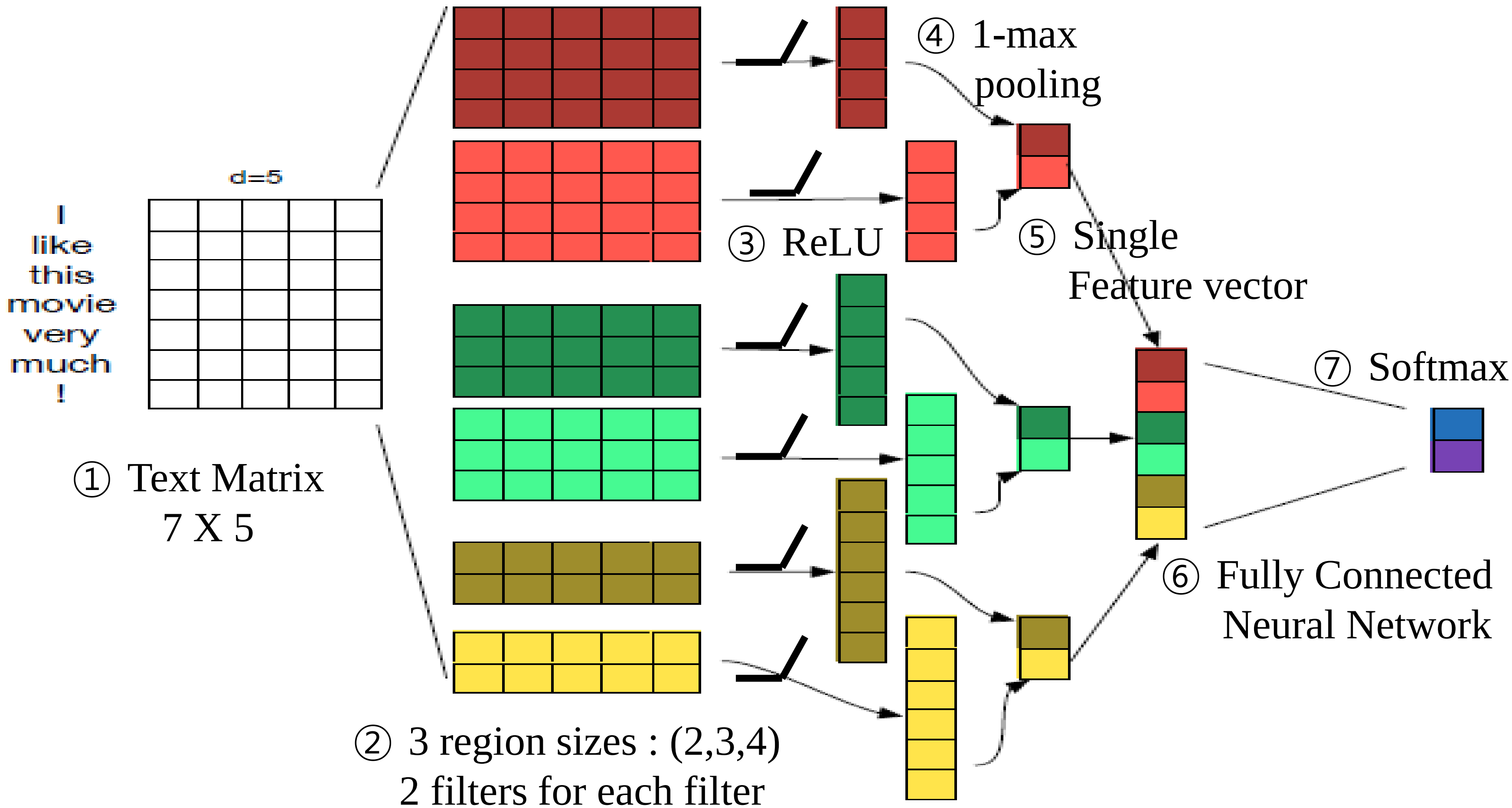


Architecture of a CNN.—Source: https://www.mathworks.com/videos/introduction-to-deep-learning-what-are-convolutional-neural-networks--1489512765771.html

Excellent performance for image classifications thanks to the capability to extract local features that differentiate one image from another

