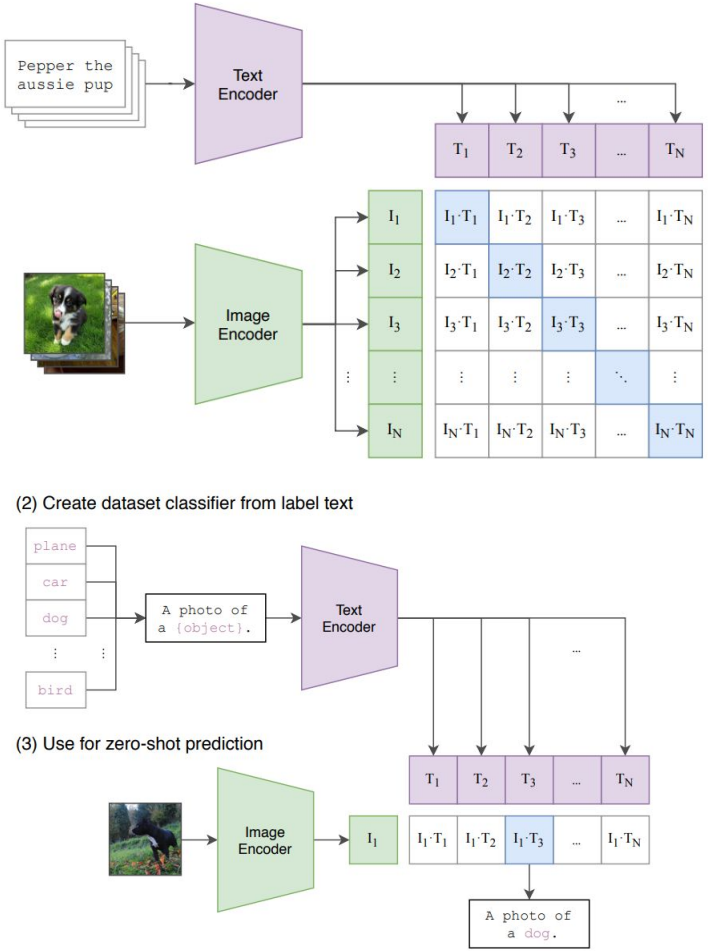
Hyperbolic Image-Text Representations

Захаров Давид БПМИ202

Вспомним CLIP

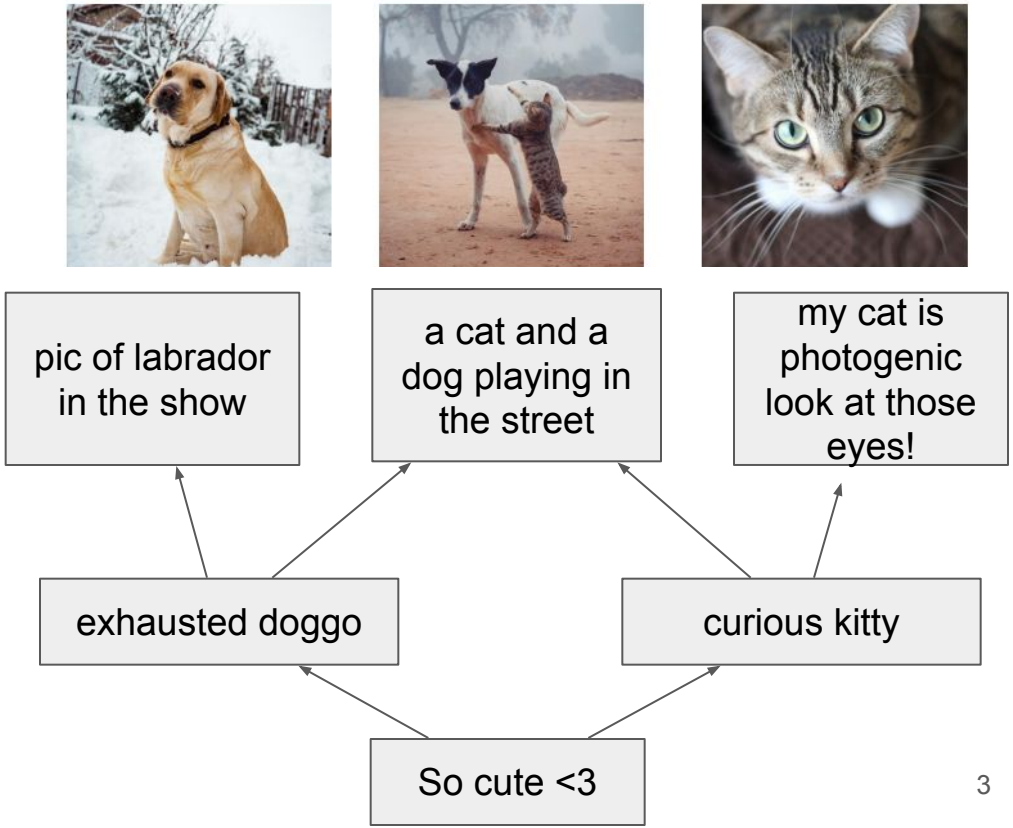
По паре(текс-картинка) создаем согласованные эмбединги

