나왔던 질문 /
궁금할 수 있는 질문 정리

1. 데이터 엔지니어링과 데이터 사이언티스트의 차이
2. 딥러닝/머신러닝 분야에서 일하기 위해 석/박사를 반드시 취득해야하는지
3. 딥러닝/머신러닝을 할때 추가적인 프로그래밍 언어를 알아야 하는지

기타 도움 되는 자료 정리와 및 개인적 생각 정리, 학습 방법

1. 데이터 엔지니어링과 데이터 사이언티스트의 차이

1. 데이터 엔지니어링과 데이터 사이언티스트의 차이

회사마다 데이터 엔지니어링과 데이터 사이언티스트를 정의하는 방법은 다를 수 있습니다.

회사가 다루고 있는 도메인/비즈니스 모델마다 데이터를 처리하는 목적과 방법이 다를 수 있습니다.

다음 내용을 참조하시면 좋을 것 같습니다.

https://tech.kakao.com/2020/11/30/kakao-data-engineering/

2. 딥러닝/머신러닝 분야에서 일하기 위해 석/박사를 반드시 취득해야 하는지?

엔지니어링의 경우는 반드시 필요하지는 않습니다. 특이 케이스지만 유니스트를 졸업한 김태훈 씨는 학부 졸업 후 오픈 AI에 입사하기도 하였고

(오픈 소스 활동을 많이 하셨고 스카웃을 받아 입사하셨습니다. https://www.joongang.co.kr/article/22918688#home)

제 주변에도 학부 졸업 후 딥러닝 분야 네이버 체험형 인턴을 거쳐 삼성 리서치에 입사한 케이스가 있지만

냉정하게 어느 정도 괜찮은 회사를 목표로 데이터 사이언스/ 분석가로 지원하길 원하신다면 연구 실적이 있어야 하기 때문에

석사 이상은 필요할 가능성이 높고 그렇지 않다면 그 수준 이상의 개인적인 노력이 필요합니다.

따라서 프로젝트를 진행하기 이전에 자신이 백엔드 분야에 취업을 원하는지, 데이터 엔지니어링 분야에 취업을 원하시는지 아니면

데이터사이언스/ 분석가로 지원을 원하시는 건지 진로 설정을 명확히 할 필요가 있습니다.

현재는 너무 광범위하게 학습 중이기 때문에 과정을 진행하면서도 자신이 가고자 하는 분야에

대해 **~~정말 올바른 방법으로 철저하게 준비~~**하는 것을 권고 드리고 싶습니다.

(정말 좋은 회사에 가고자 한다면 엔지니어링도 석/박사 출신 들과 경쟁할 수 있고 어떤 기업이던지 기초 분야에 대한 탄탄한 지식과 관련 실무능력이 필요합니다.
물론 석박사 출신분들도 자신의 전공에만 지원하시지는 않습니다. 예를 들어 제 전직장 동료분의 경우 강화학습을 전공하지고 입사하셔서 이직하실때는 컬리의
데이터 엔지니어링 분야로 이직한 적이 있습니다.)

다음 영상에서도 해당 내용에 대해 언급하고 있는데요, 시간이 되신다면 처음부터 다 들어보시면 좋을 것 같습니다.

https://youtu.be/5xXEC4PEX7E?t=345 (딥러닝 분야에서 석사 이상의 학력이 반드시 필요한가요?)

https://www.youtube.com/watch?v=NpuZk1lhG1o

3. 딥러닝/머신러닝을 할때 추가적인 프로그래밍 언어를 알아야 하는지

데이터의 경우 파이썬은 거의 보통 활용되고 회사마다 추가적인 언어가 필요할 수 있습니다.

엔지니어링의 경우 다양한 데이터를 처리하는데 관련 어플리케이션과 연계된 언어를 사용할 수 있습니다. 보통은 python을 메인으로, java, scala 등과 c#등을 활용하기도 합니다.

그러나 하나의 언어를 깊이 있게 알아 놓으면 다른 언어를 익히는 것은 보통은 쉽기 때문에 처음부터 여러 언어를 공부하는 게 아니라 추후 빠른 학습에 도움이 될 수 있는 탄탄한 기본기에 집중하시는 것이 가장 중요합니다.

나중에 현업으로 가시면 아마 하나의 언어만 할 확률은 낮고 보통은 언젠가 다른 언어도 결국 사용할 가능성이 높습니다. 그 전에 하나의 언어는 확실히 알아두는 것이 경험상으로 좋았던 것 같습니다.