Classification

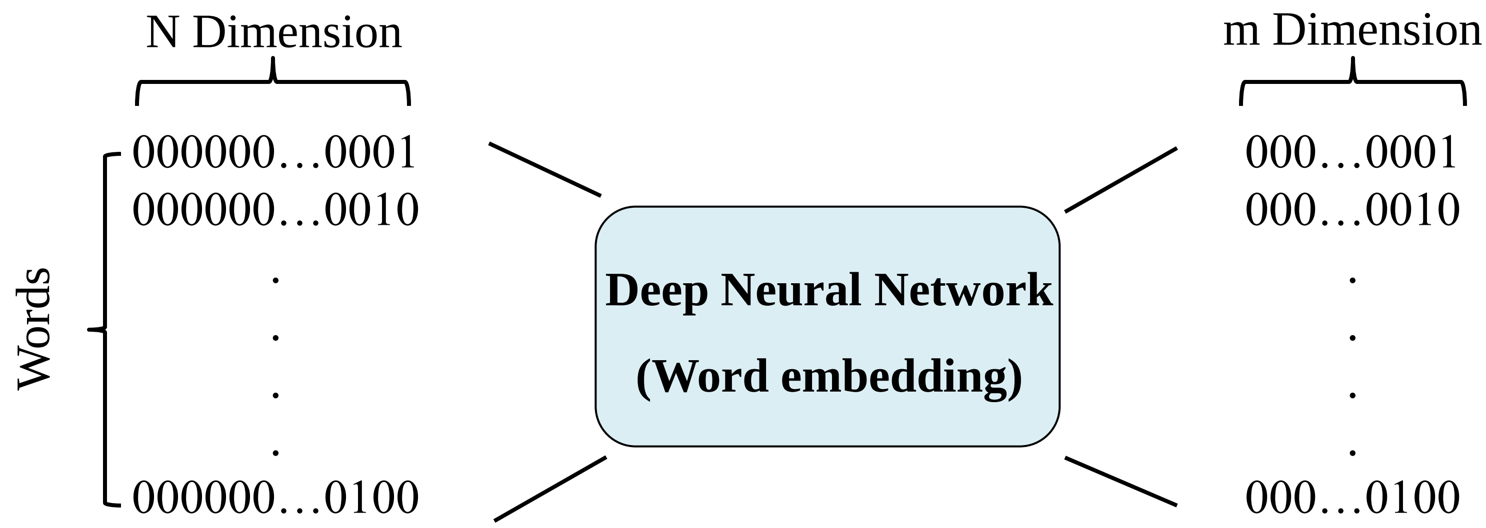
- Convolutional Neural Network (CNN) for Text



(Zhang and Wallace, 2016)

Classification

➢ Word Embedding - Introduction



(1) Efficiency

- Usually N >> m → Huge size of vocabulary to reasonable size of vectors

(2) Effectiveness

- Extracting features of texts based on semantic features of words

Due to the close relationship between 'classification' and feature extraction

Classification

➢ Word Embedding – Further gains

Model	MR	SST-1	SST-2	Subj	TREC	CR
CNN-static	81.0	45.5	86.8	93.0	92.8	84.7
CNN-non-static	**81.5**	**48.0**	87.2	**93.4**	**93.6**	84.3
CNN-multichannel	81.1	47.4	**88.1**	93.2	92.2	**85.0**

- Models
  . CNN-static : A model with pre-trained vectors from word2vec
  . **CNN-non-static : Same as CNN-static with additional fine tuning for each task**
  . CNN-multichannel : A model with two sets of word vectors
- Benchmark Datasets
  . MR : Movie reviews with one sentence per review
  . SST-1 : Stanford Sentiment Treebank
  . SST-2 : Same as SST-1 but with neutral reviews removed and binary labels
  . TREC : Question datasets – task involving classifying a question into 6 types
  . CR : Customer reviews of products – classifying positive/negative reviews

→ We'd better fine-tune the word embedding but we do not need multichannel

(Kim,2014)

