

김윤환, 김홍철, 이종탁, 고준영, 김재현
아주대학교 전자공학과

요약

- 본 논문은 텍스트 기반의 메일을 주제별로 분류하기 위한 목적으로 DNN 구조의 메일 분류 모델을 설계하고 제작함
- 수집한 메일 데이터를 학습이 가능하도록 KoNLPy 패키지와 Gensim의 Doc2vec 라이브러리를 활용하여 형태소 분석 및 벡터화를 진행
- 최적의 성능을 위해 DNN 구조의 은닉 층 수, 노드 수, 학습량을 변화시키며 메일 분류 정확도를 확인

서론

- 메일 분류는 일반적으로 관련 키워드 검색이나 중요한 발신자를 따로 저장하여 분류하는 방법을 이용함
- 웹 메일의 스팸 메일 필터링 서비스가 스팸 또는 비 스팸 메일 구분의 이원적인 서비스를 제공하고 있지만, 효과적인 다원적 분류 시스템에 대한 필요성이 높아짐
- 머신 러닝 기법의 하나인 DNN을 적용한 주제별 메일 분류 시스템을 설계하고 형태소 분석 및 벡터화된 데이터를 입력하여 시스템의 성능을 검증

형태소 분석 및 벡터화

- 시스템의 전체적인 구조 및 기능

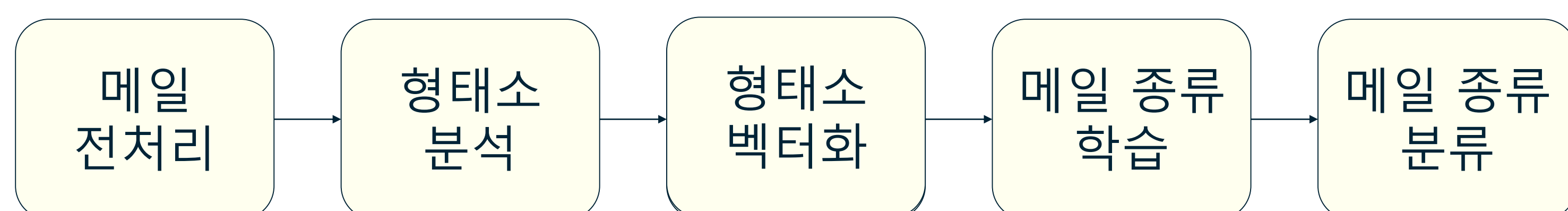


Fig 1. 시스템의 전체적인 구조 도식화

단계	기능
메일 전처리	<ul style="list-style-type: none"> ● 제목 및 보낸 이 제거 : 제목 및 보낸 이만으로 판단하는 경우 배제 ● 메일의 의미 추출 : 내용의 공백, 몇 가지의 특수문자 제거
형태소 분석	<ul style="list-style-type: none"> ● KoNLPy 패키지 사용 → 의미를 가지는 최소 단위인 형태소로 분석 ● Twitter 태거 : 클래스 로딩 시간 및 실행 시간 절약 ● 분석 후 조사, 어미, 숫자, URL 정보등의 품사 제외
형태소 벡터화	<ul style="list-style-type: none"> ● 형태소를 학습이 가능한 형태인 숫자로 벡터화 (크기 : 1X100) ● 벡터화 모듈 : Gensim 라이브러리의 Doc2Vec 모듈
메일 종류 학습	<ul style="list-style-type: none"> ● 벡터화된 메일을 입력으로 주제 예측을 하는 학습과정 → Learning rate, step, 노드 수로 DNN 구조 결정
메일 종류 분류	<ul style="list-style-type: none"> ● 1번(개인 정보 주제)부터 7번(취업 주제)으로 정해진 메일 주제 분류 ● softmax를 수행하고 예측한 값과 실제 정답을 비교

Table 1. 각 단계별 기능 요약

- 전처리 및 형태소 분석 후 검증

순위	취업/Noun	설문/Noun
1	기업/Noun 0.99976	하셔서/Verb 0.99821
2	채용/Noun 0.99904	여러분/Noun 0.99809
3	대기업/Noun 0.99879	설문조사/Noun 0.99793