(예를 들어 직무 별로 백엔드라면 자바/스프링이, 데이터라면 파이썬이, 게임 혹은 클라이언트 등 임베디드와 연관있다면 c++, c# 계열 및 유니티(게임의 경우) 처럼

분야마다 보통 메이저한 언어가 정해져있으므로 우선 해당 하나 언어에 대해 깊이 있게 학습하는게 좋을 것 같습니다.)

기타 추천 방법 및 개인적 생각 정리
1. 입사 조건에 관한 생각

저는 사실 누구나 가고 싶어하는 대기업 정도가 아니면 데이터 사이언티스트도 틈새 수요를 노려볼 수 있다고 생각합니다. 작은 스타트업 등에서 프로젝트 등의 실적을 쌓고 회사에서 팀 단위로 연구, 논문을 제출하는 등의 제안이나 진행을 통해 연구실적까지 쌓아서 실력을 인정받고 이직하는 경우 인데요, 이 경우는 개인적인 노력이 굉장히 많이 필요하고 주변 네트워크 시스템을 갖춰서 자신이 올바른 길로 가고 있는지 끊임없이 확인해야 하기 때문에 쉽지만은 않습니다. 그러나 요즘은 커뮤니티 등이 잘되어 있고 인터넷이 발달되어 있기 때문에 자료나 커뮤니티 정보 또한 찾아보기 쉬운 편이며, 캐글 등 온라인 대회 또한 잘되어 있기 때문에 해당 대회에서 수상 등 실적을 내도 굉장히 좋은 회사에 입사할 자격 조건에 들어 갈 수 있습니다.

또한 각 회사마다 도메인이 굉장히 다르고 지원하고자 하는 회사에 대한 프로젝트 경험이 있다면 굉장히 유리할 수 있기 때문에, 지금 교육 과정을 거치면서도 지원하고자 하는 회사에 대한 분야와 연관된 주제를 고민해보고 실제 사용하고 있는 기술 스택을 적용해보면 유리한 점이 있을 것 같습니다.

기타 추천 방법 및 개인적 생각 정리

2 . 공부는 어떻게 해야 하나요?

굳이 두가지로 나눈다면 거시적 관점으로서의 공부와 단기적 관점에서의 공부가 있을 수 있을 것 같습니다.

자신이 현재 필요로 하는 것을 해당 두가지 관점으로 나눠 ' 야생형 학습' 이라는 학습 방법을 적용하는게 가장 빠른 학습 방법이 될 수 있다 생각하는데요,

해당 내용에 대해서는 '함께 자라기'라는 책을 읽어보시는 것을 추천드리고 싶습니다.

인공지능 분야에 있어서 학습 방법은 우선 앞서 말씀드렸던 부분인 백엔드/ 데이터 엔지니어링/ 데이터 사이언티스트, 분석가 에 대한 장단점 및 필요 조건에 대해 먼저 확실히 고민해 중요할 것 같은데요,

각 분야의 공부 내용이 완전히 다르고 맡은 부분이 다르기 때문에 현재 진행하고 있는 프로젝트를 진행하면서도 작게씩은 다 도움이 되겠지만 정말 중요한 핵심 부분들을 다 차이가 있기 때문에 해당 핵심 부분들에 집중하면서 학습을 진행하는 것을 추천하고 싶습니다.

도움되는 글 및 커뮤니티 정리

https://www.wanted.co.kr/wd/42707?utm_source=google&utm_medium=sa&utm_campaign=kr_recruit_web_dsa_apply&utm_term=&utm_content=dsa_socar&gclid=Cj0KCQjwgO2XBhCaARIsANrW2X1HZ3Ut5s_GgQIQf8TOz9lh85-EMu1Y-WQiszVFQ1Fnodu-_3iqqb0aAoOBEALw_wcB

(쏘카 데이터 분석 채용 예시)

https://www.youtube.com/watch?v=MQI-6P9RJZk

(데이터 사이언티스트 방안들)

https://www.youtube.com/watch?v=fLVEfcCAA2Q

문과로 시작한 데이터사이언티스트

https://www.facebook.com/groups/TensorFlowKR/

(텐서플로우 코리아 - 커뮤니티)

https://www.ibatstudio.com/%EB%A7%88%ED%9D%94%EC%82%B4-%EA%B8%B0%ED%9A%8D%EC%9E%90-%ED%94%84%EB%A1%9C%EA%B7%B8%EB%9E%98%EB%A8%B8-%EB%90%98%EB%8B%A4/

(마흔살 기획자 프로그래머 되다)

(또한 취업 시 사람인보다는 원티드 같은 플랫폼을 활용하시는게 상대적으로 좋은 기업 조인에 유리한 것 같습니다. (개인적 의견입니다.)

