BERT는 구글이 만든 딥러닝 모델입니다. 먼저 위키피디아 같은 텍스트 코퍼스로 사전훈련을 합니다. 그다음 이 모델을 기반으로 새로운 데이터에 맞게 다시 훈련합니다. 이렇게 전이학습을 하면 훨씬 좋은 성능을 낼 수 있습니다.

Hugging Face가 공개한 패키지를 사용하면 BERT를 아주 쉽게 적용할 수 있습니다. 네이버 영화리뷰 데이터로 감정분석을 하는 코드를 구현해보았습니다. Colab으로 되어있어 바로 실행해보실 수 있습니다. 관심있는 분들에게 많은 도움이 되었으면 합니다.

< 네이버 영화리뷰 감정분석 with Hugging Face BERT >

-> https://colab.research.google.com/drive/1tIf0Ugdqg4qT7gcxia3tL7und64Rv1dP

솔트룩스 이경일 대표님의 설명으로 듣는 GPT-3 이야기입니다.

이미 공유가 됐는지 모르겠는데,

재미있으면서도 깊이 있게 잘 설명해주셔서 공유합니다.

제목은 '한방에 이해하기'지만 약 2시간짜리입니다. ㅎㅎㅎ

<https://m.facebook.com/story.php?story_fbid=120095536507305&id=104193434764182>

자연어처리 논문 읽기 모임 :

<https://www.facebook.com/groups/TensorFlowKR/permalink/1374921782848876/>

전이 학습 기반 NLP

<https://brunch.co.kr/@learning/12>

Bert

<https://velog.io/@gwkoo/NLP-transfer-learning-history>

안녕하세요. Tensorflow KR. 자연어처리 리뷰모임 [집현전]의 운영자 고현웅이라고 합니다. ***이번 12월 한 달 동안 집현전에서 만든 자연어처리 논문 리뷰 비디오 8편 (초급반 4편, 중급반 4편)* 을 공유**하기 위해 글을 작성하였습니다. 관심 있으신 분들은 아래 설명과 영상을 참고하시면 좋을 것 같습니다!

**[집현전 Links]**

집현전 Github : [https://github.com/jiphyeonjeon/nlp-review](https://github.com/jiphyeonjeon/nlp-review?fbclid=IwAR3r0evRC0THtIo_XqrpF3yTL7WfgN8pgzzMhkeX7vn-dPqBlc-gseksmi4)  
집현전 초급반 Youtube : [https://www.youtube.com/playlist...](https://www.youtube.com/playlist?list=PLsXisDblbLJ_UaUPi0iThUQKgSbhIk5ho&fbclid=IwAR1yCbGNZNeqGnKNVqJSo5dQZGhzJ24Je0Ce2jLiPIKnafZy1XZj5HE5qag)  
집현전 중급반 Youtube : [https://www.youtube.com/playlist...](https://www.youtube.com/playlist?list=PLsXisDblbLJ8msozFyA8o2zf3zfbORHWd&fbclid=IwAR0ZKEd-RPwJyADYIeM6OQ5vtvi6HTMiERfFCoLVj7Ff4aBqE99wfHPAJ-s)

**[초급반 12월 발표]**

**1. 송석리 - Long Short Term Memory**  
[https://youtu.be/HHKSCkPEQfw](https://l.facebook.com/l.php?u=https%3A%2F%2Fyoutu.be%2FHHKSCkPEQfw%3Ffbclid%3DIwAR0itEs3MtDOGqXz90JTREiSz6378QOfnMy4iLXGhWrxgWExXXuFa1TFoY0&h=AT31llOde2ZeSuFsHv7_dQ-hq5UK4Cx0wWmV6VwN12-mw-wdrhFF2DrFnl6ufKLZBah7doMsESXuZvv5L1p0PKHtJOfwqfQMI20YXr7z21Cympv896HFTbvtch9DpR8_c73d&__tn__=-UK-R&c%5b0%5d=AT3mSSRLkF_-Zil10dRK7lBO7urX5cXtnQ6r8Js7c3R-oMh6MN2Ruqc2gFCgKRyAYOyyCLgL5-o-CZzFo5AwTyrDmLRm0nGL6EBjzZSzIFzL8WPEKUgOd6FCsOuONsXZttEs3lnooDTkfc1V36j8d-GZNdkcmePVbGg)  
  
자연어처리에 대표적으로 사용되는 시퀀스 처리 모델인 LSTM에 대해 리뷰합니다. LSTM을 알아보기 이전에 RNN에 대해 알아보고, LSTM의 각종 게이트의 역할, 그리고 실험 결과를 깔끔한 일러스트와 함께 설명합니다.

