IJCV 2022

Learning to Prompt for Vision-Language Models

Kaiyang Zhou · Jingkang Yang · Chen Change Loy · Ziwei Liu

- Problem / objective
  - Prompt Learning
- Contribution / Key idea
  - Context Optimization (CoOp)

Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision* 130.9 (2022): 2337-2348.

CLIP

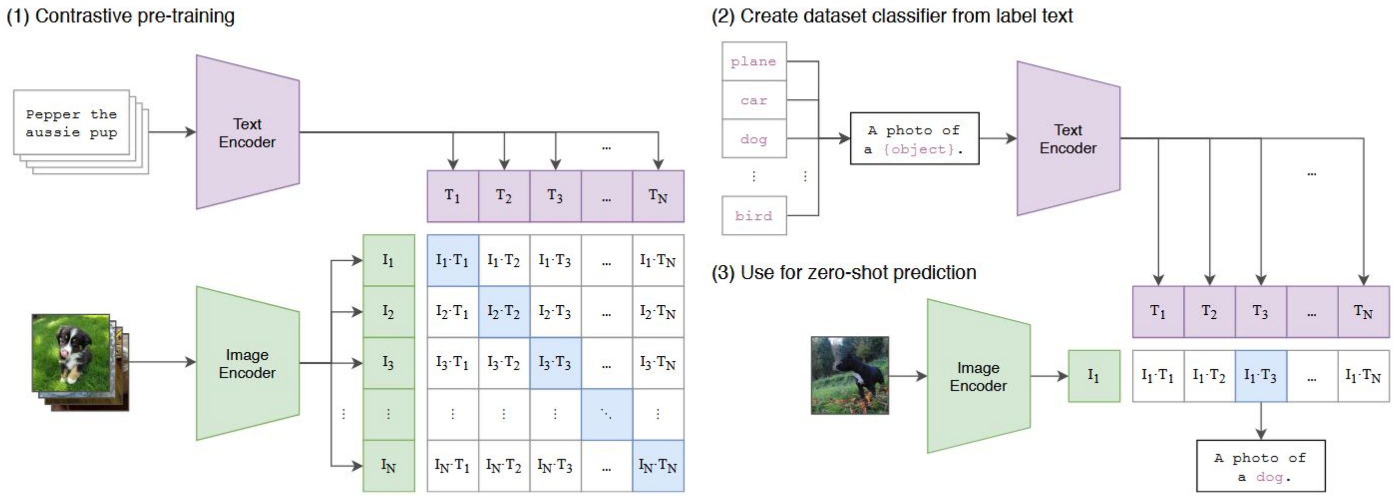


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

전유진

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PmLR, 2021.

Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision* 130.9 (2022): 2337-2348.

Prompt template used in CLIP

- 기존 연구 (그래프에서 파랑선)

  0. 기존 연구에서 사용한 템플릿은 그냥 '레이블'
  "{label}."

- CLIP (그래프에서 초록선)

  1. 디폴트 템플릿
  "A photo of a {label}."

  2. Category 구체화
  예시 1) Oxford-IIIT Pets 데이터셋 -> "A photo of a {label}, a type of pet."
  예시 2) Food101 데이터셋 -> "A photo of a {label}, a type of food."
  예시 3) EuroSAT 데이터셋 -> "a satellite photo of a {label}."

  3. 앙상블
  'A photo of a big {label}", "A photo of a small {label}" 등 포함한 여러 프롬프트들 사용하여 앙상블