또한 다음 링크 또한 유용합니다..)

https://github.com/jojoldu/junior-recruit-scheduler **(이동욱님의 주니어 개발자 채용 자체 서비스 운용, 평점 3점이상, 어느정도의 트래픽 처리 및 규모 있는 데이터 확보가 가능한 기업)**

캐글 필사 및 공부 방법
(https://www.kaggle.com)

(권철민님의 글을 참조하였습니다.)

필사하는 방법이 따로 있는건 아닙니다만, 약간의 노하우를 정리해봅니다.

1.너무 쉬운것 부터 하지 마시고, 자신의 수준에서 약간 어려운 것 부터 도전해 봅니다.

2. 하시다가 분명히 막히는 부분이 나옵니다. 이때 포기하지 말고 조금 더 코딩을 해봅니다. 왜 그런지 몰라도 일단 코딩 부터 먼저 해봅니다.

3. 모르는 부분은 구글링을 해봅니다. 모르는 부분은 QA로 남겨 질의응답을 받거나 방안을 찾고 방안을 못찾거나 그래도 모르겠으면 일단 엑셀 같은데다 표기하고 다음 코드로 넘어 갑니다.

4. 해당 커널이 모르는게 난무한다 싶으면 보다 쉬운 다른 커널을 찾습니다. 특히 어떤 커널들은 시각화라던가, EDA에 너무 많은 코드를 할애하기도 합니다. 이런 커널 보다는 데이터 전처리나 모델을 효과적으로 구현한 코드를 찾아보시면 됩니다. 커널 검색해서 Most votes 순으로 찾아보셔도 됩니다.

5. 반복, 또 반복합니다. 여러번 다른 경연대회 코드를 필사하지만, 세개 정도 해보시면 내가 제대로 하고 있는지, 잘 이해가 안되는 부분이 많은 경우가 있습니다. 이럴때는 다시 새로운 경연대회를 하지 마시고, 기존에 했던 경연 대회 코드를 다시 복습해 보세요. 세개 경연 대회 코드정도 마스터 하시면 이제 스스로 노하우가 생깁니다.

6. 가장 중요한건 , 꾸준히 얼마나 오래 하는가 입니다. 지치지 않도록 유의 하시고, 하루에 짧은 시간이라도 필사를 하는 버릇을 만들다 보면, 어느 순간에 수준이 확 오릅니다.

실력은 점진적으로 늘지 않고, 한동안 안늘다가, 갑자기 늘어납니다. 꾸준히가 제일 중요한 요소입니다.

비전분야의 경우, 준비 내용 (1)

1. 기본적인 CNN 지식과, Object Detection/Segmentation에 대한 이해, 이론 및 실습, - 핵심과 역사를 이해해야 합니다.

2. 요새 컴퓨터 비전 분야가 많이 딥러닝으로 흐르고 있지만, 딥러닝외에 기존 컴퓨터 비전 영역에서 잘 활용되는 패키지나 프로그램 언어를 배우면 좋을 것 같습니다.

취업까지 시간이 얼마나 있으신지는 모르지만, 아래 정도 하면, 신입 스펙으로는 차고도 넘칠거라 생각됩니다(이 이상 바라면 회사가 도둑놈 입니다 ^^;;)

- Deep learning 기반 컴퓨터 비전

a. CNN 이미지 분류, Object Detection, Segmentation 등의 활용, 다양한 데이터 세트로의 적용, 성능 향상 최적화 경험,

b. 논문 이해 능력: 이 분야가 하루가 다르게 발전하다 보니, 최신 논문을 이해하는 것이 중요하게 되었습니다. 하지만 논문 자체를 해석한다기 보다 기술이 어떻게 발전되어 왔는지에 대한 맥락을 이해하고, 어떻게 개선을 하는게 좋을 건지에 자신만의 방향성을 갖추는게 더 중요합니다.

c. 논문 이해 능력과 더불어 구현 능력을 갖추는게 좋습니다. 가장 좋은것 git에 있는 소스코드를 따라해보는 겁니다.

