



LangCon 2019

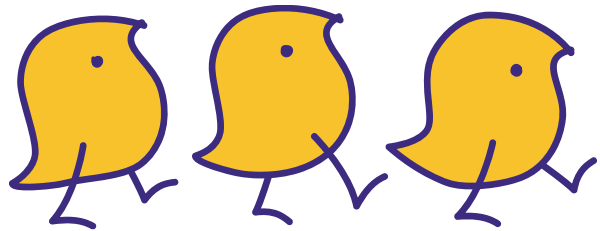


1인 미디어를 위한 자동 하이라이트 알고리즘 을 향해..

홍지민 곽현석

발표자 : 홍지민

1. 배경
2. 문제 제기
3. 데이터 전처리 및 Labeling 기준
4. 모델링
5. 성능 및 기대효과
6. 한계점



배경

공중파 TV채널 에서도 활용 되는 1인 스트리머 콘텐츠



스트리머 & 크리에이터 전성시대




대도서관TV (buzzbean11)
구독자 1,919,825명

홈 동영상 재생목록 커뮤니티



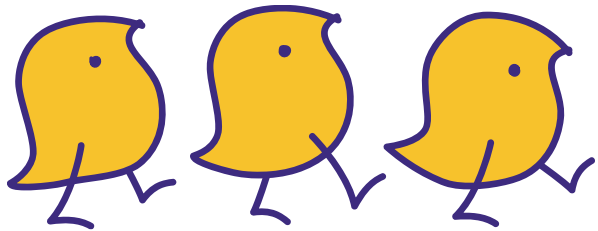
창현거리노래방KPOP COVER
구독자 2,025,859명

홈 동영상 재생목록 커뮤니티

RISABAE
구독자 2,101,640명

홈 동영상 재생목록 커뮤니티



문제 제기

• 축구, 야구 중계 등 전통적인 콘텐츠의 하이라이트 편집

토트넘 3:1 레스터 경기기록



[PL] 손흥민-에릭센 인터뷰 "아직 모든 것은 가능하다&..."

▷ 13,665 | 2019.02.12



[PL] 현지해설 "센세이션얼한 SON! 얼마 나 훌륭한 마..."

▷ 919,651 | 2019.02.11



[PL] 리그 11호골! 손흥민 "다이빙 경고? 억울하지..."

▷ 193,227 | 2019.02.11

SK 5:4 두산 경기기록



[하이라이트] ACE의 귀환 - 김광현

▷ 74,340 | 2018.11.13



[네.야.수] "SK 한국시리즈 우승 축하드립니다"

▷ 40,630 | 2018.11.13



한동민 결승포+김광현 SV SK, 두산 꺾고 KS 우승

▷ 505,114 | 2018.11.12

- 88 원정팀, 허비 반스나가고 오카자키 신지들어옵니다. 클로드 위엘 감독, 오늘 세번째 선수 교체.
- 88 토트넘의 선수교체, 대니 로즈나가고 카일 워커 피터스 들어옵니다.
- 85 레스터 시티의 유리 티엘레만스, 런던에서 경고를 받습니다.
- 84 코너킥을 얻어내는 레스터 시티.
- 83 런던에서 레스터 시티의 웰레치 이헤아나초 슛을 했지만 빗나갔습니다.
- 82 레스터 시티의 제이미 바디, 웬블리 스타디움에서 득점 기회를 맞았으나 슛이 빗나갑니다.
- 80 홈팀, 페르난도 오렌테나가고 빅터 완아마들어옵니다. 마우리시오 포체티노 감독, 오늘 두번째 선수 교체.
- 79 레스터 시티의 허비 반스, 오프사이드에 걸립니다.
- 77 레스터 시티, 공격상황에서 유리 티엘레만스의 슛까지 이어졌으나 토트넘의 수비수가 막아냅니다.
- 76 리카르도 페레이라의 도움으로 기록됩니다.
- 76 골! 제이미 바디의 골로 추격하는 레스터 시티 스코어는 1 - 2.
- 73 레스터 시티의 제이미 바디, 오프사이드.
- 73 토트넘의 손흥민, 웬블리 스타디움에서 오프사이드 판정을 받습니다.

영상편집 | 도비는 원합니다 돈 잘주는 주인님을 (sbs 방송국 유튜브 편집자 역임)

방송하기 | 유튜브 외주편집자를 구한다는 것. (채산성 관련 정보)

안녕하세요. ■■■입니다. 일주일에 하이라이트 영상 3개, 5~10분 분량의 편집 해주실 재택근무 편집알바 구합니다.
(정직원으로 채용 가능합니다. 정직원으로도 재택근무 가능)

동영상 하이라이트 편집하는데

5분당 약 5만원 1분 추가될 때 마다 약 +5천원 ~ 1만원


보통 유튜브를 하시는 분들은 영상 제작이 아주 간단하며 쉬울 것이라 생각하지만,
편집의 힘으로 재미없는 영상을 재미있게 만들어주는 편집은 되게 어려운 것이다.

이런 센스있는 편집은 편집자의 센스가 매우 가미된 것이라 볼 수 있지만
이러한 편집 뿐 아니라
단순한 자막 삽입, 편집, 효과 등을 동시에 하게 된다면,
거의 중노동이나 다름없는 시간을 할애하게 된다.

예를들어보자면 BJ의 영상을 편집하는 편집자는
재미있는 부분을 찾아내기 위해 그 BJ의 영상을 계속해서 봐야하기에 **시청시간이 매우 길다.**
시청시간은 비제이의 팬이라면 자주 보는 것이겠지만 아무리 팬이라도,
의무적으로 비제이의 방송을 보고 재미있는 **하이라이트 부분을 편집**하기 위해 눈을 부릅뜨고 찾아보면


당연히 스트레스가 발생된다. 그러므로 일단 그 영상을 시청하는 시간도 당연히 시급으로 환산해야한다.
대략 5시간 풀 스트리밍 영상을 보고, 편집 시간을 아무리 적게 잡아 2시간이라고 칠 때,
(보통 영상 업계에선 렌더링 시간도 가격에 포함된다.)