Classification

➢ Word Embedding - Choices

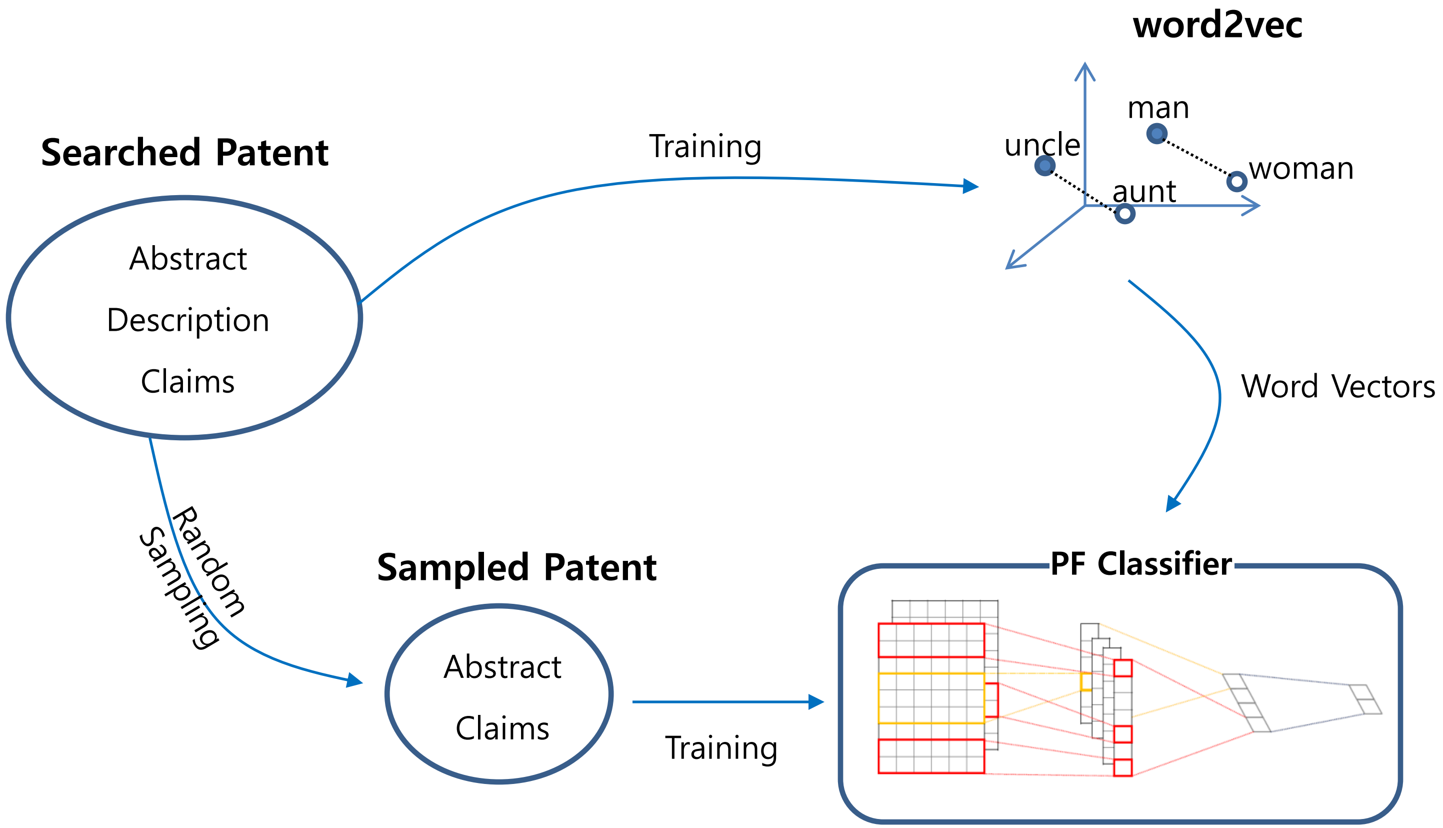
Dataset	word2vec	Glove	word2vec+Glove
MR	**81.24**	81.03	81.02
SST-1	**47.08**	45.65	45.98
SST-2	**85.49**	85.22	85.45
Subj	93.20	93.64	**93.66**
TREC	**91.54**	90.38	91.37
CR	83.92	84.33	**84.65**

(1) Word2vec performs consistently better than Glove only or multichannel word embedding

(Zhang and Wallace, 2016)

