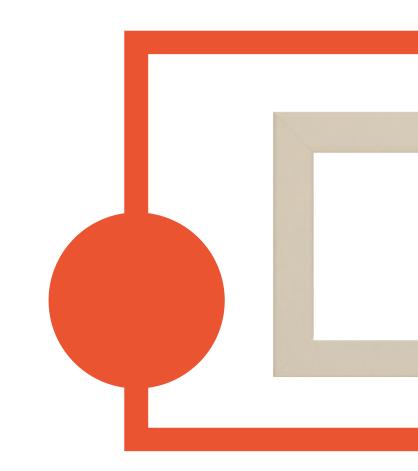
PaperFlow

2021 캡스톤 디자인 8조

20191632 윤상건 (팀장)

20160729 김태영

20181656 이민종



01 Problem

시장 조사를 통한 문제점 확인

설문 조사를 통한 문제점 확인

기존 문제점

02 Solution

PaperFlow

핵심 서비스 소개

기대 효과

03 Development

데이터셋 협업 툴

개발 기술 멘토링

역할 추가 개발 요소

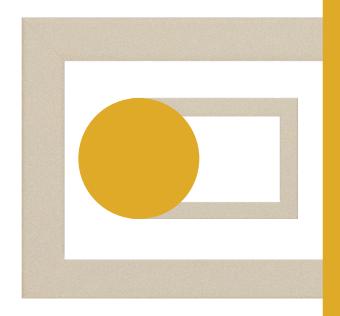
04 Appendix

논문흐름 그래프 생성 알고리즘

Use Case Diagram

배포 시스템 구조

o Problem



시장 조사를 통한 문제점 확인

대표적인 논문 검색 서비스인







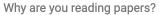
를 조사함.

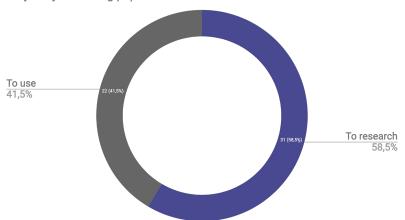
• 기존 논문 검색 플랫폼은 특정 논문에 대한 정보 전달에 집중하고 있다.

그렇기 때문에

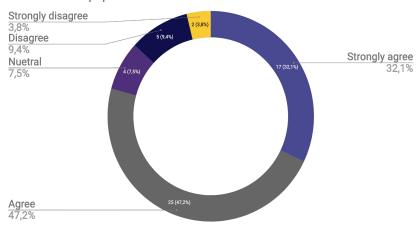
- 여러 논문 간 정보 전달 요소가 존재하지 않아 흐름 파악 어렵다.
- 특정 논문과 **연관된 핵심 논문**을 찾기 어렵다.

설문조사를 통한 문제점 확인





Have you ever thought it's hard to figure out the flow of research due to the number of papers or unfamiliar with that area?



설문 조사를 Reddit, Facebook Group에서 직접 진행

20년 1월 19일 ~ 20년 1월 23일 설문 진행 / 총 53개의 응답

Reddit Machine Learning 채널: https://bit.ly/37Mcbby / Facebook Groups: https://bit.ly/2UerC8d

기존 문제점



기존 분야의 최근 연구 흐름을 따라잡기 힘들다.

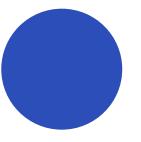


수많은 논문 중, 어떤 논문이 **핵심 논문**인지, 어떤 논문이 관심사와 가까운지 **파악**하기 힘들다.

※ 2019년, arxiv.org에서 CS분야 논문만 46,514건 (모든 분야에 대해서는 약 16만건)

02

Solution



PaperFlow

특정 논문이나 연구 분야에서의 동향에 대해

쉽고 빠르게 파악할 수 있는 자료를 제공하는 웹 서비스

핵심 서비스





논문 흐름



연구 키워드 시각화



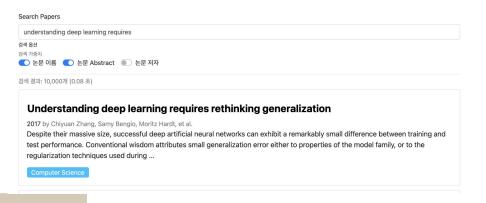
연구 통계



뉴스레터

핵심 서비스 1 - 기본 정보 열람

- **논문 검색 기능** 제공
- 검색한 논문에 대한 **기본 정보 제공**



Inderstanding deep learning requires rethinking generalization

2017

Publisher: ICLR

Published in: ArXiv • abs/1611.03530

Fublished III. Al AlV * abs/1011.03530

Authors: Chiyuan ZhangSamy BengioMoritz HardtBenjamin RechtOriol Vinyals

DOI:

bengio.abracadoudou.com			_		_		web.mit.edu 🗷
bengio.abracadoudou.com 🗵	arxiv.org 🗇	www.cs.ubc.ca		bracadoudou.com 🗷	crcv.ucf.edu 🗷		

