뉴랄 브리핑
Neural Briefing
제 0002호

by 워니스홍

2019년 5월 31일 금요일

1 토막상식

Transformer: Vaswani et al. [2017][Paper-link] [Blog-link]
Self-attention 을 사용해서 Language understanding 문제를 푸는 친구임. 번역기를 떠올리면 됨 (e.g., 한국어→영어). 입력으로 들어가는 모든 단어가 서로서로 모든 단어를 참조함. 참조하는 가중치가 다름 (i.e., 어느 단어이냐에 따라서 주의 (attention)을 더 기울이고 덜 기울이고 함)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

, where a set of quaries Q, keys K, values V, and dimension d_k. Q, K, V are matrices.

"An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key."Vaswani et al. [2017].

2 대형 회사들 Blog

Google AI Blog

No issues

Microsoft Research

Fashion forward: Researchers, designers debut new tech on New York City runway

엔지니어랑 디자이너랑 테크가 결합된 패션쇼를 했다는 조금 시덥잖은 내용.

Facebook Research Blog

Mark Harman on receiving the SIGSOFT Outstanding Research Award at ICSE 2019

Mark Harman라는 아저씨는 SBSE (Search-based software engineering)이라는 분야를 만든 선구자급 연구원임. 그래서 좋은 상을 두개를 받았다는 조금 시덥잖은 서론으로 글이 시작됨. 그래서 그 분야를 개척하게 된 스토리 를 들려줌. 이분은 진화 알고리즘 (evolutionary algorithms)에 대한 관심이 매우 많은 분임. 90년대에 소프트웨어 엔지니어링 분야에서 탐색 (search) 알고리즘에 대한 연구가 매우 시덥잖았다는 것에 충격을 잡숴음. 탐색 알고리즘은 매우 고리타분하게 오래된 것인데, 그래서 이분야 저분야에 사용이 많이 되었으나 유독 소프트웨어 엔지니어링에는 잘 사용되지 않았음. 그래서 이상타 해갖고 연구를 막 했음. 그래갖고 SBSE라는 분야를 창시하게 되어버림.

AWS Machine Learning Blog

Automatically extract text and structured data from documents with Amazon Textract

세상에는 문서 (document)라는 것이 매우 천지에 널렸음. 부동산 문서 같은거 말하는 것임. 근데 양은 무진장 많은데 형식이 다 제각각이라서 정보가 있어도 제대로 뽑아 활용을 못하는 단점이 있었음. 그래서 Amazon Textract라는, 문서로부터 텍스트 및 정형화된 정보(=테이블 형태)를 뽑아내는 툴을 만들어 보았음. 이 포스팅에서는 아마존의 Textract라는 것을 어찌 사용하는가를 설명해줌. 클릭클릭 스텝바이스텝으로 하면 참 쉽게 따라할 수가 있음.

DeepMind

Unsupervised learning: the curious pupil (2019년 4월 10일 글)

이 글은 딥마인드의 연구가 어떤 의의가 있는가를 줄줄이 설명돕기 위한 시리즈물임. 사람은 정답지가 있는 상황 (supervised) 말고, 정답지 없는 상황 (unsupervised)

에서도 매우 잘 배움. 왜 근데 기계는 그걸 잘 못할까? 고심해봄.

Decoding the elements of vision: 본다는것이 무얼 의미하는가를 설명해봄. 우리의 눈과 뇌는 시지각을 어떻게 처리할까. 그거를 컴퓨터로 어떻게 시늉낼 수가 있을까.

Transfer learning: 여기서 배운 것을 전혀 다른 저따가 갖다가 척 하니 붙여서 사용할수 있으면 좋겠음. 예컨대 사람아이가 사람을 작대기 몇개로 그리는 법을 배우면 그 방식으로 친구도 그리고 안경도 그리고 별거별거를 다 그려낼 수가 있음. 사람자식은 그거를 하는데 왜 기계는 그걸 못할까? 를 고심해봄

Learning by creating: generative models. "뭔가를 만들 수 없으면 그걸 이해한게 아니다." 라는 리처드 파인만의 어록이 있음. 그래서 머신러닝 모델이 글이나 그림을 '만들도록' 시켜서 배우게 하는 방식이 있음. GAN (Generative Adversarial Network)가 대표적임.