**2. 이영빈 - Word2Vec**  
[https://youtu.be/o-2kfiZP7Z8](https://l.facebook.com/l.php?u=https%3A%2F%2Fyoutu.be%2Fo-2kfiZP7Z8%3Ffbclid%3DIwAR0vzSmx3c5PvZcNnoJPN2z6mdcVuFcKvyzHtpjvyLrkwEUalFlP5DDz9bA&h=AT13exz5XvzyhdTw5M3Df5vH3LcjiXMqERq1l9f7WierzmDJ-yglqxWB6jOuXGkniBFQaCsGkXlsZzAdRZea9VEiDiDRqni4cimPp2_L3aM3Rl_EsLkfYZtZ-MtrD8DeCw1_&__tn__=-UK-R&c%5b0%5d=AT3mSSRLkF_-Zil10dRK7lBO7urX5cXtnQ6r8Js7c3R-oMh6MN2Ruqc2gFCgKRyAYOyyCLgL5-o-CZzFo5AwTyrDmLRm0nGL6EBjzZSzIFzL8WPEKUgOd6FCsOuONsXZttEs3lnooDTkfc1V36j8d-GZNdkcmePVbGg)  
  
워드 임베딩 모델인 Word2Vec에 대해 논의합니다. 특히 Skip-gram (주변 단어 예측) 과 CBOW (중심단어 예측) 등의 방식에 특성에 대해 논의하고 Word Analogy 등의 태스크에서 성능을 비교하였습니다.

**3. 이기창 - FastText**  
[https://youtu.be/7pDB9zqwxhs](https://l.facebook.com/l.php?u=https%3A%2F%2Fyoutu.be%2F7pDB9zqwxhs%3Ffbclid%3DIwAR3tfeQlbDTr3KO8ckPSCWSVtv2UOiUOKUu-spXzE2S8Ec6uCnKwehNlyfQ&h=AT3Gkf5uwoLQVNAO1yL7qk92zxyiIm6DQ0S_v0cnlGmlI62m-cHnfatmyFZLmE36o81IxIfonIXHK-WZGQDYLiojLwYG0mBFAJ2UCoLASQ9r9csaXixVcJES5ykGs3cPBvYC&__tn__=-UK-R&c%5b0%5d=AT3mSSRLkF_-Zil10dRK7lBO7urX5cXtnQ6r8Js7c3R-oMh6MN2Ruqc2gFCgKRyAYOyyCLgL5-o-CZzFo5AwTyrDmLRm0nGL6EBjzZSzIFzL8WPEKUgOd6FCsOuONsXZttEs3lnooDTkfc1V36j8d-GZNdkcmePVbGg)  
  
Word2Vec을 개선한 FastText 임베딩에 대해 논의합니다. OOV 문제를 완화하기 위해 도입된 서브워드 임베딩의 개념에 대해 이야기하고 실험 성능을 비교합니다. 특히 Semantic 영역의 성능과 Syntactic 영역의 성능을 비교하여 평가합니다.

**4. 허치영 - Neural Probabilistic Language Model (NPLM)**  
[https://youtu.be/EWMNCTpfsLI](https://l.facebook.com/l.php?u=https%3A%2F%2Fyoutu.be%2FEWMNCTpfsLI%3Ffbclid%3DIwAR0itEs3MtDOGqXz90JTREiSz6378QOfnMy4iLXGhWrxgWExXXuFa1TFoY0&h=AT2mcJ9hRDyqlYptFXgzbplxzXyEgS0zY2H8kpgIcQm4Rrqu8kbSnvvsesBCepZ4zRzO2GB6sacOBF2iG7LHE7zsS2RmoLRuMC8fkfJ_ouI37wVP5oDRMdrv8-PTnfmdUfwH&__tn__=-UK-R&c%5b0%5d=AT3mSSRLkF_-Zil10dRK7lBO7urX5cXtnQ6r8Js7c3R-oMh6MN2Ruqc2gFCgKRyAYOyyCLgL5-o-CZzFo5AwTyrDmLRm0nGL6EBjzZSzIFzL8WPEKUgOd6FCsOuONsXZttEs3lnooDTkfc1V36j8d-GZNdkcmePVbGg)  
  
뉴럴 기반의 언어모델의 시초 격인 NPLM에 대해 리뷰합니다. NPLM의 학습 오브젝티브를 자세한 수식으로 알아보고 NPLM이 Long term dependency에 효과적인 이유에 대해 논의하고 실험 성능을 비교합니다.

**[중급반 12월 발표]**

**1. 진명훈 - If Beam Search is the Answer, What was the Question?**  
[https://youtu.be/KJClfF\_nJj0](https://l.facebook.com/l.php?u=https%3A%2F%2Fyoutu.be%2FKJClfF_nJj0%3Ffbclid%3DIwAR2kEpGgpv9D15ERwa5kCZypuqbu9rrnN2fCSbWOPmHnfCS0tGmwE1nEonA&h=AT1OhTDjy0dwvW8gzZcDky6KfMKwAGBFVoKGv09Ie1BGxXgl5mHKQJnbQwd28pDGP3v_TUrxnSt4SwvzPFfxx4jHBG_gqqER_gw5tidK0ENvg-JUG3B0K-6TE4O6Knul7OMo&__tn__=-UK-R&c%5b0%5d=AT3mSSRLkF_-Zil10dRK7lBO7urX5cXtnQ6r8Js7c3R-oMh6MN2Ruqc2gFCgKRyAYOyyCLgL5-o-CZzFo5AwTyrDmLRm0nGL6EBjzZSzIFzL8WPEKUgOd6FCsOuONsXZttEs3lnooDTkfc1V36j8d-GZNdkcmePVbGg)  
  
디코딩 전략 중 한가지인 빔서치에 대해 논의합니다. 빔 사이즈를 최대로 높히는 것과 동일한 MAP 생성이 실제로는 잘 안되는 문제를 지적하며 빔서치가 좋은 문장을 생성해내는 이유를 인지심리학적인 요인을 통해 분석합니다. 추가로 UID 디코딩이라는 새로운 기법에 대해서도 논의합니다.