하지만 수입이 없는 스트리머들은 어떻게 해야할까?



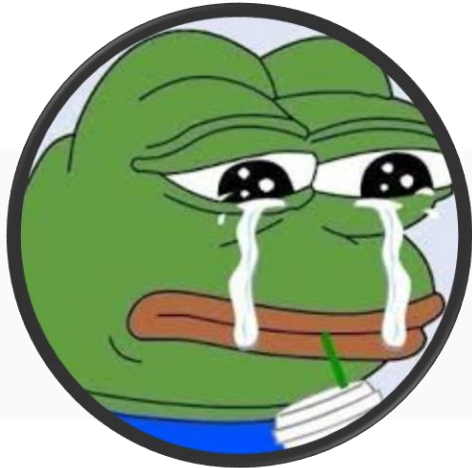
RISABAE
구독자 2,101,640명

홈 동영상 재생목록 커뮤니티

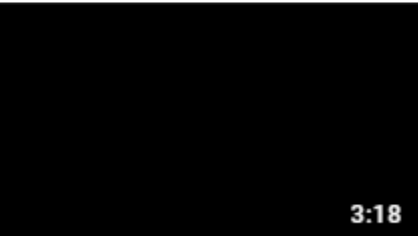
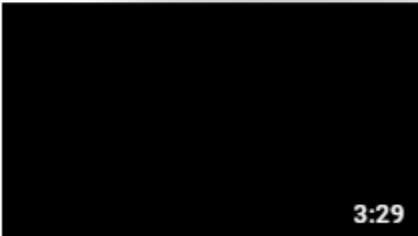


지민 홍지민

홈



업로드한 동영상 ▶ 모두 재생

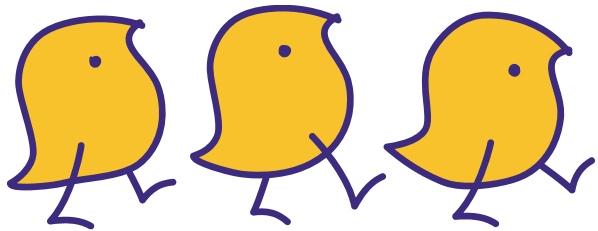
 <p>3:18</p>	 <p>3:29</p>
객석소리 조회수 11회 • 9개월 전	미끄럼틀 조회수 없음 • 9개월 전

하이라이트를 예측해보자!



The image shows a live stream of a man with glasses in a room. On the right, a chat window is overlaid with a red border. The chat contains the following messages:

- 행복한귀지 강유튜브로 가져
- 김승민 그냥 유튜브하죠
- 유리 환유화 저는 유튜브브 잔성입니다
- 최다빈 이게무슨 갑질인지
- 트아 와.. 진짜 짜증나고 억울하시겠어요ㅠㅠ 아 듣는 저도 짜증나네요
- 서지혜 미진놈들
- 지현 서지현 유튜브 생방송은 댓글창 관리 가안되잖아??
- 미승 김미승 #오노답
- 태영 신태영 옹
- 전승민 시노자키하이가 뭐예요??
- 계 미겨랄개 10일이면
- 준우 김준우 아프리카 갑질 ㅋㅋ
- An! 유튜브에서 하셔도 계속 할줄게요! 정보보내거지만 유튜브 오히려 궁금없구름네요
- 이자영 어쭙놈들 쳐먹고 있어
- 임희이 완전 웃긴다
- 김포류집 어후
- Bion 류류 인정



데이터 수집 및 Labeling 기준

데이터 선정의 어려움..

1인 스트리머의 가장 재미있는 부분(하이라이트)를 어떻게 판단하나?

현실적으로 Labeling이 불가 하므로 하이라이트라고 공식적으로 편집해 놓은 영상을

토대로 Labeling을 하자!

데이터 선정 : 롤 챔피언스 리그 코리아 경기

Train set (Past)



Dev set (Relatively Recent)



Test set (Most Recent)



파이썬 Selenium 패키지

```
In [16]: while True:
html=webscrape.page_source
#현재 페이지 소스 html에 저장

soup=BeautifulSoup(html, 'html.parser')
#beautifulSoup 함수를 통해 위에 페이지 소스 html parsing 하기

star=soup.select('#root > div > div.tw-flex.tw-flex-column.tw-flex-nowrap.tw-full-height > div > div.right-column.tw-flex-shrink-0.tw-full-height.tw-relative > div > div > div ')
#css selector를 통해 별점 정보만 긁어오기

li=[]
for i in star:
    li.append(i.text.strip())
a=li[0].split('에게 회신링크 복사')

for n in a:
    example.append(n)

example=list(set(example))

print('-----')
print(len(example))

# for i in a:
#     print(re.split('동영상 채팅|동영상으로 이동|에게 회신링크 복사|게시됨 3개월 전|', i))

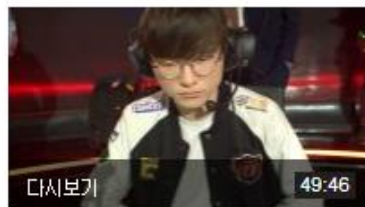
# for i in a:
#     total_review.append(re.split('동영상 채팅|동영상으로 이동|에게 회신링크 복사|게시됨 3개월 전|', i))
#
time.sleep(0.5)
```