추천 드릴만한 것은

- Keras ResNet 구현 소스
- SSD, RetinaNet 구현 소스(Yolo는 좀 어렵지만, Yolo 부터 구현을 도전하는 것도 괜찮습니다)
- 필요하다면 관심 분야 구현 소스

-다음장으로

비전분야의 경우, 준비 내용 (2)

d. OpenCV
opencv를 다 아실 필요는 없지만, opencv 기본서 하나 정도는 알고 계시거나, 자주 활용해 볼수 있는 자그마한 프로젝트 정도는 해보시는게 좋을 거 같습니다.

e. 프로그램 언어: python, C#, Android/iOS
C#이 의외로 컴퓨터 비전 쪽에서 잘 사용됩니다(PC용 클라이언트 만들기 위해서).
Android/iOS를 잘 하실 필요까지는 없는데, 컴퓨터 비전 모듈을 모바일에서 연동하기 위한 기본 활용 능력 정도만 있으시면 될 것 같습니다.

f. 마지막으로 많은 컴퓨터 비전 영역에서 특정 한 분야는 남들보다 더 뛰어나게 이룰 필요가 있습니다.
적다보니 많이 적었군요. 이 정도면 거의 슈퍼맨 급이지만, 다 잘하라는 게 아니라, a~e까지는 거시적으로, f는 집중해서 노력하신다면서 프로젝트 구현 위주로 자신 만의 포트폴리오를 만들어 나가신다면, 취업은 크게 어렵지 않을 것 같습니다.
감사합니다.

(권철민 님 글을 참고하였습니다.)

도움되는 강의 및 책 정리 (기본기) 1

모두의 딥러닝 시즌1 or 2

https://www.youtube.com/watch?v=dZ4vw6v3LcA&list=PLlMkM4tgfjnKsCWav-Z2F-MMFRx-2gMGG&index=2

CS231N(혹은 영문 버전)

https://www.youtube.com/watch?v=3QjGtOlliVI&list=PL1Kb3QTCLIVtyOuMgyVgT-OeW0PYXl3j5

논문읽기 스터디 모임 리뷰 PR000

https://www.youtube.com/watch?v=auKdde7Anr8&list=PLlMkM4tgfjnJhhd4wn5aj8fVTYJwIpWkS

도움되는 강의 및 책 정리 (기본기) 2

한요섭님의 딥러닝 할껀데, 실습만 합니다.

시리즈

https://www.youtube.com/watch?v=kVaBDpwgsGg&list=PLqtXapA2WDqbE6ghoiEJIrmEnndQ7ouys

각 SOTA 및 깃헙 볼수 있는 사이트

https://paperswithcode.com

공부 내용 예시(텐서플로우 코리아)

https://theonly1.tistory.com/1564

직접 보고 추천하는 머신러닝 & 딥러닝 & 수학 총정리(2022)

https://xe.obg.co.kr/programming/4562?order_type=desc&listStyle=viewer&page=12

스파크 엔지니어 예시 (1)

(권철민 님 글을 참고하였습니다.)

이 분야가 워낙 빨리 변하기 때문에 제가 섣불리 어떤 분야의 수요나 전망을 예측하지는 못할 것 같습니다만
먼저 oracle sql과 tensorflow정도를 아시면, 신입으로 취직하기에는 충분한 스펙으로 개인적으로 생각합니다.

1. **채용공고를 찾아보니 spark 수요가 생각보단 없는것(?)같은데 spark와 머신러닝/딥러닝 엔지니어면 경쟁력이 좋을까요?**

=> 충분히 경쟁력이 있다고 생각합니다.

2. **머신러닝 파이프라인, 데이터 처리, 자동화 쪽 엔지니어 수요나 비전이 어떨까요?**

현재 한줄 아는건 oracle sql, tensorflow(머신러닝/딥러닝) 정도입니다.

=> 데이터 처리와 자동화 쪽이 너무 광범위한데 어떤 분야를 말씀하시는지 정확히 잘 모르겠습니다만,
머신러닝 파이프라인이 그렇게 어려운 분야가 아니고, 짧은 시간을 투자하면 충분히 스펙으로 가져 갈 수 있는 분야로 보입니다. 다만 머신러닝 파이프 라인이 충분히 전문성있는 분야라기 보다는 원천 데이터를 머신러닝 모델로 만들기 위한 일련의 자동화된 프로세스 이므로 수요적인 측면은 제한정이 있을 것 같습니다.