**2. 박동주 - SSMBA : Self Supervised Manifold Based Data Augmentation for Improving Out of Domain Robustness**  
[https://youtu.be/1IwHZ\_4uPK0](https://l.facebook.com/l.php?u=https%3A%2F%2Fyoutu.be%2F1IwHZ_4uPK0%3Ffbclid%3DIwAR1yCbGNZNeqGnKNVqJSo5dQZGhzJ24Je0Ce2jLiPIKnafZy1XZj5HE5qag&h=AT2EWcDjEsSUJJTAajQnMSJ3F3JsvyOYqcg0GmO0w33aANKUPVoSHnY44_sZaLfP7Ty_X_QN_xQqQ-wIYU2VZQoZMIfV1oVav_HyGTS5iGfK_tlIAO8dmFZ5By0so9dqQRx6&__tn__=-UK-R&c%5b0%5d=AT3mSSRLkF_-Zil10dRK7lBO7urX5cXtnQ6r8Js7c3R-oMh6MN2Ruqc2gFCgKRyAYOyyCLgL5-o-CZzFo5AwTyrDmLRm0nGL6EBjzZSzIFzL8WPEKUgOd6FCsOuONsXZttEs3lnooDTkfc1V36j8d-GZNdkcmePVbGg)  
  
Vicinal Risk Minimization의 일환으로 BERT의 MLM을 활용한 Data Augmentation 기법에 대해 논의합니다. Machine Translation, NLI, Sentiment Analysis 등의 Downstram Task에서 실험을 진행하였으며, 특히 In domain과 Out of domain 세팅에서의 성능을 비교합니다.

**3. 고현웅 - Machine Translation Survey**  
1편 : [https://youtu.be/KQfvEg-fGMw](https://l.facebook.com/l.php?u=https%3A%2F%2Fyoutu.be%2FKQfvEg-fGMw%3Ffbclid%3DIwAR1yCbGNZNeqGnKNVqJSo5dQZGhzJ24Je0Ce2jLiPIKnafZy1XZj5HE5qag&h=AT15liHvs5F9rs2VjQxv4s8h82RAaUGIWy4R4b4oramSkkpBag5FdI1Mzus6MqTxp2fg_r12q1HCTpheYoNozSWSOzTCZu2ZPFhjPAym3ZeuPEnIu82lpFtMga2AdcjKM31a&__tn__=-UK-R&c%5b0%5d=AT3mSSRLkF_-Zil10dRK7lBO7urX5cXtnQ6r8Js7c3R-oMh6MN2Ruqc2gFCgKRyAYOyyCLgL5-o-CZzFo5AwTyrDmLRm0nGL6EBjzZSzIFzL8WPEKUgOd6FCsOuONsXZttEs3lnooDTkfc1V36j8d-GZNdkcmePVbGg)  
2편 : [https://youtu.be/18iH6VX-IU4](https://l.facebook.com/l.php?u=https%3A%2F%2Fyoutu.be%2F18iH6VX-IU4%3Ffbclid%3DIwAR2R2fhB1AVDwduyg_AvOY7dohXg6_qpwShcTOnIZO_NRjLOZvBvshAnoGA&h=AT0Zayu10L7wwQCJGG8dGr3hRWEPhr9zHSlESpyIXrTQtZ5j_jpiCVPUfM5rKKV06QW3ssZ26nFm2eJGOczH8um70gDeYyJd58L76GCtf6o4TZL6YxCSLI80_FFL-VXFneCc&__tn__=-UK-R&c%5b0%5d=AT3mSSRLkF_-Zil10dRK7lBO7urX5cXtnQ6r8Js7c3R-oMh6MN2Ruqc2gFCgKRyAYOyyCLgL5-o-CZzFo5AwTyrDmLRm0nGL6EBjzZSzIFzL8WPEKUgOd6FCsOuONsXZttEs3lnooDTkfc1V36j8d-GZNdkcmePVbGg)  
  
총 4시간에 걸쳐서 Seq2Seq부터 최신 연구(mBART, M2M-100, M2 등...)까지 23편의 논문을 참고하여 신경망 기계 번역의 역사와 최신 연구 동향에 대해 논의합니다. Background, Dataset, Efficiency, Training, Architecture 등에 대해 논의를 진행하고 성능, 서비스, 앞으로의 방향성에 대해 디스커션을 진행합니다.

자연어 전처리 모음집

<https://colab.research.google.com/drive/1FfhWsP9izQcuVl06P30r5cCxELA1ciVE?usp=sharing>