Table 2. '취업/Noun', '설문/Noun'과 연관성이 있는 단어 상위 3개

형태소	(1)	(2)	(3)	...	(100)
취업/Noun	0.1840	-0.0749	0.0323	...	-0.2506
기업/Noun	0.1265	-0.0465	0.0162	...	-0.1633
채용/Noun	0.2835	-0.1203	0.0483	...	-0.3478
대기업/Noun	0.0562	-0.0213	0.0143	...	-0.0693
설문/Noun	0.2706	0.0279	0.0104	...	-0.4508
하셔서/Verb	0.0496	0.0077	0.0070	...	-0.0831
여러분/Noun	0.1816	0.0179	0.0017	...	-0.3173
설문조사/Noun	0.1669	0.0044	0.0148	...	-0.2870

Table 3. 형태소 분석을 마친 단어들을 벡터화한 결과

- ▶ Gensim 라이브러리에서 제공하는 연관 단어 추출 기능을 사용한 결과, 각 단어가 연관성이 있음을 확인
- ▶ 수치적으로 벡터 값들을 비교한 결과, 벡터의 부호가 동일한 것으로 단어간 연관성을 확인
- 형태소 분석이 잘 되었음을 검증

제안한 최적의 DNN 구조

Learning Rate : 0.003, Dropout : 0.5, 200,000Step

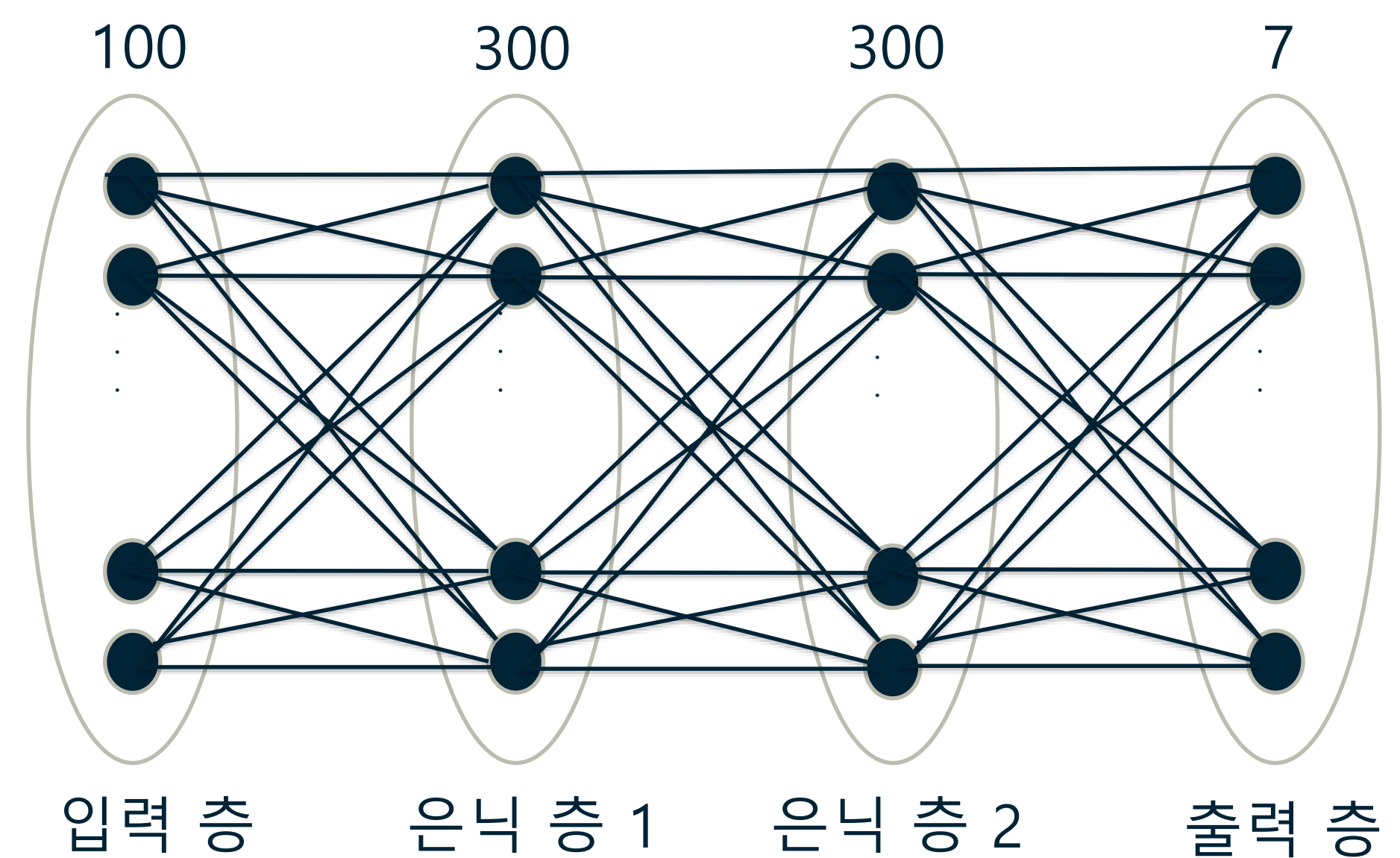


Fig 2. 제안한 DNN 구조

- ▶ Learning Rate = 0.004 이상에서 Overshooting 발생 → 0.003으로 설정
- ▶ Dropout = 0.5 적용 → Over fitting 방지

주제	인식률
개인 정보	67.5
공지, 안내, 알림	54.7
광고	72.5
설문	67.3
인증	86.2
주문, 결제, 환불, 예약	66.9
취업	70.0

Table 4. 제안한 DNN의 주제별 인식률

	정답 주제	가장 혼동하는 주제	오차율
1	개인 정보	공지, 안내, 알림	21.5
2	공지, 안내, 알림	개인 정보	16.3
3	광고	취업	14.8
4	설문	공지, 안내, 알림	17.9
5	인증	공지, 안내, 알림	5.5
6	주문, 결제, 환불, 예약	공지, 안내, 알림	14.0