Classification

➢ Classifier Training



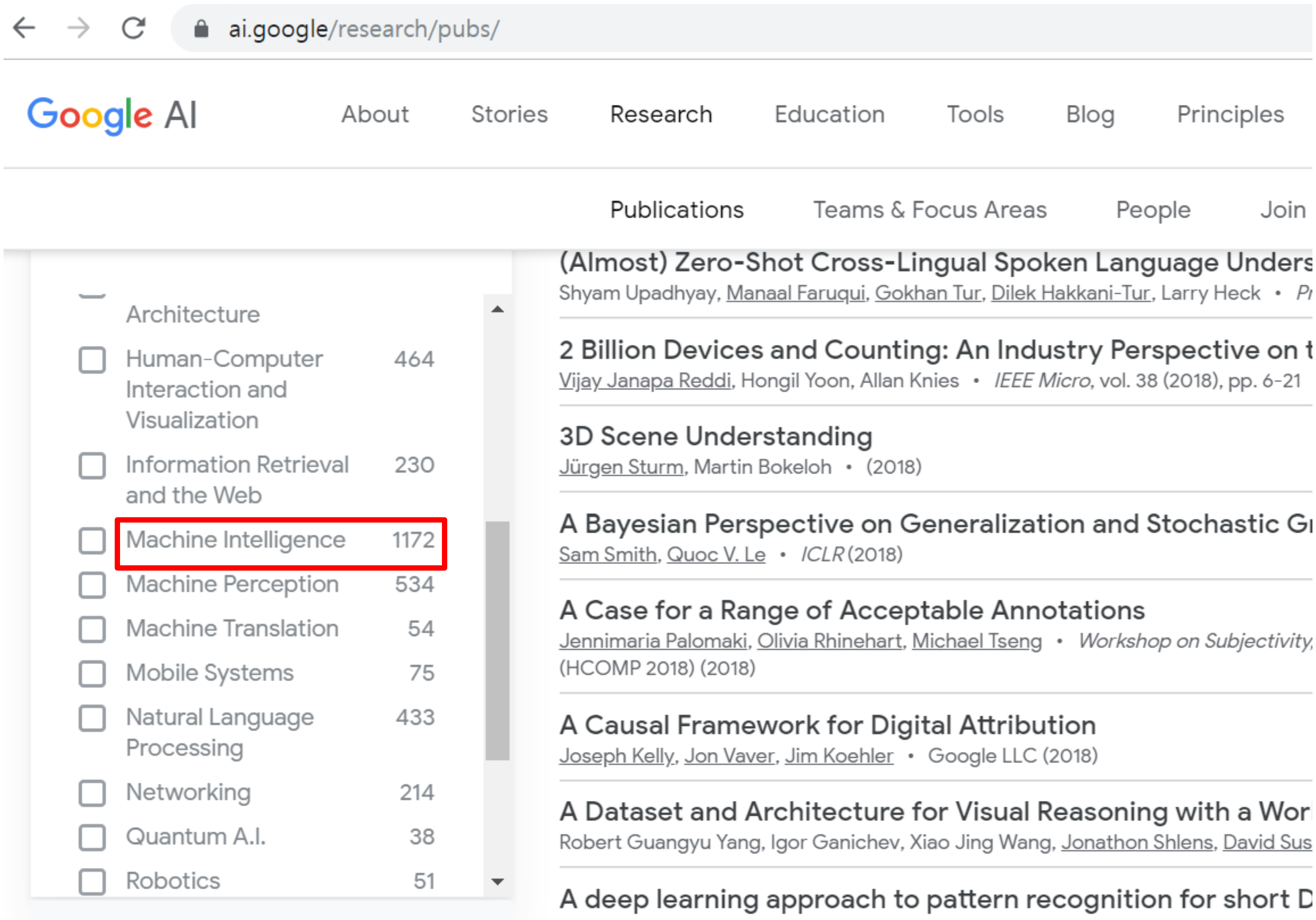
(1) Abstract, description, and claims are used for training word2vec