— 2,456 Citations

22 31 References

Abstract:

Despite their massive size, successful deep artificial neural networks can exhibit a remarkably small difference between training and test performance. Conventional wisdom attributes small generalization error either to properties of the model family, or to the regularization techniques used during training. Through extensive systematic experiments, we show how these traditional approaches fail to explain why large neural networks generalize well in practice. Specifically, our experiments establish that state-of-the-art convolutional networks for image classification trained with stochastic gradient methods easily fit a random labeling of the training data. This phenomenon is qualitatively unaffected by explicit regularization, and occurs even if we replace the true images by completely unstructured random noise. We corroborate these experimental findings with a theoretical construction showing that simple depth two neural networks already have perfect finite sample expressivity as soon as the number of parameters exceeds the number of data points as it usually does in practice. We interpret our experimental findings by comparison with traditional models.

Paper Topics

Computer Science

〈논문 검색 기능 〉 〈논문 기본 정보 〉

핵심 서비스 2 - 논문 흐름

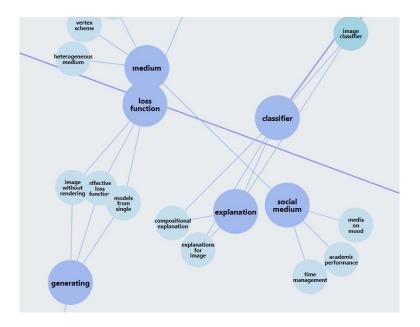
- 각 논문 별로 관련 핵심 과거 논문 및 최신 논문을 시간 순으로 제공
- 연구의 진행 흐름 파악을 도움
- 레퍼런스 기반 더자세한 원리는 Appendix 참고

Paper Flow Programming Technique: An improved hash code for 1968 scatter storage 1968 by Ward Douglas Maurer **Computer Science** Although scatter storage tables are used widely in system programming, they are subject to various drawbacks. One of these is that the size of the table cannot be arbitrary, but is restricted to powers of 2 by the hash coding method. In this note we present a new hash coding method that, besides bei ... Space/time trade-offs in hash coding with allowable 1970 1970 by Burton H. Bloom Computer Science In this paper trade-offs among certain computational factors in hash coding are analyzed. The paradigm problem considered is that of testing a series of messages one-by-one for membership in a given set of messages. Two new hash-coding methods are examined and compared with a particular conventional ... Compressed bloom filters 2001 2001 by Michael Mitzenmacher Computer Science A Bloom filter is a simple space-efficient randomized data structure for representing a set in order to support membership queries. Although Bloom filters allow false positives, for many applications the space savings outweigh this draw-back when the probability of an error is sufficiently low. We i ... **Theory and Network Applications of Dynamic Bloom** 2006 2006 by Deke Guo, Jie Wu, Honghui Chen, et al. Computer Science

A bloom filter is a simple, space-efficient, randomized data structure

핵심 서비스 3 - 연구 키워드 시각화

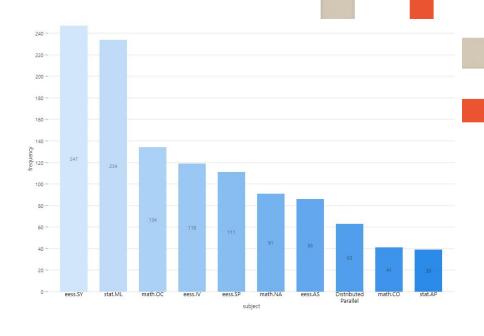
- 논문에서 등장한 **키워드**들의 **관련성**을 네트워크 형태로 **시각화** (월 단위)
- **생소한 분야**에서 **새로운 연구 분야를 탐색**하고 **발견**할 수 있도록 도움