유저가 채팅을 한 시간, ID, 채팅을 Table 형식으로 저장

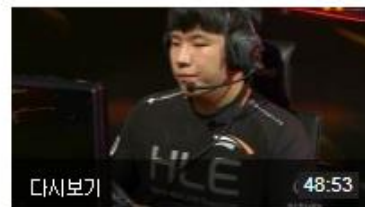
	time	ID	comment
0	4:06:27	초보누에 (dskstkwl)	와 진짜 꼬갓 넘넘 미쳤음 파파파파 ㅠㅠㅠㅠㅠㅠㅠ 개머쓱어
1	4:03:52	남천등 (smartboy1203)	ㄱㅈㅇㅇ
2	3:56:55	운타라최고다 (sogogi0816)	갓구
3	3:41:53	귀여운아롱이 (ero100)	혜지 굿~
4	3:24:25	jang2031	크 역시 동준
5	52:48	lolFranco	야스오 바로리쉬중 ㅋㅋㅋ
6	1:50:19	arnapola	대—————퍼
7	1:14:41	차냥한다 (t53171213)	????
8	4:04:25	naaaaamoooo	소라카도 하는거없어 왜픽함
9	1:58:36	마트로제시카 (matroska94)	장로 미훤
10	2:04:42	ggdog444	대퍼의 기적
11	3:52:20	exteriorglass	트할템머야 저거왜감
12	4:36:30	wbo0622	잼구 : 던짐
13	3:55:12	밤에피는장미란 (kyant13)	뱅 점멸힐 ㅋㅋㅋ
14	50:48	선지쉐이크 (blessthief)	대퍼타임
15	3:18:20	제프티 (leinpols)	애들 신났네 ㅋㅋ

Labeling :
네이버 e-스포츠에서 선정한
하이라이트 기준으로 Labeling

정말 많이 고심했던 부분!



HLE vs SKT 2세트
▷ 144,578 | 2019.02.03



HLE vs SKT 1세트
▷ 145,013 | 2019.02.03



클기문 감수하셨습니다' 오늘의 단독
MP, Clid 인터뷰
▷ 27,361 | 2019.02.03



HLE vs SKT 2세트 하이라이트
▷ 66,411 | 2019.02.03



흔들리는 HLE' 연이어 킬을 만들어내는
SKT
▷ 16,826 | 2019.02.03

HLE vs SKT 2세트 하이라이트



게임 또 타졌다' 한타 승리하며 격차를
벌리는 SKT
▷ 18,826 | 2019.02.03



정글러의 달레마' Clid의 기습에서 겨우
벗어나는 bonO
▷ 18,475 | 2019.02.03



HLE vs SKT 1세트 하이라이트
▷ 57,706 | 2019.02.03



반전은 없다' 1세트 완승을 거두는 SKT
▷ 12,905 | 2019.02.03

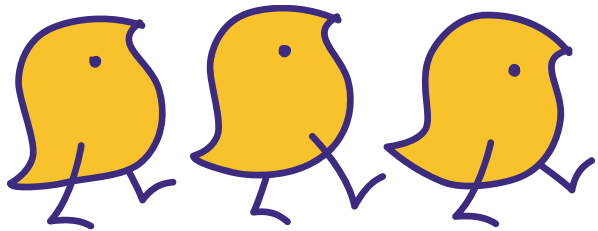


초단위로 예측하기!



네이버 하이라이트 전체

이벤트 발생 후 <채택>



모델링

〈Data 설명 및 Main Task〉

1개의 동영상 당 평균 6시간

하이라이트인 시간대 : 12 % (약 50분) (Positive)

하이라이트가 아닌 시간대 : 88% (약 5시간 이상) (Negative)

-> **Imbalanced Data Problem..**

편중된 데이터의 Positive Class 인지 아닌지 평가를 해야하기에
F1_score를 평가지표로 진행!

Binary Classification!

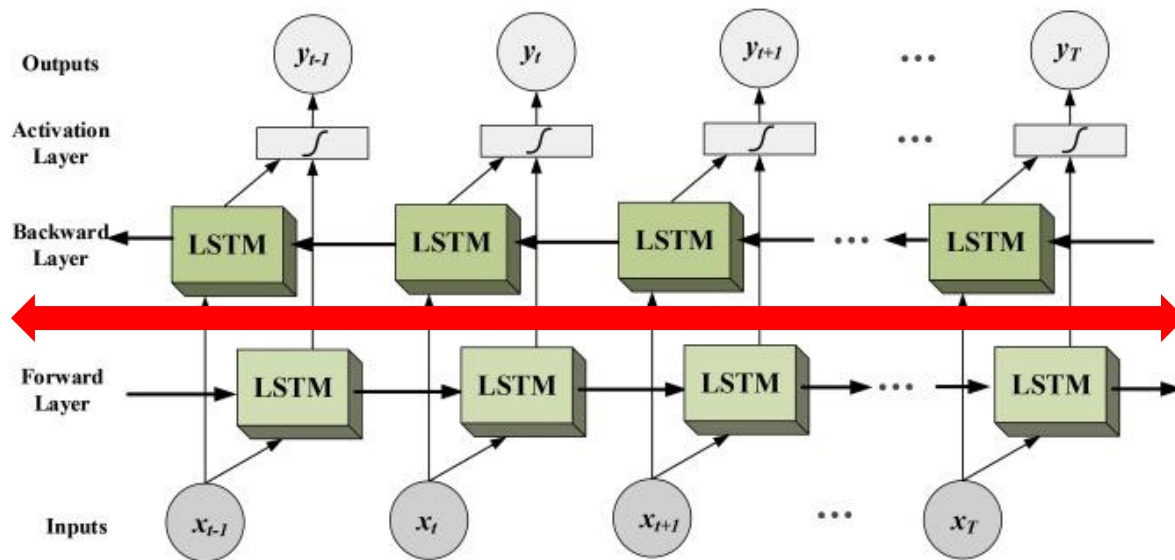
해당 시간대의 채팅 로그 개수

해당 시간대 주변의 채팅로그 개수 합산

+ 지수평활법, 이동 평균법, 형용사 개수, 'ㅋ,ㅎ' 개수 ..등등 다양한 피쳐로 모델링

이후 BI - LSTM 모델로 진행

Time step이 너무 길어 모델
자체에 학습이 안된다!
(6시간 * 60분 * 60초)



생성해낸 피쳐로 Logistic , SVM, Random - Forest 적용

최대

F1 score = 20%.. (in Dev set)

애초에 Data가 imbalance 했기에 생긴 문제가 아닐까 싶습니다.