(2) Abstract and claims are used for training CNN-based classifier

Contents

1. Motivation
2. Methodology
3. Illustrative Case
4. Conclusion

Publications from Google Machine Intelligence



As of Sept. 2018

Data Collection

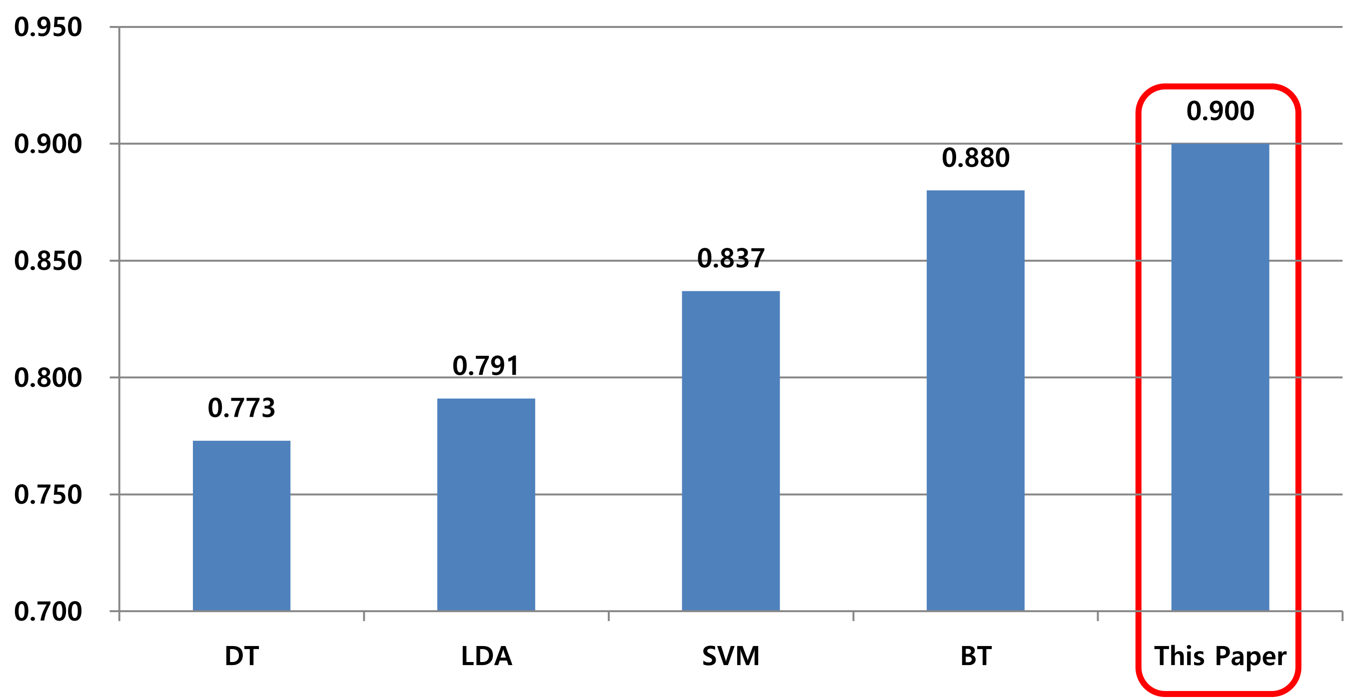
➢ Results

PF	Corresponding Google Product	CPC		Searched Patents	Papers (Machine Intelligence)
		DCPC	ECPC (Total Number)		
Ad	AdWords, AdSense	G06Q30/02*	**No Extension**	18,156	771
Image	Image Search, Google Photo	G06K9/00* G06T[1,3,5,7]/* G06F17/30047 G06F17/30247 G06F17/30256 G06F17/30265 G06F17/30271	G02B2027/0138 … H04W84/12 (111)	28,937	
Mail	GMail	H04L51/* G06Q10/107 H04L41/26	G06F17/30424 ... H04W4/206 (17)	8,470	
Map	Google Maps	G09B29/* … G06T17/05 (21)	G01C21/00 … H04W4/02 (32)	5,195	
Search	Google Search	G06F17/30*	**No Extension**	40,929	Collected as of Mar. 2018. Published only after 2010
Video	YouTube	H04N21/* … G06F17/30858 (15)	G06F17/30038 … H04N9/8205 (107)	20,724	
Total				**122,411**	**771**

1) * means a wild card. For example G06K9/00* includes all the codes starting with **G06K9/00** such as **G06K9/00**087 and **G06K9/00**14
2) [ ] means a union. For example G06T[1,3,5,7]/* includes G06T1/*, G06T3/*, G06T5/*, and G06T7/*
3) When there are too many DCPCs or ECPCs to be displayed, only the first and the last are displayed in sorted order and total number follows in parenthesis