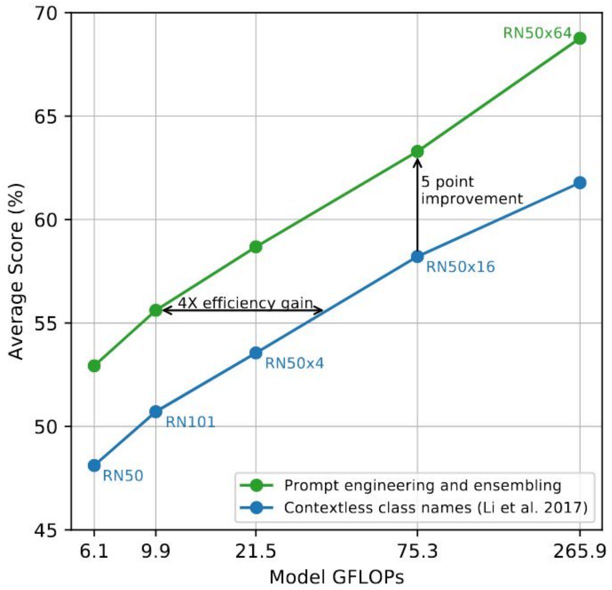


Figure 4. **Prompt engineering and ensembling improve zero-shot performance.** Compared to the baseline of using contextless class names, prompt engineering and ensembling boost zero-shot classification performance by almost 5 points on average across 36 datasets. This improvement is similar to the gain from using 4 times more compute with the baseline zero-shot method but is "free" when amortized over many predictions.

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PmLR, 2021.

Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision* 130.9 (2022): 2337-2348.

IJCV 2022

Motivation

- Pretrained 된 Vision Language 모델을 downstream task 에 사용하려면 'prompt engineering' 잘하는것이 중요.
    - 왜냐하면, VLM의 성능이 프롬프트에 굉장히 예민하게 반응함.
- Prompt engineering 하기 어렵고, 열심히 해도 그게 최적인지 알기 어려움.
- 본 논문 : prompt learning 을 제안 (: prompt engineering 자동화)
- 본 논문의 가치 : NLP 에서 사용되던 prompt learning 개념을 컴퓨터 비전 분야에 처음 도입한 논문.

Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1 [V]_2 \ldots [V]_M$ [CLASS].	**91.83**

(a)

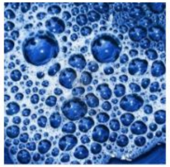
Flowers102

Prompt	Accuracy
a photo of a [CLASS].	60.86
a flower photo of a [CLASS].	65.81
a photo of a [CLASS], a type of flower.	66.14
$[V]_1 [V]_2 \ldots [V]_M$ [CLASS].	**94.51**

(b)

'a' 하나 붙였다고 정확도 5% 이상 증가.

Describable Textures (DTD)



Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1 [V]_2 \ldots [V]_M$ [CLASS].	**63.58**

(c)

EuroSAT

Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1 [V]_2 \ldots [V]_M$ [CLASS].	**83.53**

(d)

Fig. 1 Prompt engineering vs Context Optimization (CoOp). The former needs to use a held-out validation set for words tuning, which is inefficient; the latter automates the process and requires only a few labeled images for learning.