→ 유저들이 쓴 채팅 단위로 Classification 해보자!

〈Data 설명 및 Main Task〉

1개의 동영상 당 평균 6시간

하이라이트인 시간대에 채팅 : 22 % (Positive)

하이라이트가 아닌 시간대의 채팅: 78% (약 5시간 이상) (Negative)

-> 여전히 Imbalanced Data Problem..

F1_score를 평가지표로 진행!

Binary Classification!

Tokenize문제.. 아무래도 채팅 기록이다 보니 비속어, 은어, 오타 남발로 인해
Token이 제대로 형성이 안되는 문제가 발생

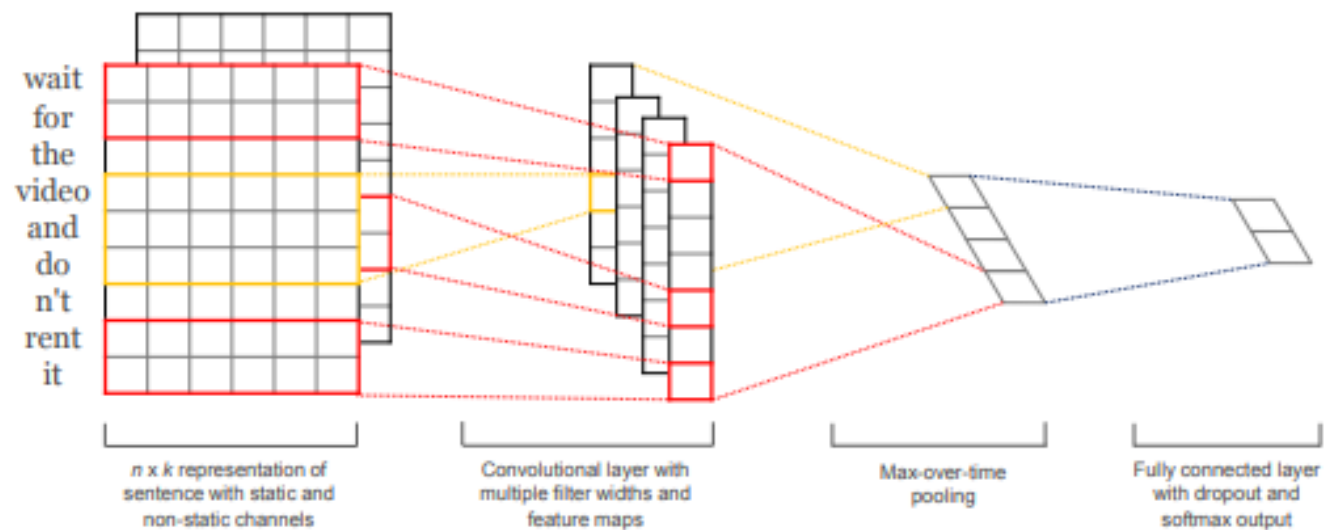
Word2vec 결과 Embedding 되는 단어가 60%가 채 안되는 문제 발생

음절 단위로 Tokenize 해보자!

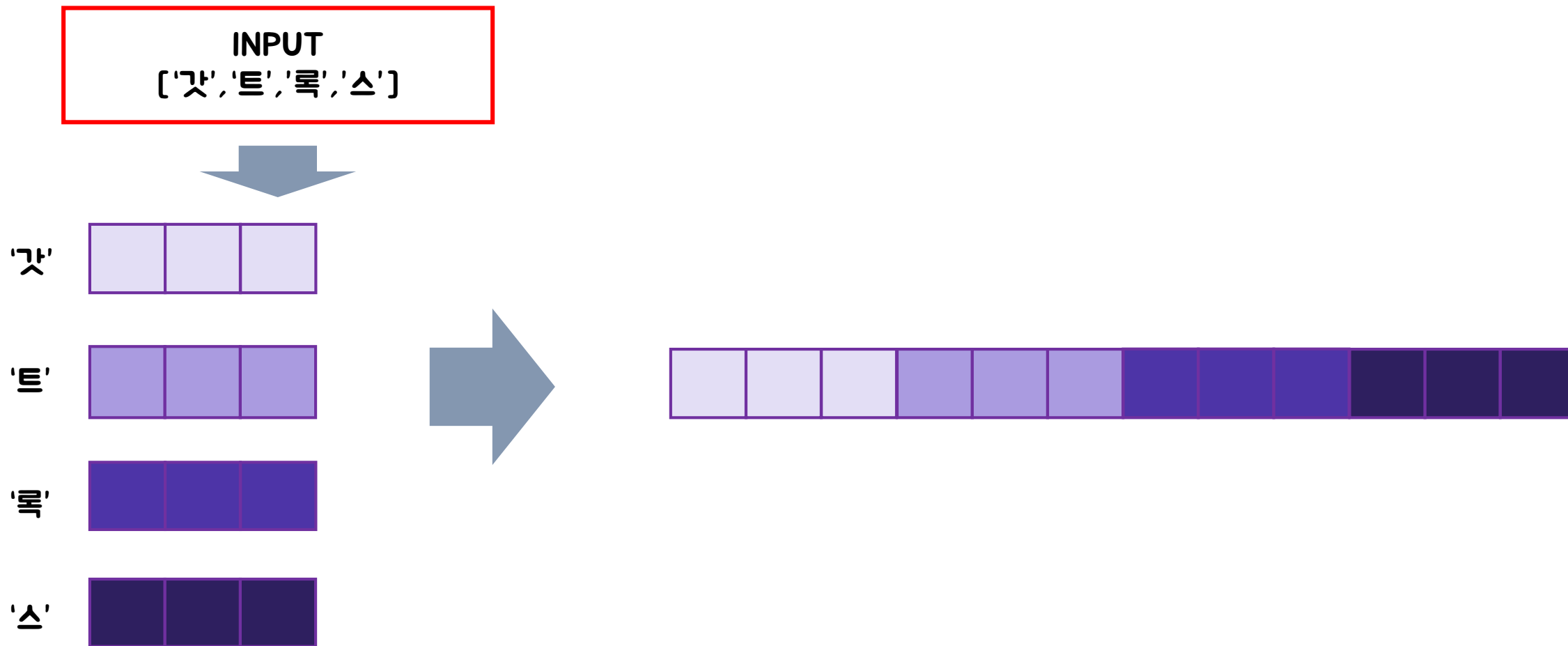
예시 : ['갓트록스'] -> ['갓', '트', '록', '스']