Classifier

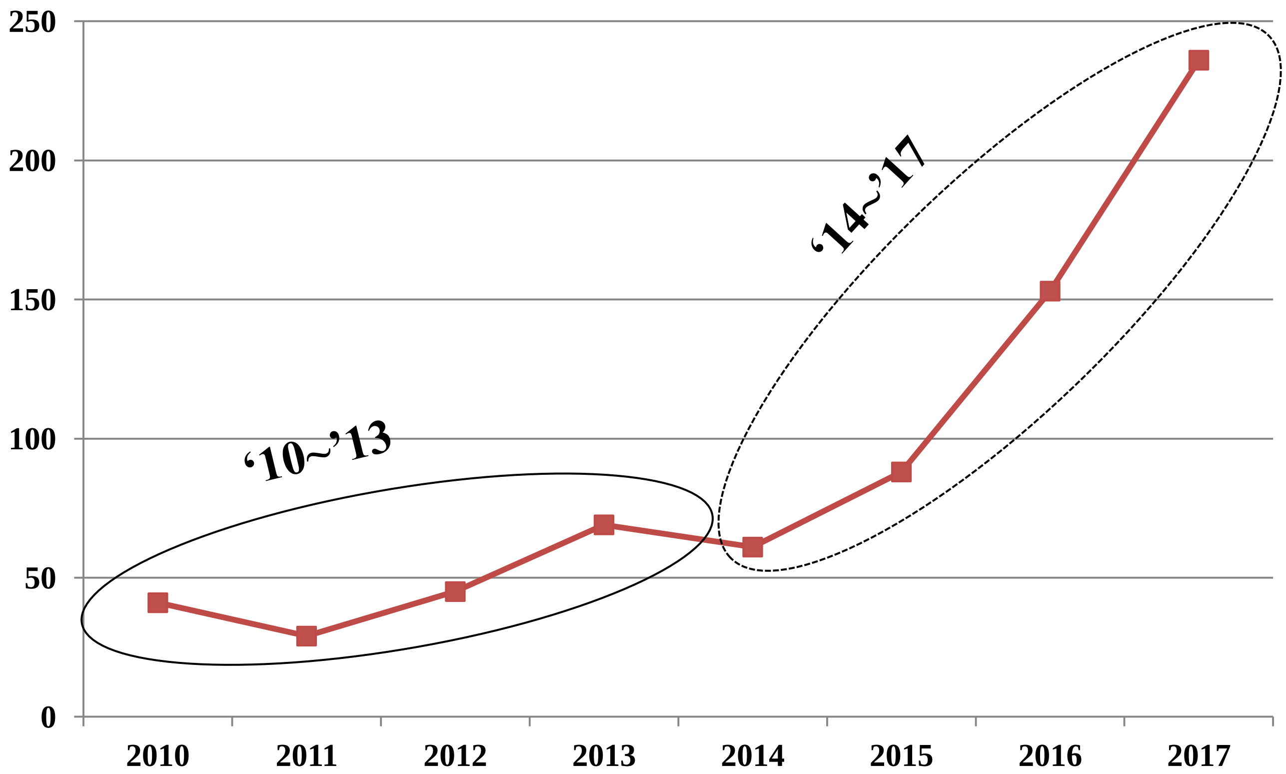
➢ Performance Comparison



(1) 5 classes for Decision Tree (DT), Linear Discriminant Analysis (LDA) Support Vector Machine (SVM1, SVM2) and 2 for Boosted Tree (BT)