핵심 서비스 4 - 연구 통계

- 월 단위로 투고된 논문들의 정보에 대해 **통계** 및 **시각화**
- 최근 **연구 동향 파악**에 도움

핵심 서비스 5 - 뉴스레터



• 통계 및 이슈에 대해 소개하는 뉴스레터를 이메일을 통해 구독 및 발송

기대 효과

비전문 분야라도 관심 논문에 대해 과거부터 현재까지의 연구 흐름을 쉽게 파악

시각적 도구로 특정 분야, 특정 기간에 진행되었던 연구의 동향을 쉽게 파악

비전문 분야라도 키워드를 이용하여 최신 관련 분야를 쉽게 발견 및 탐색



데이터셋

- 1. Semantic Scholar
 - 14,116,789 개
 - 전체 CS 분야 논문



- 기본 논문 정보
- 논문 흐름

- 2. ArXiv
 - 214,623 개
 - 2017 ~ 2021.04 CS 분야 논문

월별 데이터



- 연구 키워드 시각화
- 연구 통계
- 뉴스레터

개발 기술

Front End





웹개발





UI/시각화

배포 더 자세한 구조는 Appendix 참고







Amazon Elasticsearch Service

Docker 및 AWS를 이용한 배포

Back End





API 서버





데이터 분석





DB



김태영

백엔드 DB 구축 데이터 분석



윤상건

풀스택 AWS 배포 데이터 분석

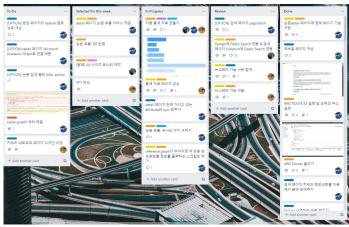


이민종

풀스택 데이터 시각화 데이터 분석

협업 툴











멘토링

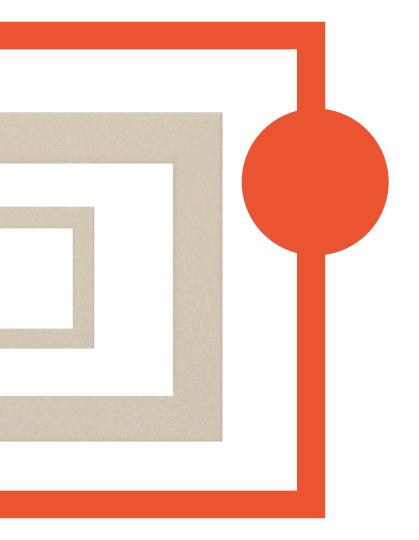
- 조헌기 멘토님
- 40회 이상의 메일을 주고받으며, 여러 도움을 받을 수 있었음
 - 프로젝트 스코프에 대한 의견 등을 질문 드리고 도움을 받음.
 - 논문에 친숙하신 멘토님께서 소비자 입장에서의 경험 공유를 통해 서비스에 대한 피드백 반영
 - 참고할만한 타 서비스 추천 (Microsoft Academic)
 - 기술적인 조언
 - Ex) 도메인에 맞는 워드 임베딩 학습을 위한 데이터셋 추천

Hunkee Cho - printing-grant.com B-0710-EE-9107-WLU, 201210-EE-9107-WL Huny 971 -								
인생하네요 음당건 님.								
혹시 중한 발표 자료 어느 정도 준비가 되었으면 이리 공유 제주실 수 당사합니다.	있나요? 현료형 세선 가지가 전에 제가 더리 한번 노현 중대 효율적으	로 대칭 전병 할 수 있을 것 같습니다.						
조전기 도립 Best Regards, Hardan Cho								
manage Coo								
출살권[소프트웨이학부] rgg[00251@kookmin.ac.krr Hunkes, 나, 김학경(소프트웨이학부], Haley에게 ~								
네, 어제 발표자로 준비가 어느렇도 마무리 되었습니다. 예상 구글 프	원전테이션 문서를 이메일로 군유해 드렸습니다.		19/1 Haley (12) +					
또한 이번 미팅에서 저희가 준비한 질문은 다음과 같습니다.			도 다른 특별한 방법은 없는 첫 같습니다. 대신 일만 자카드나 모시한	맛식을 이용했던 유사도를 꾸하시고 이 유사도를				
 카워도 노도 그래프를 만들기 위해 작 논문의 abstract에서 가 문 것이라고 생각할 생각입니다. 이때, 자금은 단순히 된 4 단역 각 카워드와 관계가 나타난 그래프가 있다고 만들 때, 도 2는 문의 40억만(1042년) 이용하여 여구가에를 보신했다. 		용 가서 할 것 같습니다. 장희를 취임 여러와 같습니다. 기가 전 프 사건 발식 중 높은 유사도를 역해서 전혀 데이지성이 대해서 유사도 계신 ameliation 고 에 보면에 서로 집에 되었다게 나온다는 것을 모여들으로 유사도 계산에 대한 집중 102						
BABUG								
		COURSE CES						