Creating by predicting: 예측을 시켜서 배우는, 예컨대 OpenAI's GPT-2같은 pretrained language model은, A문장 다음에 B문장이 온다는 사실을 가지고 또 그다음 문장 그다음 문장을 예측하도록 시킴. 뭔가를 예측함으로 배우게한다는 논리임. 멀쩡한 글 중간중간에 일부러 빈 구멍을 뚫어서 머신러닝 모델보고 맞춰봐(=예측해봐) 시켜서 배우게 만듬.

결론 Re-imigining intelligence: 생성모델 (Generative models)는 멋진 것이지만 우리 딥마인드에서는 이놈을 첫걸음마 정도로밖에 취급하지 않음. 우리는 기계한테 상상력 (imagination)을 부여하고픔. 그래서 미래에 대한 계획 (planning)이랑 추론 (reasoning)을 하게 만들고픔. 연구를 좀 해보니깐 계속 달라지는 환경을 예측하는 식으로 배우게 하는게 모델이 가진 지식이나 문제해결능력을 매우 풍부하게 만들어준다는 것을 깨달았음. 사람도 그렇잖음. 그래서 우리의 목적은 인간의 머리통을 고대로 쏙 빼닮은 일반 인공지능을 만들어내는 것임. 거기서 젤루 중요한것은, 누가 명시적으로 알려주고 시키고 그러지 않아도, 답없이도 (unsupervised) 제 스스로 깨치도록 하는게 옳다고 믿음. 우리 사람도 가만 놔두면 자기 주변 환경에 계속 주의를 기울이면서 배우잖음. 누가 정답지를 쥐어줘서 배우는게 아니고.

OpenAI Blog

No Issues

3 Curated Papers

What do you learn from context? probing for sentence structure in contextualized word representations

Tenney et al. [2019] [Link]

이 논문에서는 edge probing tasks 라는 것을 제안함. 이게 뭐냐하면 기존의 대표적인 contextual pre-trained모델 네가지 (ELMo, BERT (base, large), OpenAI GPT, CoVe)가 encode하는 고유한 성질이 무엇인가를 탐색하는 것임. 일반적인 word embedding이 안하는 그 무엇인가를 이놈들(=contextual embeddings)은 잘할듯 해서 그 속을 들여다보고 싶은 것임.

결론1 First, in general, contextualized embeddings improve over their non-contextualized counterparts largely on syntactic tasks (e.g. constituent labeling) in comparison to semantic tasks (e.g. coreference), suggesting that these embeddings encode syntax more so than higher-level semantics.
요약 contextualized embeddings에서 contextualized된 부분은 non-contextualized된 부분보다 syntactic tasks가 반영되어 contextualized된 것임. semantics tasks는 contextualized되는데에 효과를 덜 일으켰음. 뭔말이냐면, contextualized embedding이라는 놈은 의미 (semantics)보다는 구문 (syntactic)을 가지고 학습을 한다는 것임.

결론2 Second, the performance of ELMo cannot be fully explained by a model with access to local context, suggesting that the contextualized representations do encode distant linguistic information, which can help disambiguate longer-range dependency relations and higher-level syntactic structures.
요약 ELMo의 성능에 대한 설명으로는 local context의 개념만 가지고는 부족하고, 서로 멀리 떨어진 context 정보까지 있어야 설명이 잘 됨. (ELMo가 전체 corpus를 모두 반영해서 계산하는 좋은 모델이라는 뜻임)

Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding.

Xiaodong Liu [2019] [Link]

이 논문의 주안점은 MT-DNN + Knowledge distillation 을 했다는 것임. 그래서 Distilled MT-DNN을 제안하였음. Vanilla MT-DNN이 baseline임

knowledge distillation effectively transfers the generalization ability of the teachers to the student. (c.f.) Hinton et al. [2015][Link]

아래 그림 Figure 1이 MT-DNN모델임. 아래쪽은 pre-training을 시켜서 일반적인 컨텍스트를 다 포함하도록 만들어놓고, 위쪽은 여러가지 서로 다른 tasks마다 다르게 적용할 수 있게 분리해놓았음. 그래서 contextual embedding이라고 부름.