<CNN for Text- Classification>

(by Yoon kim 2014)

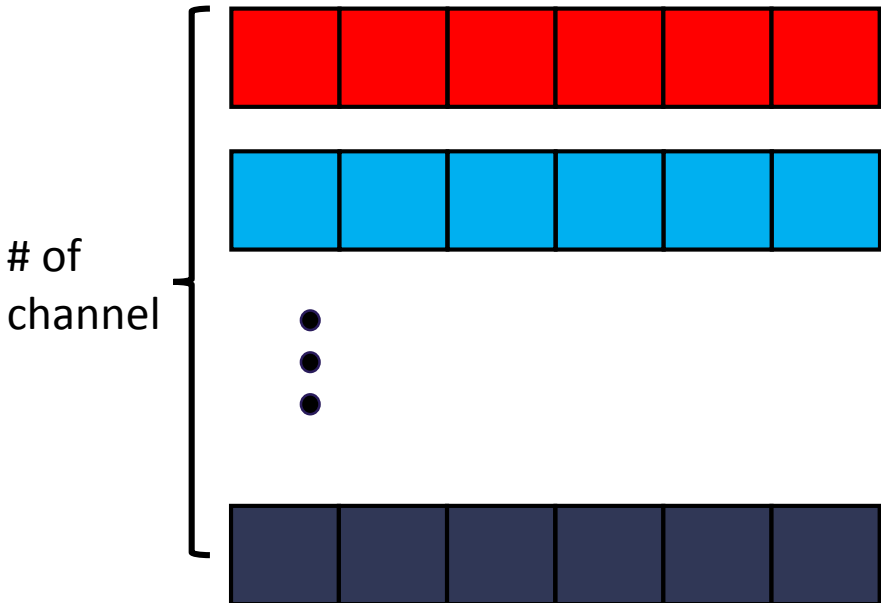


<CNN for Text- Classification>



<CNN for Text- Classification>

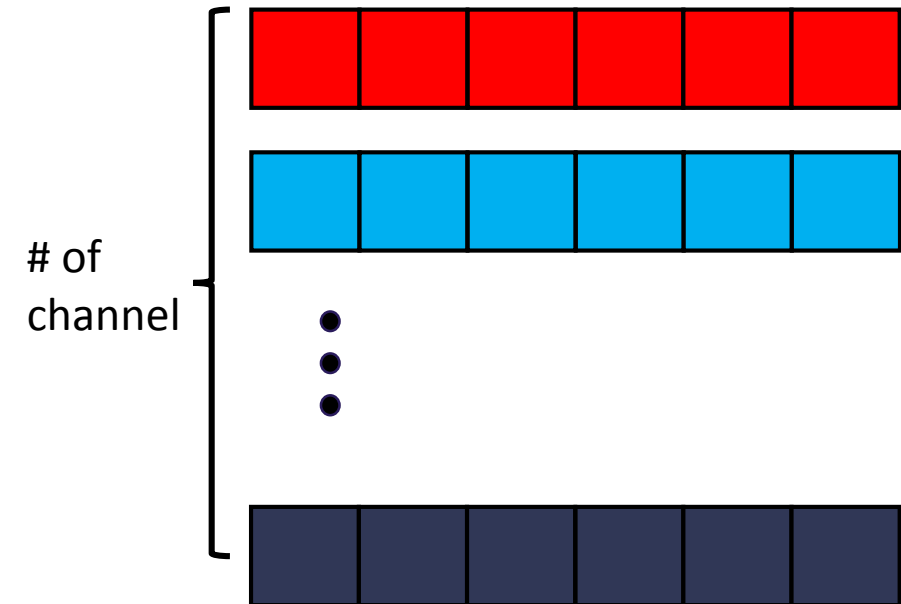
- MAX_SEQ_LENGTH : 4
- EMBED_SIZE : 3
- Filter_sizes = [2, 3, 4] → 한꺼번에 고려할 단어 갯수
- Out_Channels = 10
- Kernel_size = 3 (=word2vec dim)