전유진

Method

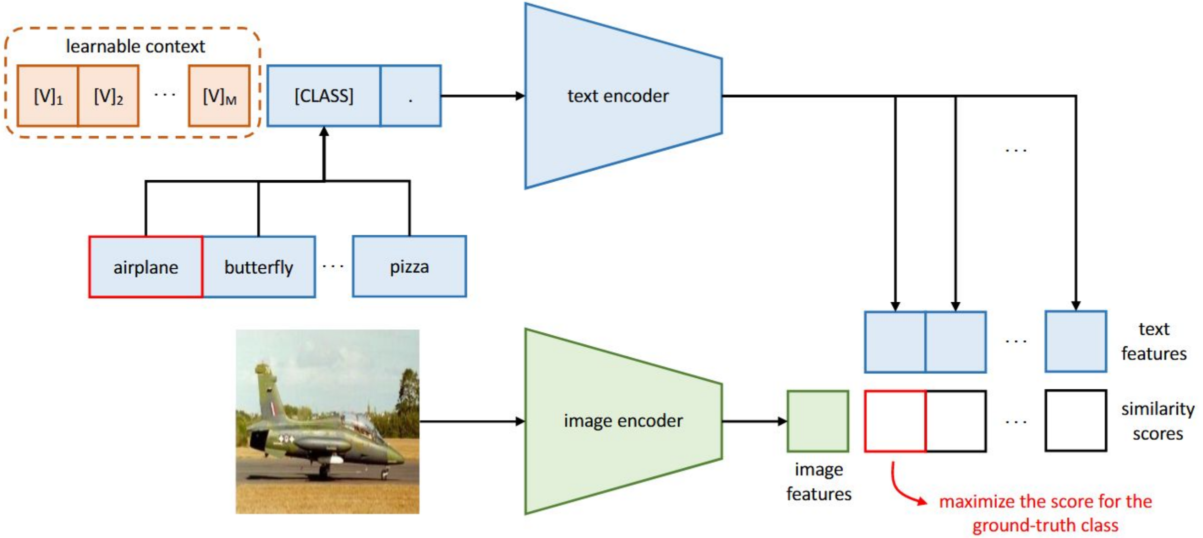
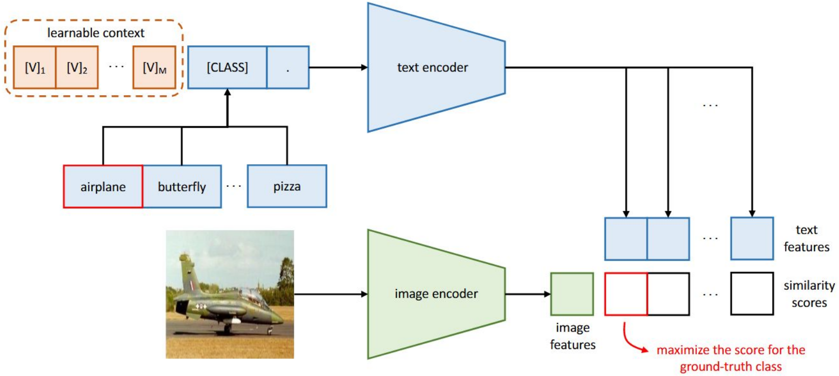


Fig. 2 Overview of Context Optimization (CoOp). The main idea is to model a prompt's context using a set of learnable vectors, which can be optimized through minimizing the classification loss. Two designs are proposed: one is unified context, which shares the same context vectors with all classes; and the other is class-specific context, which learns for each class a specific set of context vectors.

Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision* 130.9 (2022): 2337-2348.

IJCV 2022

Method



- 2가지 유형의 Context vector 제안.
  1. Unified Context : 하나의 context vector
  2. Class-Specific Context : 클래스마다 다른 context vector
- CLASS 토큰을 뒤에 말고 중간에도 넣어보고, 뒤에 description 추가하든 알아서.

$$\boldsymbol{t} = [\mathrm{V}]_1[\mathrm{V}]_2 \dots [\mathrm{V}]_M[\mathrm{CLASS}], \quad (2)$$

$$\boldsymbol{t} = [\mathrm{V}]_1 \dots [\mathrm{V}]_{\frac{M}{2}}[\mathrm{CLASS}][\mathrm{V}]_{\frac{M}{2}+1} \dots [\mathrm{V}]_M, \quad (4)$$

- CE loss 로 학습.

$$p(y = i|\boldsymbol{x}) = \frac{\exp(\cos(g(\boldsymbol{t}_i), \boldsymbol{f})/\tau)}{\sum_{j=1}^{K} \exp(\cos(g(\boldsymbol{t}_j), \boldsymbol{f})/\tau)}, \quad (3)$$

전유진

Zhou, Kaiyang, et al. "Learning to promp ... 2022): 2337-2348.