Уже работает на zero-shot задачах



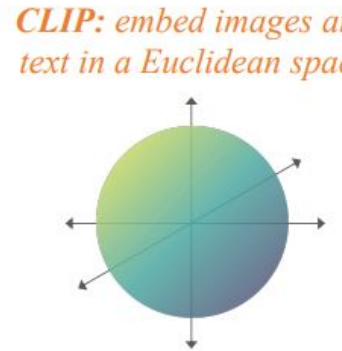
Специфичность текста и картинок

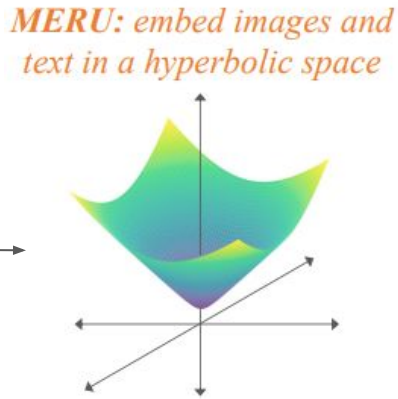
Стремимся учитывать древовидную иерархию текстовых описаний, переходя от общих концепций к более специфическим.

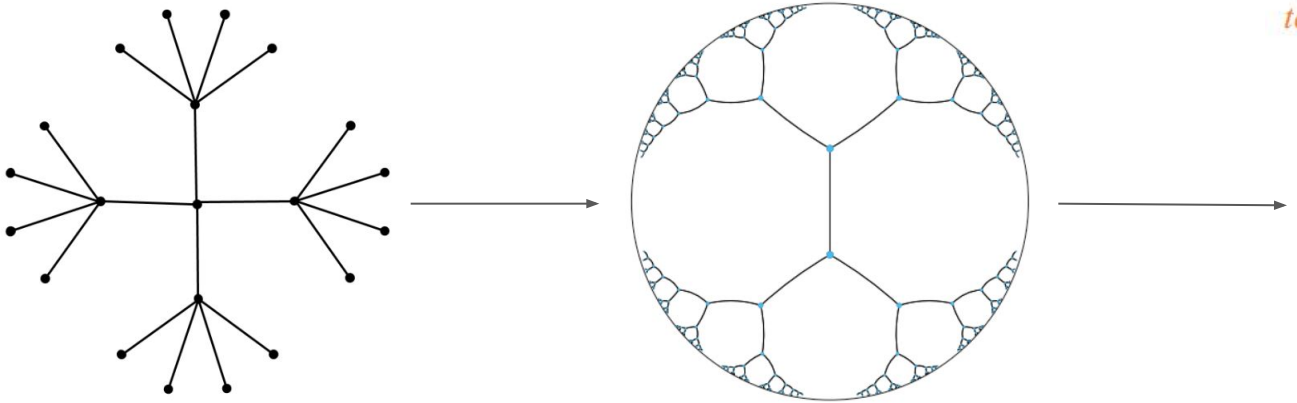


Почему гиперболоид?

Переведем эмбединги в гиперболическое пространство
По мере удаления от "корня", объем в этом пространстве увеличивается экспоненциально,
По сути это непрерывный аналог дерева