<CNN for Text- Classification>

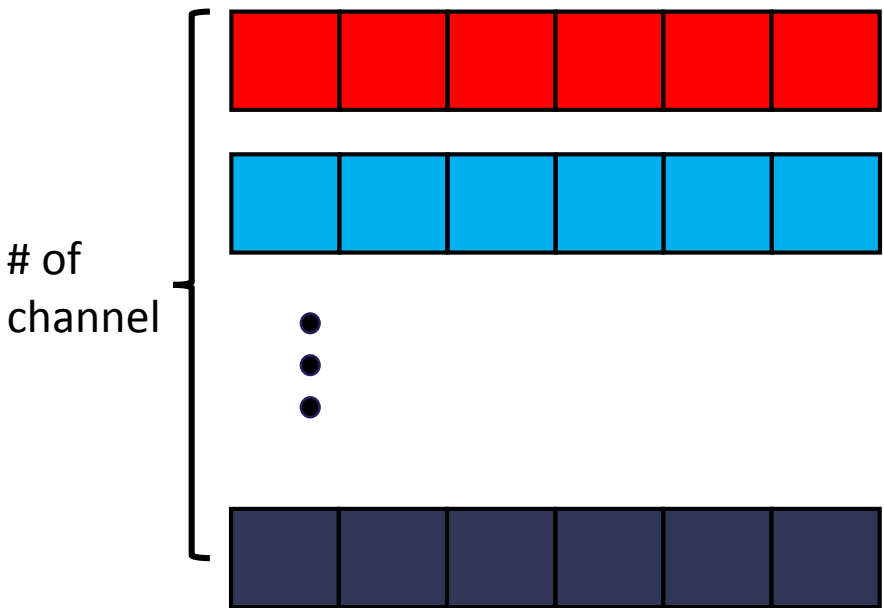
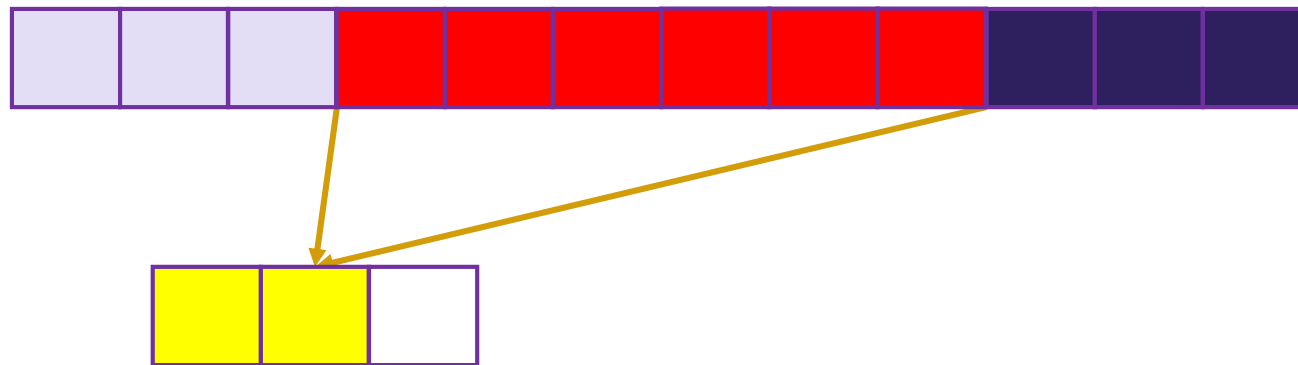
- MAX_SEQ_LENGTH : 4
- EMBED_SIZE : 3
- Filter_sizes = [2, 3, 4]
- Out_Channels = 10
- Kernel_size = 3 (=word2vec dim)

Ex) filter_sizes = 2 (token을 두개 씩 합성곱 진행)



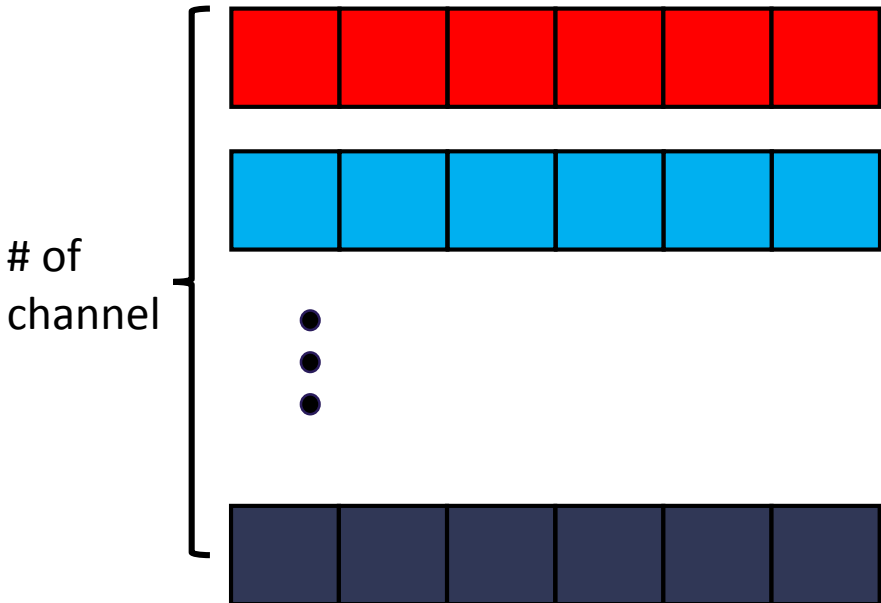
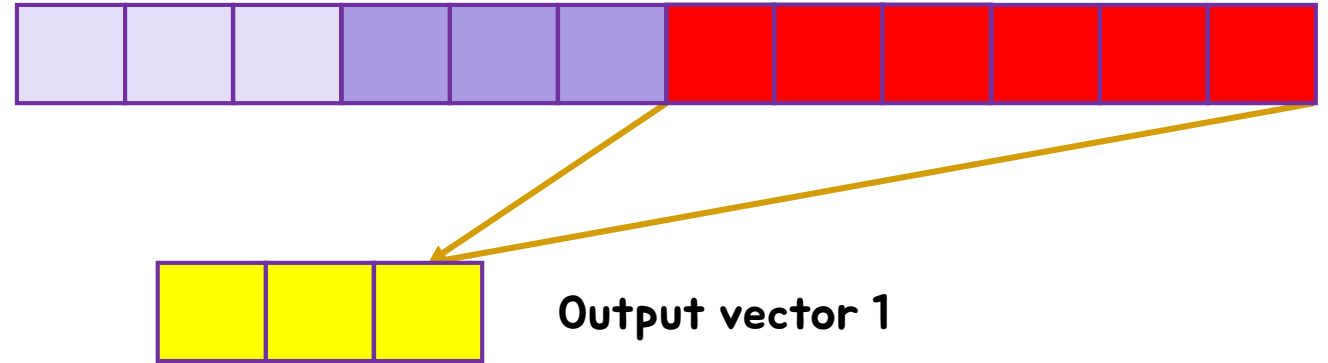
<CNN for Text- Classification>