		용상진(소프트웨이락부) -pg/05281gs						
Hanker Cho - gostontyggman com- 8 (2) G. R. R. R. R. R. R. L. L. G. R.			(0.2) +					
인명하세요 준당한 병.			52 9900					
계속 호텔에 합니다 한다시 어떤 일을 대표합니다. 한다시 회사들이는 이 대로 및 함께하다다. 현대가 인한 크 전체에 보내는 이 대한 의 교육을 받는 병료에는 보는 전체에 모든 경계에서는 이 등에 가는 하는 이 한 등에 나는 이 에 대한 시간 이 가는 이 에 대급하게 구축한 하는데, 다. 그리고 하나는 군반에서 기업의 부명이고 있는데 확인하기, 하면 windows는 한 한경에 한답기 이에 대한 계속이 있으면 가지나 1억에 한 한 등 없는 사람이나, 한 제가의 구축한 생활하게 하는 하게 되어 모든 분명이 가지 있을 것 같습니다.								
조단기 유럽			Dho siptinfinite@gmail.com 사용이 작성.					
East Regards.								
Huston Cho								
			I트립어탁위, 나에게 +					
준상건[스프트웨이학생] - 10000201(8000000000000000000000000000000			9.					
화근까지 등문 데이터 수집에 많은 시간이 들어서 답답을 이제다 어때까지 본무 소집 관련 지원 안물을 공용 요리면	EBUR.							
 등을 데미터의 결후, 비간장의 용제로 우선 그것이 수한 entro에서 소립한 데미터의 문제같은 reference 집에, cit 그리다 ceresants establish 에서 제공하는 API Datas Ref. 을 수 있었습니다. 요서 그래요 소개한 anable 등 serents' substantial. 	법원 and 아이 이를 사용하기도 법으니다. (CS필의 20%-2015로 2017-2 sides 전로, 중인 전체증한)ears 등의 작은 전체를 중가하는 전체가 차 semanticulation copie, 이동하는 돈 전도 제 2호를 통하는 전체가 차 수입한 아이네고 대학교 수 있는 요즘이 보통 전체를 통한되고, 매력 약을 중 138,444대로 하려면 요니다. 교육는 전문 순구 교육을 중심하는 작업을 통한된다. 학교 약을 교육는 전문 순구 교육을 중심하는 작업을 통한 환경 약동합 전체 교육는 전문 순구 교육을 중심하는 작업을 통한 항상 하는 다음을 받았다.	 이 있고신가요? 우시도가 높은 논문들을 보았음데 설계로도 관한당이 높다고 불 로도 집중이 된 Entraset 이 있고한 물을 것 같습니다. 등미하면 서로 연극당이 있 나나다	5월 7월 [월] 의한 927 ☆ ♣ [1수 있는 수지화된 데이티를 가지고 계신가요? I다 없다를 할 수 있는 liabel 이 있어 있는 데이					
여기서 질문은, 연료님께서 비대주인 validation문제와 연란이 집: 문에 validation를 이렇게 할 수 있을지 그만이 됩니다.	로 작는 단관 논문 그리프를 계산에 내는 작업을 수정할 예정입니다. 문데, 논문간 유사료를 계산해낸 결과된 valid이부를 어떻게 수지와 하는가	 습니다. 지난번과 비슷한 시간대에서 활용분들지의 가능하신 날짜 알려주시면 이	IS 장도록 하건입니다.					
해당 부분에서 참고할 수 있는 충분 방법이나 시의기 있을까요?	Hustine Cho							
	유상제소트트웨이턴틱 -opticinateoimicas in-			\$10.500 (10.520 12.27 · dy · dy · 1				
•	Harden, U. CORDIS IN BROTHE Harry O'T -			1818(8) X5 (12) W + :				
	전에 진행되면 작업은 논문의 reference를 가슬러 돌다가면서 obstend 일단 해당 작업을 진행한 위에 해당 논문 을 면에서 주루에 abstract를 d	i가 높은 논문을 주물해 내는 작업을 진행된다. loc2rec 문을 이용하여 벡디의 한 티, 주사도를	(다. 국정자에 관련성이 적은의 reference된 논문을 열다면 잘 생각을 가지고 있습니다 -	λ.				
	유사노의 선언에서 선생은 GERNAGE 가시고 있지 않았다나 다신, 개선에 면표님이 보내주신 scientific papers 데이터넷을 이용해서 현재 제표가 하나의 논문에 대해 가지고 있는 값들은 다음과 같습니다.	i docZvec를 학습시킨 위, 제당 docZvec에서 U	돈 벡터를 간의 문사도를 코사인 문사도 등으로 축정할 영작을 가지고 있습니다.					
	· 68 H							
	 는문제목 electrical pd 다스용 							
	. pd v.a.g . N.S. . R.S. b.Fille							
	· 레이언스 논문됨 · 논문문의							
	· 출한 영도 • 출한사							
	• 중단 자설 명 • 저설 물론							
	Nig Sjöl 7) dol Microsoft Academic Graph (E-15) id							
	Microsoft Academic Graph® 412 id THE UNE SEED DIS ADDRESS THEBUTE.			_				
	- \$49.953430E08 - \$49.953400E							
	* and * M.E. 301 Std.							