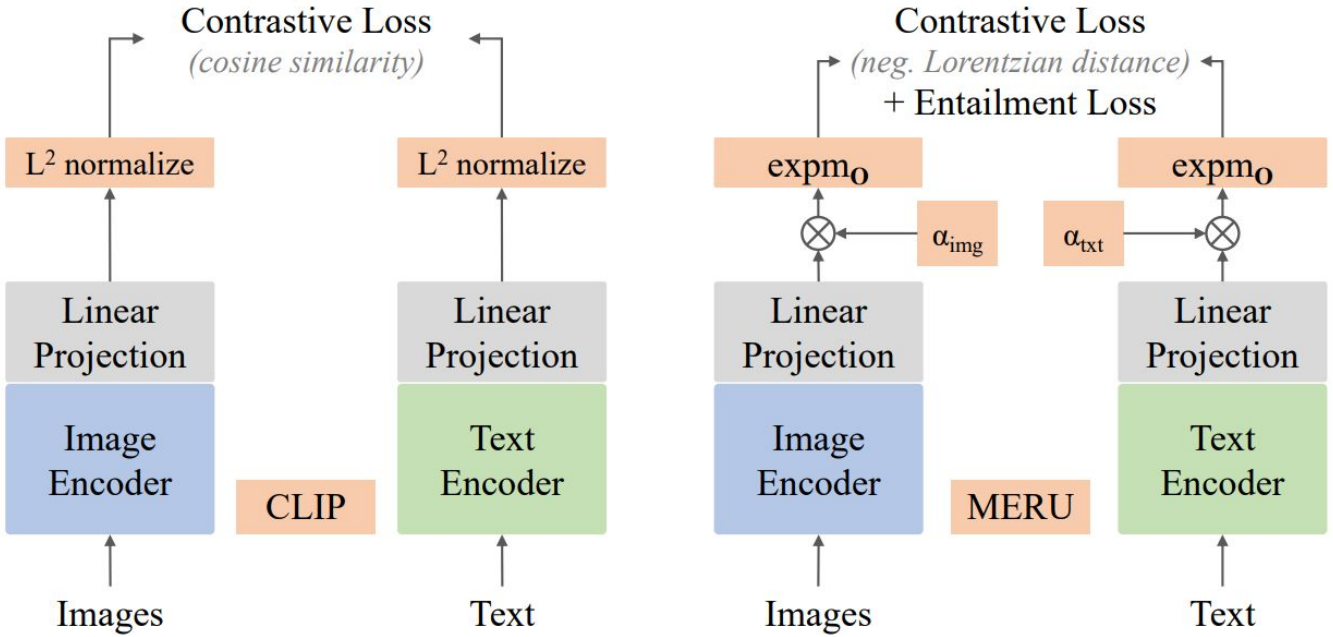




Краткий принцип работы модели

1. Переводим эмбеддинги с гиперсферы на гиперболоид
2. Введем меру расстояния между векторами на гиперболоиде для оценки их схожести.
3. Разработаем специальную функцию потерь, учитывающую иерархическую структуру.
4. В остальном следуем подходу, использованному в CLIP.

Архитектура MERU, contrastive loss

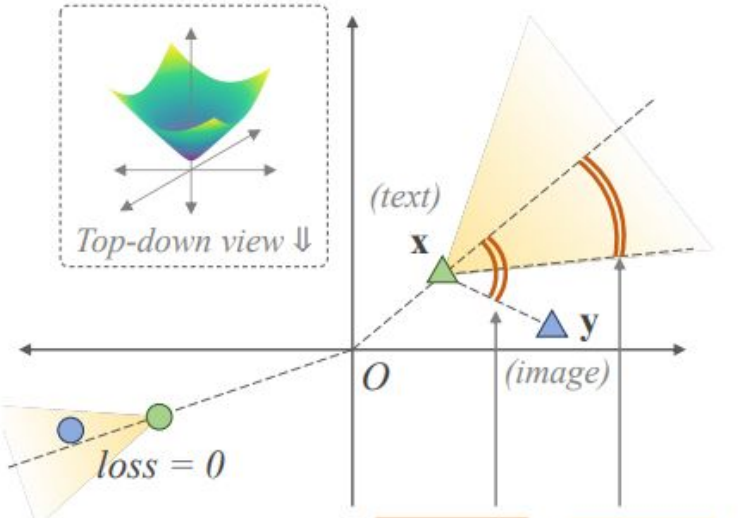


Используем отрицательное значение Lorentzian distance(кратчайший путь) взамен cosine similarities.

$$d_{\mathcal{L}}(\mathbf{x}, \mathbf{y}) = \sqrt{1/c} \cdot \cosh^{-1}(-c \, \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}})$$

Entailment loss

Устанавливаем иерархию, в которой тексты являются более обобщенными, чем соответствующие им изображения.