IJCV 2022

Experiments

- *Few-Shot Learning*

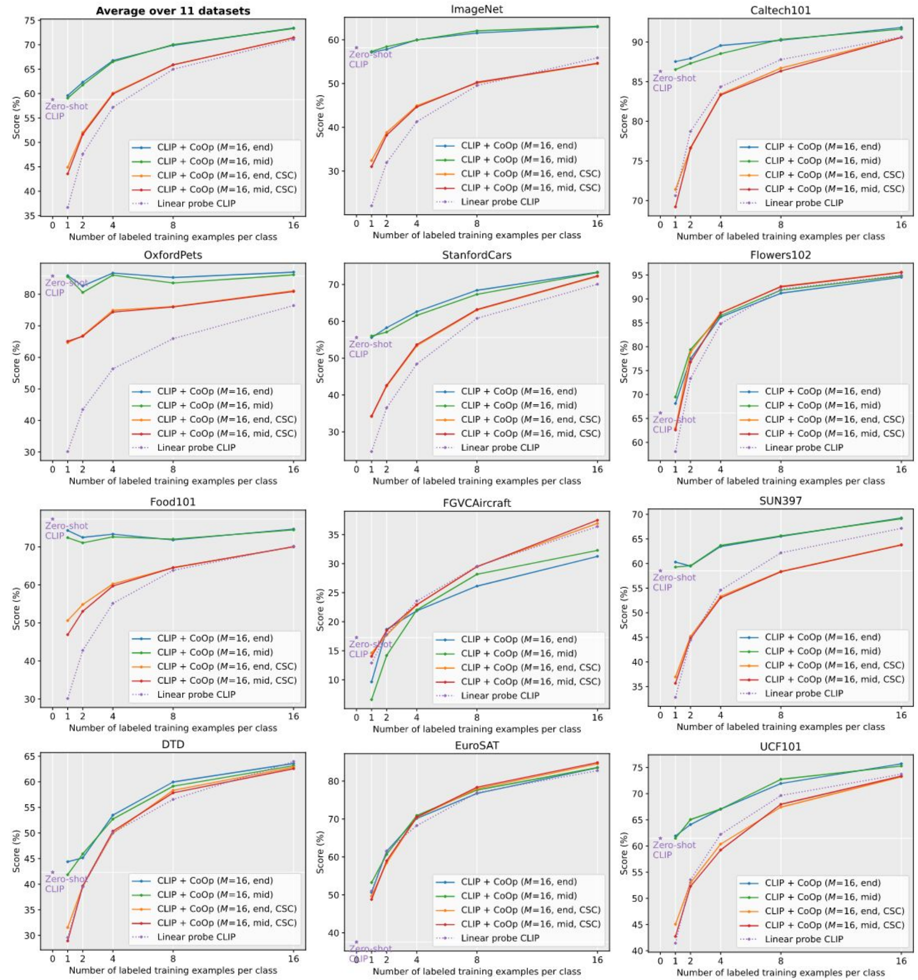


Fig. 3 Main results of few-shot learning on the 11 datasets. Overall, CoOp effectively turns CLIP into a strong few-shot learner (solid lines), achieving significant improvements over zero-shot CLIP (stars) and performing favorably against the linear probe alternative (dashed lines). M denotes the context length. "end" or "mid" means putting the class token in the end or middle. CSC means class-specific context.

전유진

Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision* 130.9 (2022): 2337-2348.

Experiments - *Few-Shot Learning*

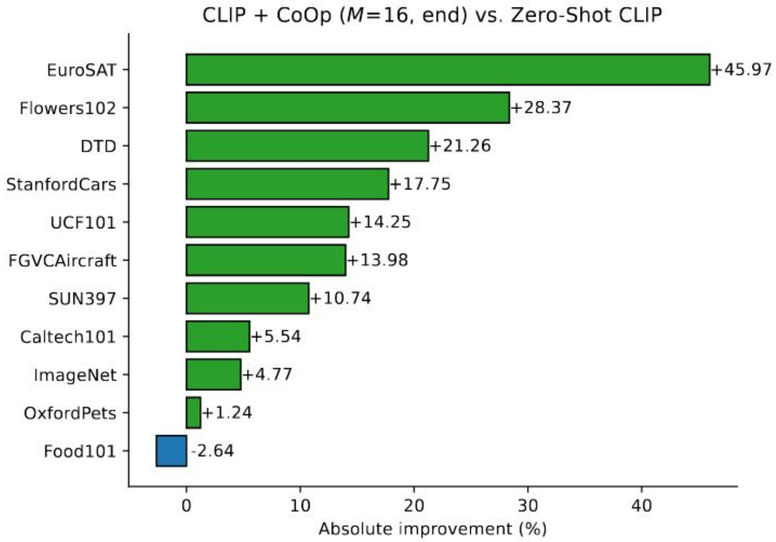


Fig. 4 Comparison with hand-crafted prompts.