- MAX_SEQ_LENGTH : 4
- EMBED_SIZE : 3
- Filter_sizes = [2, 3, 4]
- Out_Channels = 10
- Kernel_size = 3 (=word2vec dim)



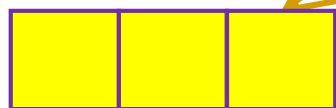
<CNN for Text- Classification>

- MAX_SEQ_LENGTH : 4
- EMBED_SIZE : 3
- Filter_sizes = [2, 3, 4]
- Out_Chns = 10
- Kernel_size = 3 (=word2vec dim)



<CNN for Text- Classification>

- MAX_SEQ_LENGTH : 4
- EMBED_SIZE : 3
- Filter_sizes = [2, 3, 4]
- Out_Channels = 10
- Kernel_size = 3 (=word2vec dim)



Output vector 1

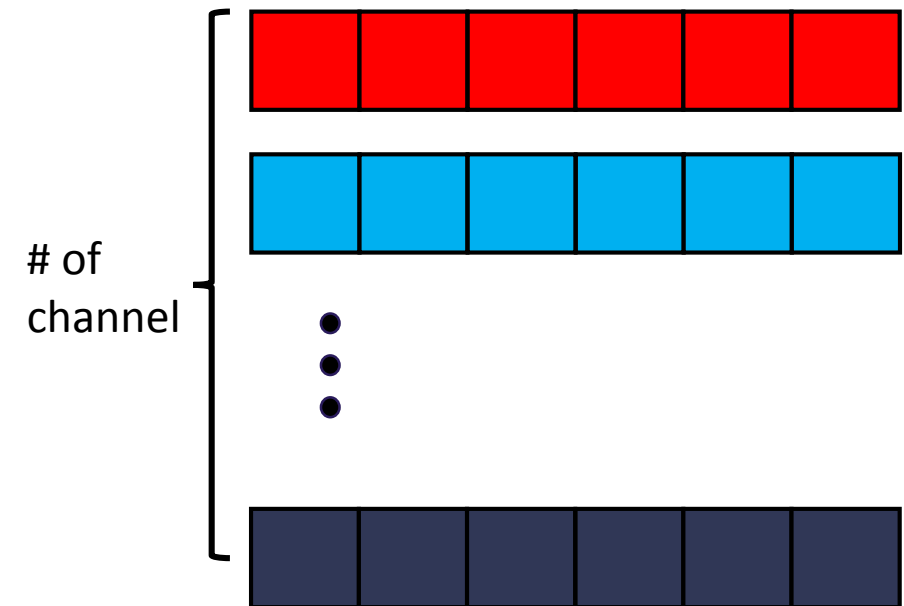


Output vector 2

⋮



Output vector 10

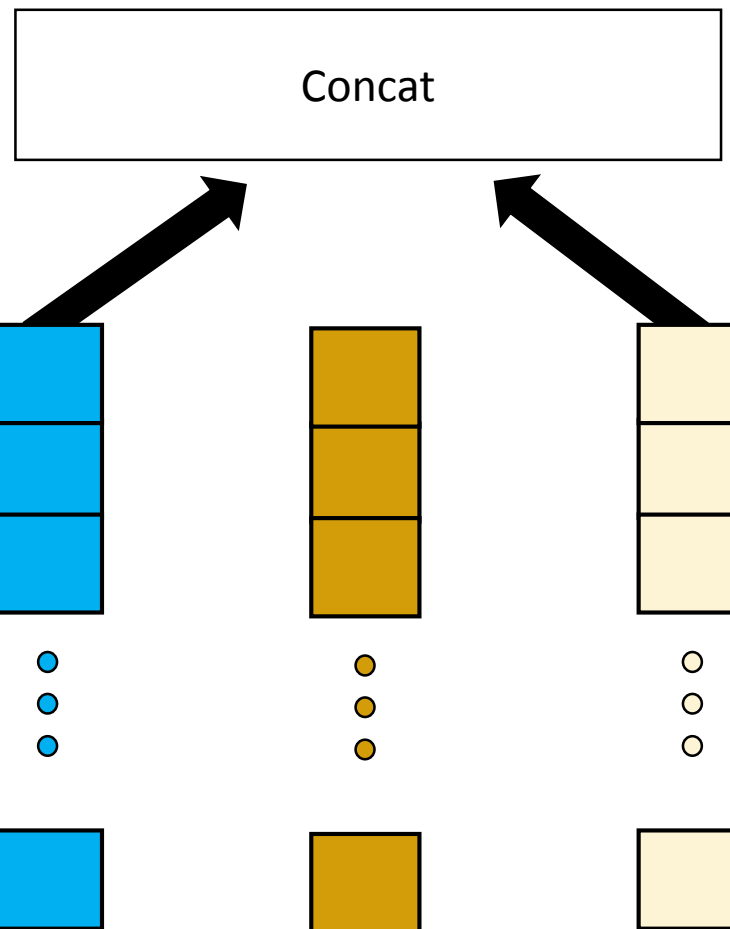


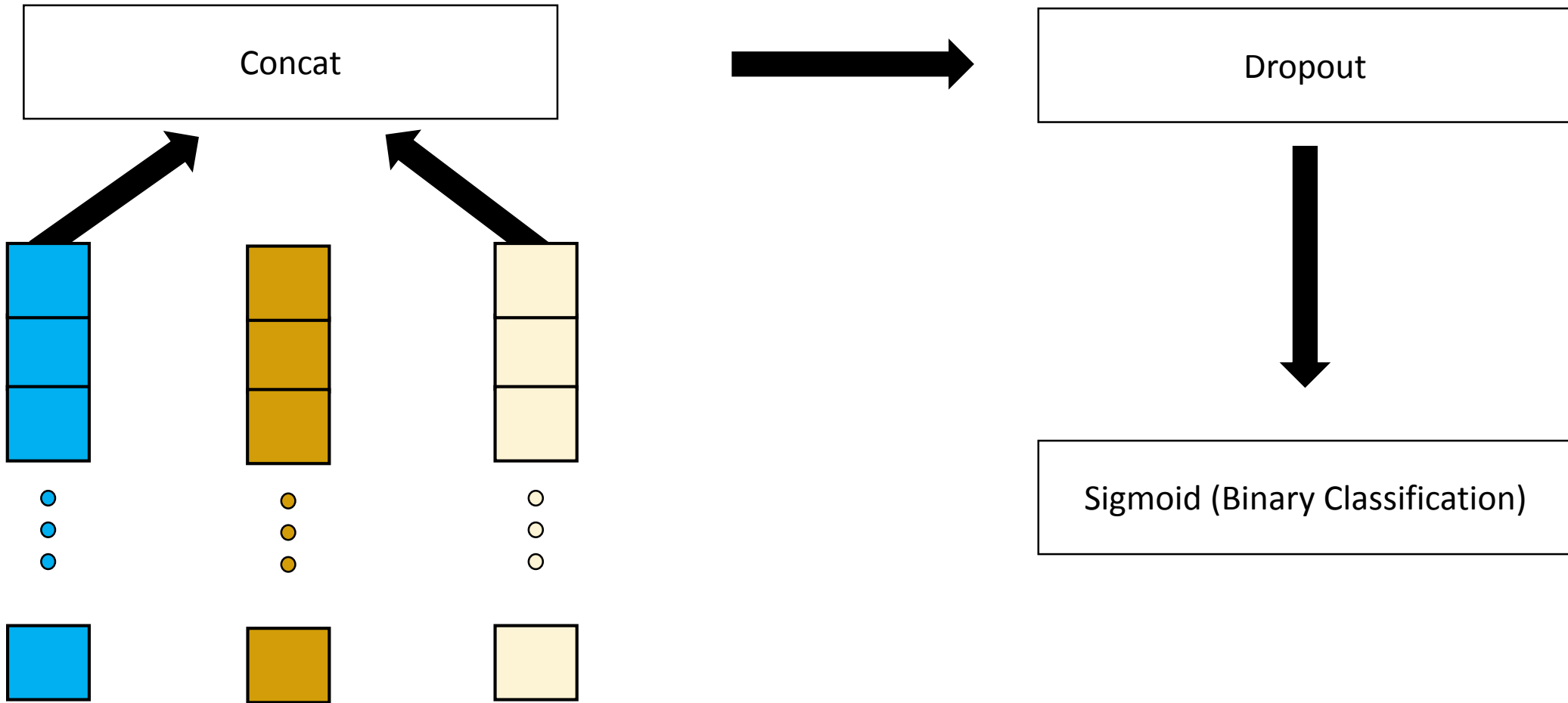