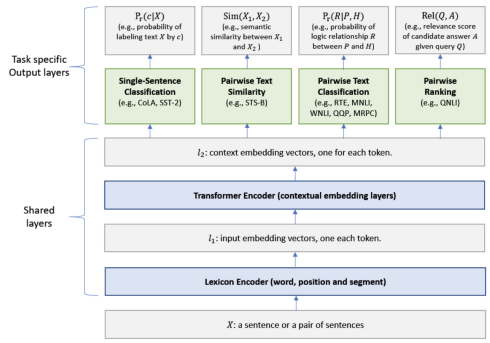

Figure 1: Architecture of the MT-DNN model for representation learning (Liu et al., 2019).

Figure 1: Distilled MT-DNN Model

논문에 나온 Terminologies

- MT-DNN := Multi-Task Deep Neural Network
- GLUE := General Language Understanding Evaluation benchmark dataset which consists of 9 NLU tasks.
- NLU := Natural Language Understanding
- Vanilla model := 아무 추가옵션을 달지 않은 기본 컨셉의 모델. vanilla RNN, vanilla CNN 등으로 부름. vanilla model을 baseline으로 놓고, 추가옵션이 달린(=제안하는) 모델이 더 좋다는것을 보여주는데 비교대상으로 사용함.
- Transformer := 변신시킨다는 뜻임. 뭣을 변신시키냐 하면, 입력(한국어)를 출력(영어로) 변신시키는 등의 번역기와 같은 일을 하는 모델을 transformer라고 부름. 번역이라 하는것은 내용물은 같고 형태만 달라지므로 그렇슴. 꼭 '번역'일만 하는것은 아니고 대표적인 task가 번역이라는 말임. 예컨대 '신체'라는 단어는 '몸'이라는 단어와 내용물이 같고 글자 생긴 모양이 다르므로 이것도 일종의 번역임. transformer라는 것은 그래서 입력과 출력의 context를 유지하면서 형태만 바꾸는 일에 사용함.
- BERT := Pre-training of deep bidirectional transformers for language understanding. 이것은 2018년 말에 나온 것인데 pre-training Language Model의 신천지를 열어주었음. 한번 큼직하게 training을 시켜놓으면 이 일에 갖다 붙이고 저 일에 갖다 붙이고 하면서 일반적으로 사용할 수 있는 모델임. Bidirectional이라는 것은 문장을 왼쪽-¿오른쪽 으로 학습하면서 오른쪽-¿왼쪽으로도 학습한다는 뜻임. 이게 가능하도록 문장의 부분부분을 Mask씌워서 스스로 맞춰보게 하였음. unsupervised learning임.
- Unsupervised learning의 의의: 자연어처리에서 unsupervised learning이 갖는 위상은 특히나 높은데 왜냐하면 세상에는 labeled 데이터보다 unlabeled데이터가 훨씬 많기 때문임. 그리고 unsupervised learning이 된다면 원어민이 작성한 온라인 그냥 아무 글이나 갖다가 데이터로 왕창 때려넣을 수가 있게 됨. supervised learning일 때는 이 문장이 어떤 문장이고 저 단어가 어떤 단어이고 하는 것을 사람이 일일이 labeling해줘야만 했으므로 데이터 양이 커질수가 없었음.
- Distilled MT-DNN에서 아래층 shared layer를 train시킬때는 BERT가 하는 식으로 함. 위층의 Multi-task learning (MTL)부분에서는 서로 다른 task를 위해서 서로 다른 embedded representation을 만들어냄.
- 그다음에 Knowledge distillation을 위해서 각각의 task마다 데이터를 따로 준비함. 아래 그림에서 task 하나를 train시킬때마다 위 그림의 모델 하나를 통째로 갖다가 사용함. 앙상블 (ensemble) 학습을 하는 중. Figure 2.