Experiments - *Domain Generalization*

Table 1 Comparison with zero-shot CLIP on robustness to distribution shift using different vision backbones. M: CoOp's context length.

	Source	Target			
Method	ImageNet	-V2	-Sketch	-A	-R
ResNet-50					
Zero-Shot CLIP	58.18	51.34	33.32	21.65	56.00
Linear Probe CLIP	55.87	45.97	19.07	12.74	34.86
CLIP + CoOp ($M=16$)	62.95	55.11	32.74	22.12	54.96
CLIP + CoOp ($M=4$)	**63.33**	**55.40**	**34.67**	**23.06**	**56.60**
ResNet-101					
Zero-Shot CLIP	61.62	54.81	38.71	28.05	64.38
Linear Probe CLIP	59.75	50.05	26.80	19.44	47.19
CLIP + CoOp ($M=16$)	**66.60**	**58.66**	39.08	28.89	63.00
CLIP + CoOp ($M=4$)	65.98	58.60	**40.40**	**29.60**	**64.98**
ViT-B/32					
Zero-Shot CLIP	62.05	54.79	40.82	29.57	**65.99**
Linear Probe CLIP	59.58	49.73	28.06	19.67	47.20
CLIP + CoOp ($M=16$)	**66.85**	58.08	40.44	30.62	64.45
CLIP + CoOp ($M=4$)	66.34	**58.24**	**41.48**	**31.34**	65.78
ViT-B/16					
Zero-Shot CLIP	66.73	60.83	46.15	47.77	73.96
Linear Probe CLIP	65.85	56.26	34.77	35.68	58.43
CLIP + CoOp ($M=16$)	**71.92**	64.18	46.71	48.41	74.32
CLIP + CoOp ($M=4$)	71.73	**64.56**	**47.89**	**49.93**	**75.14**

Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision* 130.9 (2022): 2337-2348.