Distribution Analysis

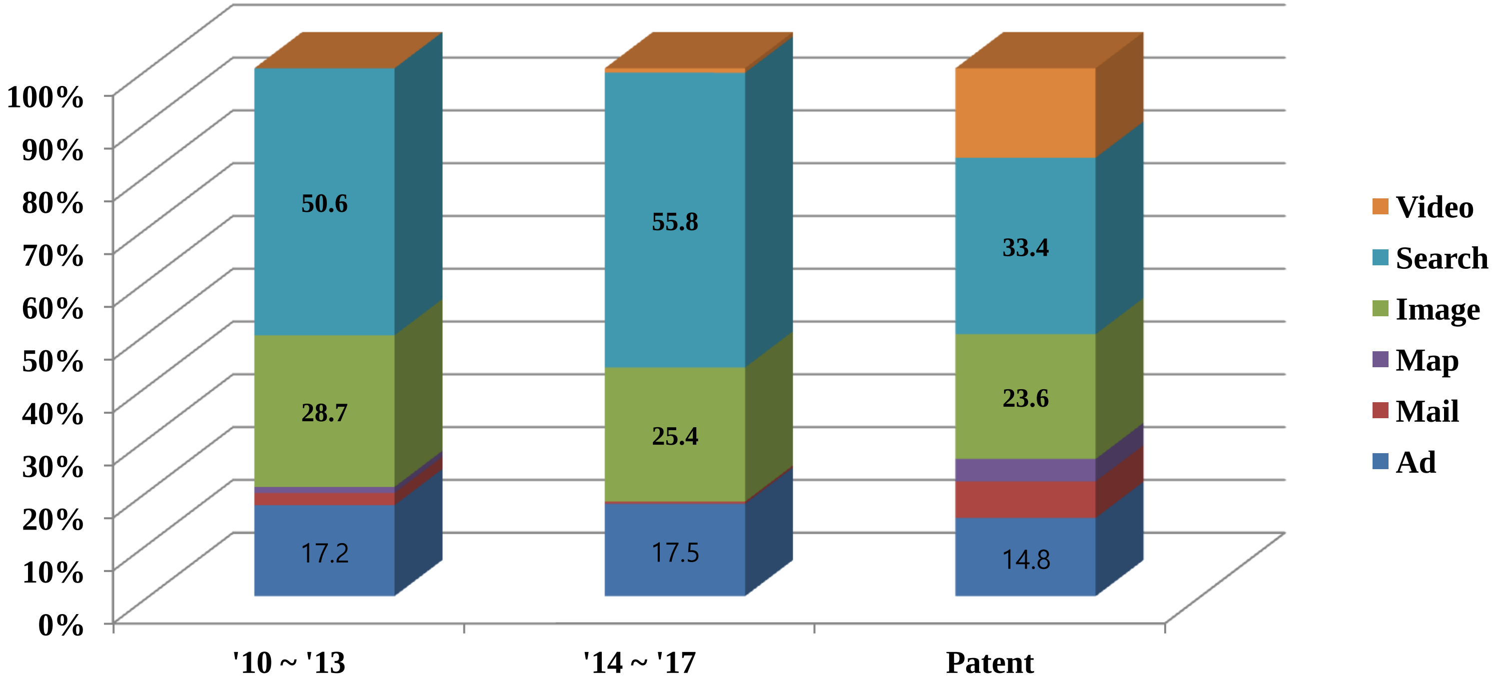
- Annual distribution of papers



(1) 722 publications in the area of machine intelligence between ‘10 and ‘17

(2) From 2014, the number increased drastically

Distribution Analysis



(1) The distributions of PFs in ‘10~’13 and ’14 ~’17 are similar

(2) Search, Image and Ad are dominant in numbers of publications

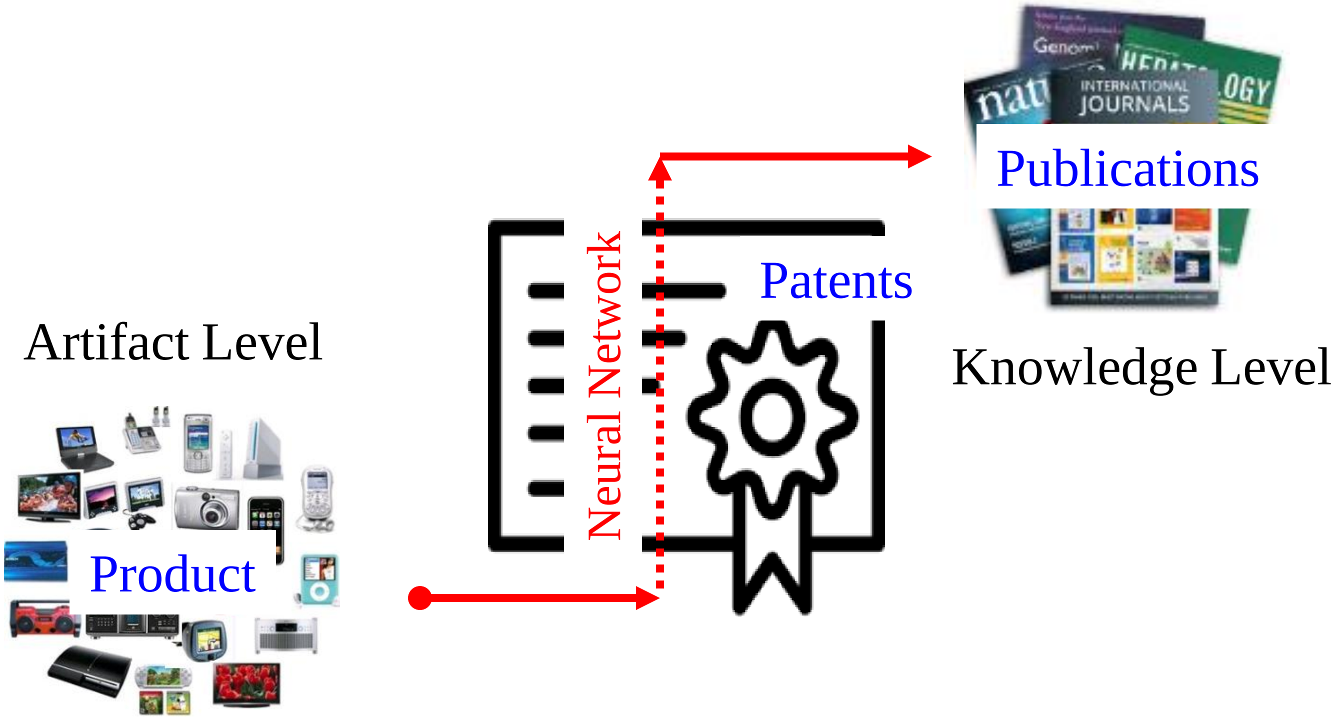
→ These fields are fundamental to other fields in Google