추가 개발 요소

- 분야 확장
 - o Computer Science 외 분야까지 포함
- 추가 기능
 - 로그인 기능
 - 로그인을 통한 개인 맞춤 기능
 - 논문 요약 서비스 제공
- 통계 고도화
- 논문 흐름 서비스 고도화
 - 사용자 경험에 따라 서비스 성능 개선

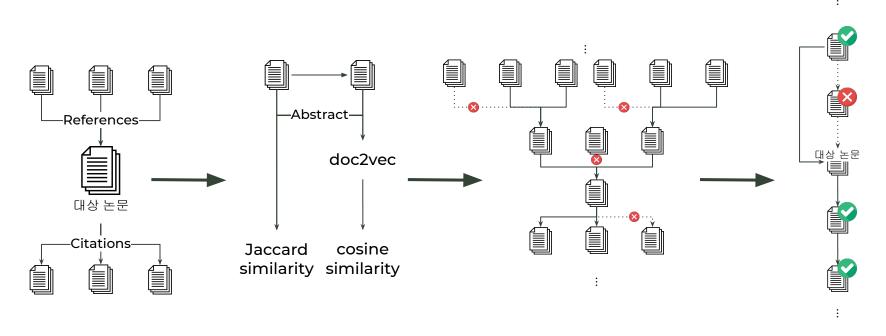


THANKS

20160729김태영20191632윤상건20181656이민종

04 Appendix

논문흐름 그래프 생성 알고리즘

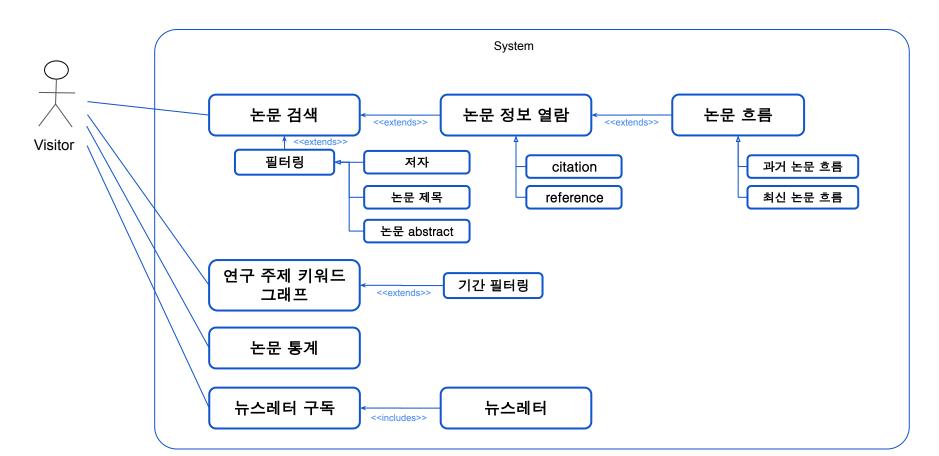


참조를 통한 Directed Acyclic Graph 생성 간선 간 유사도 계산

논문간 유사도 기반 Pruning

위상정렬 및 Citation 수 기반 상위 논문 필터링

Use Case Diagram



배포 시스템 구조