데이터 처리와 머신러닝 자동화, 파이프라인 쪽 엔지니어에 관심이 있다고 하셨으니까, 해당 분야의 온라인 강의를 들어보시는 정도 수준에서 기술 스펙을 쌓으시면 충분할 걸로 보입니다. spark, kafka 정도면 충분해 보이고, 추가하자면 flink도 괜찮을 것 같습니다만 , 이 세개를 다 할 줄 아는 사람은 경력도 흔하지 않습니다.

스파크 엔지니어 예시 (2)

또한 spark은 엔지니어링 공부를 하기에 만만하지는 않습니다. Spark가 어려워서라기 보다는 기본적으로 RDBMS나 MPP시스템에 이해가 있어야 아키텍처 측면에서 이해가 빠르게 될 수 있습니다. 제 강의에 있는 DataFrame과 SQL 정도면 신입으로 충분해 보입니다만, 엔지니어링 쪽으로 더 하실려면 Spark Engineering 공부에 시간을 투자하는것도 좋을 것 같습니다. 개인적으로는 spark보다는 RDBMS에 더 시간을 투자하는게 취업에는 더 유리할 것 같습니다. 오라클, MYSQL, POSTGRESQL 중 하나 잡아서 엔지니어링 쪽으로 더 파고 드시면 좋을 것 같습니다. 또한 데이터 엔지니어를 하시려면 SQL은 더 빠삭하게 알고 있으면 좋습니다. 단순히 SQL 기본 정도로는 응용이 힘들 수 있습니다. Spark, Kubeflow, Kafka, airflow 로 머신러닝 파이프 라인을 구축한다고, 이들 툴만 잘 사용하면 시스템 구축이 되는게 아닙니다. 데이터 파이프라인은 이름만 바뀌었지, 이미 20년도 넘은 개념이고, 과거에는 모두 RDBMS에서 SQL과 ETL 툴로 처리 했습니다. 데이터를 어떻게 잘 가공할 수 있는지가 데이터 처리의 핵심입니다. Spark, Kubeflow, Kafka, airflow에 대한 온라인 강의가 있으면 들어보시는 정도에서만 스펙을 쌓으시고, SQL공부도 더 적극적으로 해보시면 좋은 결과가 있을 것 같습니다.

3. spark 공부를 시작하면서 요즘은 TFX나 tensorRT같은것도 공부하려는데 어떨까요?

=> TFX와 TensorRT는 딥러닝 엔지니어로 분야를 정하시면 모르겠지만, 데이터 엔지니어 분야로 정하시면 크게 도움이 안될 것 같습니다.

4. 추천해주시는 공부 루트같은게 있을까요?

=> 특별한 공부 루트라기 보다, 원하는 분야를 정하시고, 팀이나 개인적으로 프로젝트를 실행해서 GITHUB 같은곳에 주기적으로 올려보시면 어떨까 싶습니다.

마무리

꼭 비전 분야를 추천드리는 것은 아닙니다. 비전 분야 또한 아직 많은 한계가 있는 현재도 이런 한계를 극복하기 위해 노력하고 있는 분야입니다.

앞서 말씀드린 내용과 같이 자신이 무엇을 준비해야할지 목표를 확실히 정리하고 지원하고자 하는 회사에 대한 job description도 한번 읽어보시고 해당 분야에 대한 기본기와 기술 스택에 대해 반드시 해야할 내용들을 정리하셔서 충분히 준비하신다면

좋은 결과 있으실 것 거라 생각합니다.

또한 어떤 준비를 하든 커뮤니티나 사람들과 스터디하고 특정 내용에 대해 논의 해보는 등

주변사람과의 경험을 쌓아나가는 경험도 한두개 정도 지속해보시는 것도 좋은 것 같습니다.

네트워크 뿐만 아니라 때로는 예기치 않은 정보를 얻기도 하고, 비슷한 주제로 다른 사람과 교류하고

토의해본다는게 생각보다 동기부여나, 정보 교류 면에서 도움이 되는 것 같습니다.

그리고 실전을 통해서 가장 빠른 실력 향상을 이룰 수 있기 때문에, 실전을 해보시면서 그때 마다 필요한 이론을 찾아가시면서 학습하시면 빠른 시간 내에 성장하실 수 있을거라 생각합니다.

추가하고 싶은 내용이 있으시다면 메일 및 슬랙으로 전달해주세요.(dharana7723@gmail.com)

감사합니다.