Contents

1. Motivation
2. Methodology
3. Illustrative Case
4. Conclusion

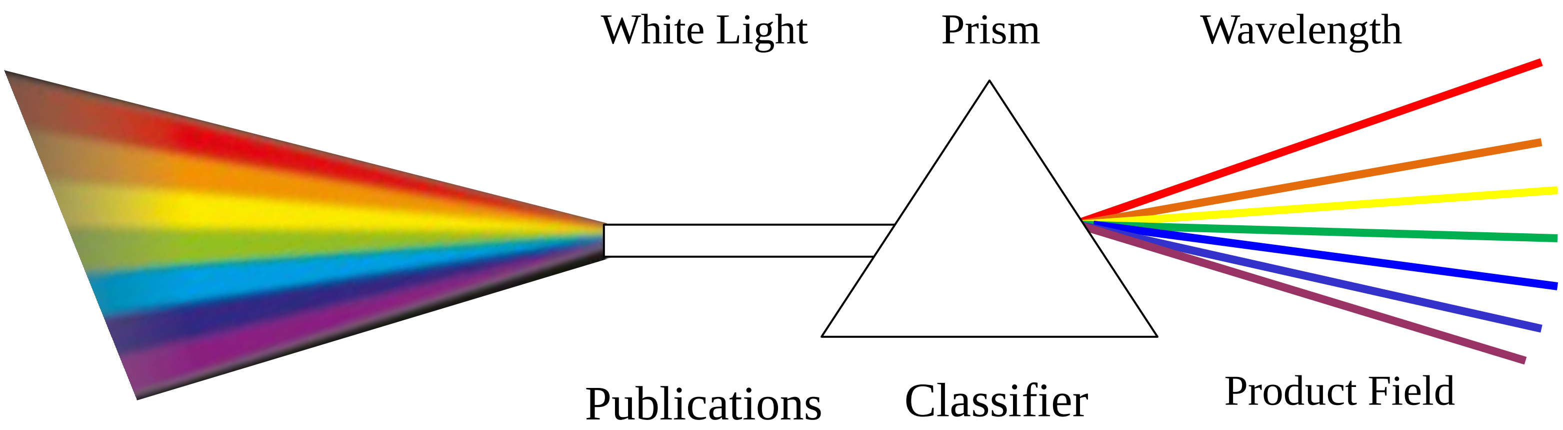
Conclusion

- ➢ Link between product (artifact level) and publications (knowledge level)
  - Patents are of both artifact and knowledge levels
  - Neural network can link data of different levels by taking advantage of patents



Implication

➢ A view on the classifier



The classifier can reveal detailed construction of publications from the product perspective like a prism for the white light

➢ Extension

. The methodology can be extended from one area in one organization to multiple areas in multiple organizations

Thank You

seonho07.hwang@gmail.com