IJCV 2022

Experiments - *Analysis on Context Length and Vision Backbones*

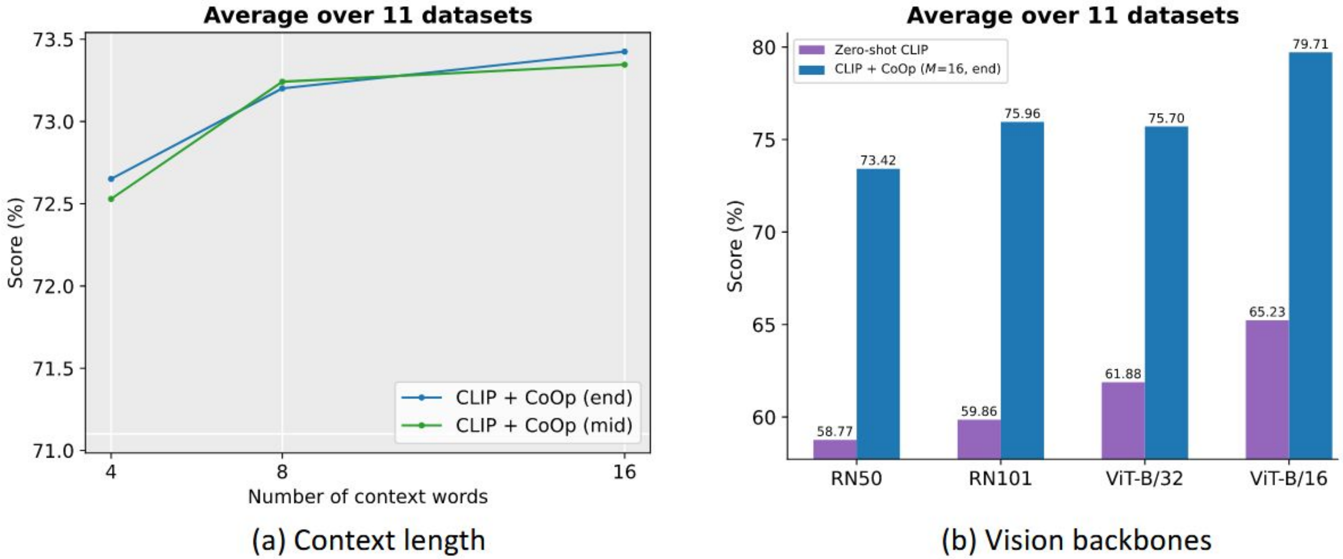


Fig. 5 Investigations on CoOp's context length and various vision backbones.

전유진

Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision* 130.9 (2022): 2337-2348.

Experiments - *Comparison with Prompt Ensembling*

Table 2 Comparison with prompt engineering and prompt ensembling on ImageNet using different vision backbones.

Method	ResNet-50	ResNet-101	ViT-B/32	ViT-B/16
Prompt engineering	58.18	61.26	62.05	66.73
Prompt ensembling	60.41	62.54	63.71	68.74
CoOp	**62.95**	**66.60**	**66.85**	**71.92**

Experiments - *Comparison with Other Fine-tuning Methods*

Table 5 CoOp vs other fine-tuning methods on ImageNet (w/ 16 shots). Δ: difference with the zero-shot model.

	ImageNet	Δ
Zero-shot CLIP	58.18	-
Linear probe	55.87	-2.31
Fine-tuning CLIP's image encoder	18.28	-39.90
Optimizing transformation layer (text)	58.86	0.68
Optimizing bias (text)	60.93	+2.75
CoOp	**62.95**	+4.77

Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision* 130.9 (2022): 2337-2348.

Experiments - *Initialization*

Table 3 Random vs manual initialization.

	Avg %
$[V]_1[V]_2[V]_3[V]_4$	72.65
"a photo of a"	72.65

Experiments - *Interpreting the Learned Prompts*

Table 4 The nearest words for each of the 16 context vectors learned by CoOp, with their distances shown in parentheses. N/A means non-Latin characters.

#	ImageNet	Food101	OxfordPets	DTD	UCF101
1	potd (1.7136)	lc (0.6752)	tosc (2.5952)	boxed (0.9433)	meteorologist (1.5377)
2	that (1.4015)	enjoyed (0.5305)	judge (1.2635)	seed (1.0498)	exe (0.9807)
3	filmed (1.2275)	beh (0.5390)	fluffy (1.6099)	anna (0.8127)	parents (1.0654)
4	fruit (1.4864)	matches (0.5646)	cart (1.3958)	mountain (0.9509)	masterful (0.9528)
5	,... (1.5863)	nytimes (0.6993)	harlan (2.2948)	eldest (0.7111)	fe (1.3574)
6	° (1.7502)	prou (0.5905)	paw (1.3055)	pretty (0.8762)	thof (1.2841)
7	excluded (1.2355)	lower (0.5390)	incase (1.2215)	faces (0.7872)	where (0.9705)
8	cold (1.4654)	N/A	bie (1.5454)	honey (1.8414)	kristen (1.1921)
9	stery (1.6085)	minute (0.5672)	snuggle (1.1578)	series (1.6680)	imam (1.1297)
10	warri (1.3055)	~ (0.5529)	along (1.8298)	coca (1.5571)	near (0.8942)
11	marvelcomics (1.5638)	well (0.5659)	enjoyment (2.3495)	moon (1.2775)	tummy (1.4303)
12	.: (1.7387)	ends (0.6113)	jt (1.3726)	lh (1.0382)	hel (0.7644)
13	N/A	mis (0.5826)	improving (1.3198)	won (0.9314)	boop (1.0491)
14	lation (1.5015)	somethin (0.6041)	srsly (1.6759)	replied (1.1429)	N/A
15	muh (1.4985)	seminar (0.5274)	asteroid (1.3395)	sent (1.3173)	facial (1.4452)
16	.# (1.9340)	N/A	N/A	piedmont (1.5198)	during (1.1755)

전유진