<CNN for Text- Classification>

<Max Pooling>



⋮

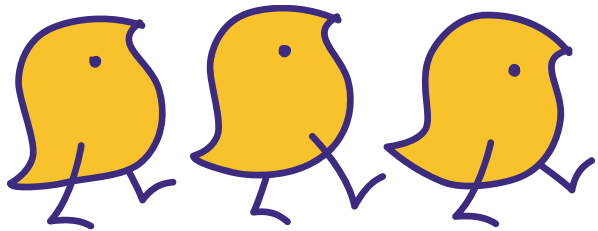




최종 스코어

58% (F1_Score)

다른 모델보다 가장 성능이 좋았다..



성능 및 기대효과

실제 적용된 동영상의 플롯

스트림 서비스 제공자 입장:

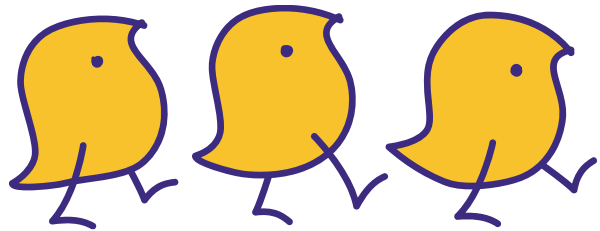
채팅 로그를 통해 자동으로 하이라이트를 생성

→ 하이라이트 편집을 위해 돈을 쓸수 없는 초기 스트리머에게 유인책이 가능

→ 새로운 스트리머로의 유입 가능성

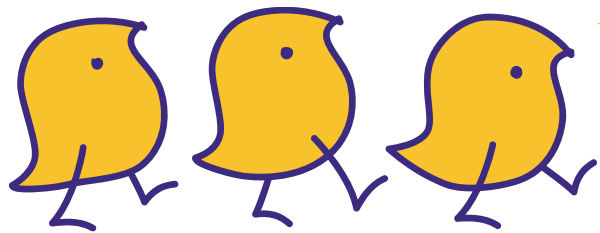
영상 편집자 입장:

하이라이트 편집을 위해 전체 영상을 보고 편집하는 수고로움이 덜어짐
더 높은 퀄리티의 영상 편집 및 디자인을 기대할 수 있음.



한계점

- 실제 개인 게임 스트리머에 적용해보니.. 이런 결과가 나왔다.. 아직 갈길이 멀구나
- 스트림 서비스가 어떤 카테고리냐에 따라 감탄사나 단어가 많이 다를 것이라 예상함. ..> 먹방 뷰티 콘텐츠 등등
- 시계열적인 피쳐를 어떻게 더 활용 했으면 좋은 결과가 나오지 않았을까..?



감사합니다!

Q & A

LangCon 2019