7	취업	공지, 안내, 알림	18.4

Table 5. 정답 예측시 가장 혼동하는 주제와 오차율(%)

- ▶ 인식률 평균이 가장 높은 주제별 인식률 평균이 가장 높은 [300, 300], Learning rate = 0.003, dropout = 0.5의 DNN 구조로 제안
- ▶ 최저 인식률 : 54.7%(공지, 안내, 알림 주제에서)
- ▶ 최고 인식률 : 86.2%(인증 주제에서)
- ▶ 혼동을 일으키는 주제의 경우 모두 공지, 안내, 알림의 큰 카테고리에 포함됨을 알 수 있음

결론

- 본 논문은 텍스트 기반의 메일 데이터를 수치화하여 DNN 구조를 이용해 학습하고, 7가지의 메일 주제에 따라 분류함
- 메일 데이터의 수치화는 KoNLPy 패키지를 활용해 형태소를 분석 후, Gensim의 Doc2Vec 라이브러리를 활용하여 벡터화를 진행함
- 최종 제안된 DNN 구조에서 '인증' 주제가 86.2%로 가장 높은 인식률, '공지, 안내, 알림' 주제가 54.7%의 인식률, 전체 주제 평균 70.0%의 인식률을 보임

참고문헌(References)

- [1] 박선 외 4명, "자동 주제 생성과 동적 분류 체계를 사용한 이메일 분류", 한국지능정보시스템학회, 제 10권 2호, pp. 79-89, 2014년 11월
- [2] 박은정 외 1명, "KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지", 한글 및 한국어 정보처리 학술대회 논문집, pp. 133~136, 2014년 10월
- [3] N. Srivastava, et al., "Dropout: A simple way to prevent neural networks by overfitting", Journal of machine learning research vol. 15, pp. 1929 - 1958, Jun. 2014.

연락처(Contact Information)

아주대학교 전자공학과 김윤환, 김홍철, 이종탁, 고준영, 김재현

E-mail : kyh1219@ajou.ac.kr, rlaghdclf12@ajou.ac.kr, rkwhr3894@ajou.ac.kr, kdb2658@ajou.ac.kr, jkim@ajou.ac.kr