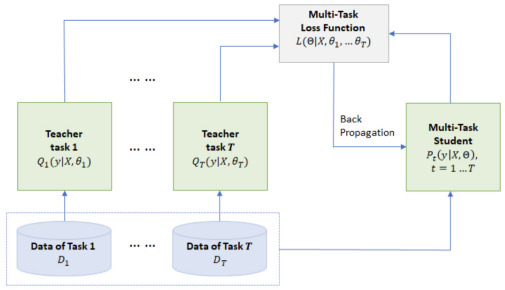

Figure 2: Distilled MT-DNN Mode, knowledge distillation

- 앙상블 (ensemble)이란, 서로 다른 모델 여러개를 갖다가 따로 훈련을 시켜갖고 자기들끼리 내놓는 결과를 평균내서 최종 결과로 사용하는 모델합체 방법임. 앙상블을 하면 테스트시에 2 3% 정도의 성능향상을 가져오고, 평균을 내므로 여러 모델들간에 혼자 성능이 툭 튀어나가는놈 없이 generalization이 잘 된다는 장점이 있음. 근데 앙상블 하나마다 모델 하나가 통으로 들어가므로 계산이 방대해진다는 단점이 있음.

$$Q = \text{avg}([Q^1, Q^2, \ldots, Q^K]).$$

- Hinton et al. [2015] [Link]에서 제안한 soft target과 hard target의 objective function 둘을 평균내서 사용함.
- **결론** 우리가만든 (=Distilled MT-DNN)모델이 제일루 좋음.
- **요약** 기존의 모델들 (BERT, GPT, ELMo, CoVe)은 성능이 좋은건 알겠는데 훈련 및 사용시 매우 육중하고 그래서 모바일 기기같은데 사용불가했음. 우리가 제안한 distilled MT-DNN을 쓰면 뉴랄넷 사이즈도 줄이고 성능도 약간이지만 올라가는것을 보였음. 그래서 잘났음.

4 Useful links

정보

- CORE Ranking portal: 학회 랭킹 보는 사이트
- Awesome-deep learning: 딥러닝 개념 리스트

작은 블로그들

- Distill
- AI Newsletter
- Andrej Karpathy Blog
- Colah's Blog
- WildML
- FastML
- The Morning Paper

뉴스

- 파비
- VentureBeat

머신러닝 논문+커뮤니티

- 레딧
- arXiv
- OpenReview.net

5 대학원생을 위한 도움말 (Advice for graduate students)

- Gradute School Survival Guide
- Ph.D. Students Must Break Away From Undergraduate Mentality
- 이지형 교수님 조언
- 김형식 교수님 조언
- FAQ for My Research Students (번역)
- 대학원생 때 알았더라면 좋았을 것들
- 대학원생을 위한 지극히 개인적인 10가지 조언
- 석사와 박사
- 내가 대학원에서 생존한 방법
- 김형식 교수님 모음집
- 저의 지도교수님이 학생에게 주는 일반 조언
- 논문의 주안점
- 대학원생 때 알았더라면 좋았을 것들
- 행복한 대학원생 되기
- [진로] 대학원 중도포기... 조언부탁합니다.
- 대학원 조언 부탁드립니다
- 대학원 진학이.. 뭔가 두렵습니다 길지만 조언 부탁드리겠습니다..
- 선배님들께 박사과정 진학 및 유학에 관한 조언을 구합니다.
- 박사 학위를 실패하는 10가지 쉬운 방법
- How to Read a Paper
- What Constitutes a Theoretical Contribution?
- Ph.D. Students Must Break Away From Undergraduate Mentality
- 10 easy ways to fail a Ph.D.
- 탐색적 논문 읽기
- QnDReview: Read 100 CHI Papers in 7 Hours
- Essential elements for high-impact scientific writing
- This Is Exactly How You Should Train Yourself To Be Smarter [Infographic]

6 딥러닝 유명한 글모음

- The Bitter Lesson, Rich Sutton, March 13, 2019
- A Recipe for Training Neural Networks, Andrej Karpathy, Apr 25, 2019. [번역]

References

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Weizhu Chen Jianfeng Gao Xiaodong Liu, Pengcheng He. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv preprint arXiv:1904.09482v1*, 2019.