X - текст, Y - картинка
Хотим, чтобы Y лежал внутри какого-то конуса по отношению к X. Задаем угол

$$\text{ext}(\mathbf{x}, \mathbf{y}) = \pi - \angle \mathbf{Oxy}$$

$$\text{ext}(\mathbf{x}, \mathbf{y}) = \cos^{-1}\left(\frac{y_{time} + x_{time}\, c\, \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}}{\|\mathbf{x}_{space}\| \sqrt{(c\, \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}})^2 - 1}}\right)$$

$$\text{aper}(\mathbf{x}) = \sin^{-1}\left(\frac{2K}{\sqrt{c}\, \|\mathbf{x}_{space}\|}\right)$$

Loss $$\mathcal{L}_{entail}(\mathbf{x}, \mathbf{y}) = \max(0,\ \text{ext}(\mathbf{x}, \mathbf{y}) - \text{aper}(\mathbf{x}))$$

Итоговый loss & Inference stage

$$\mathcal{L}_{cont} + \lambda \mathcal{L}_{entail} \quad \lambda \in [0.01, 0.3]$$

Inference

Для упорядочивания текстов (изображений) можно просто вычислить их скалярные произведения с изображением (текстом), так как все остальные используемые функции сохраняют порядок.

Zero-shot эксперименты

Классификация изображений. На датасетах, выделенных серым цветом, обе модели показали результаты, сопоставимые со случайным выбором.

Извлечение изображений и текста. Почему в задаче text2image вторая строка показывает лучшие результаты? Мы не расширяем размер кодировщика текста, только увеличиваем размер ViT.

		ImageNet	Food-101	CIFAR-10	CIFAR-100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	Country211	MNIST	CLEVR	PCAM	SST2
ViT S/16	CLIP	34.3	74.5	**60.1**	24.4	**33.8**	27.5	**11.3**	**1.4**	15.0	**73.7**	63.9	47.0	88.2	18.6	31.4	**5.2**	10.0	19.4	50.2	50.1
	MERU	**34.4**	**75.6**	52.0	**24.7**	33.7	**28.0**	11.1	1.3	**16.2**	72.3	**64.1**	**49.2**	**91.1**	**30.4**	**32.0**	4.8	7.5	14.5	51.0	50.0
ViT B/16	CLIP	**37.9**	**78.9**	65.5	**33.4**	33.3	29.8	**14.4**	1.4	17.0	77.9	**68.5**	50.9	92.2	25.6	31.0	**5.8**	10.4	14.3	54.1	51.5
	MERU	37.5	78.8	**67.7**	32.7	**34.8**	**30.9**	14.0	**1.7**	**17.2**	**79.3**	**68.5**	**52.1**	**92.5**	**30.2**	**34.5**	5.6	13.0	13.5	49.8	49.9
ViT L/16	CLIP	38.4	80.3	**72.0**	**36.4**	36.3	32.0	**18.0**	1.1	16.5	78.8	**68.3**	48.6	**93.7**	26.7	35.4	6.1	14.8	13.6	51.2	51.1
	MERU	**38.8**	**80.6**	68.7	35.5	**37.2**	**33.0**	16.6	**2.2**	**17.2**	**80.0**	67.5	**52.1**	**93.7**	**28.1**	**36.5**	**6.2**	11.8	13.1	52.7	49.3

		text → image				*image → text*			
		COCO		Flickr		COCO		Flickr	
		R5	R10	R5	R10	R5	R10	R5	R10
ViT S/16	CLIP	29.9	40.1	35.3	46.1	37.5	48.1	42.1	54.7
	MERU	**30.5**	**40.9**	**37.1**	**47.4**	**39.0**	**50.5**	**43.5**	**55.2**
ViT B/16	CLIP	32.9	43.3	40.3	51.0	41.4	52.7	**50.2**	**60.2**
	MERU	**33.2**	**44.0**	**41.1**	**51.6**	**41.8**	**52.9**	48.1	58.9
ViT L/16	CLIP	31.7	42.2	39.0	49.3	40.6	51.3	47.8	58.5
	MERU	**32.6**	**43.0**	**39.6**	**50.3**	**41.9**	**53.3**	**50.3**	**60.6**

Ablations

1. Второй loss важен для улучшения интерпретации модели.
2. Если параметр c (кривизна гиперболы) не обучается, это приводит к проблемам со сходимостью модели.
3. В CLIP используется косинусная схожесть, которая является ограниченной функцией. Применение функции cosh^{-1} помогает контролировать рост неограниченного скалярного произведения.

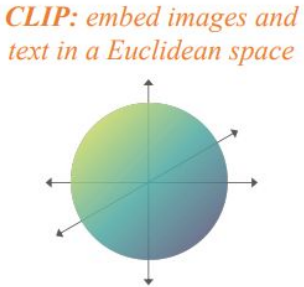
	COCO *text→image*	COCO *image→text*	ImageNet
MERU ViT-B/16	33.2	41.8	37.5
1. *no entailment loss*	33.7	43.5	36.2
2. *fixed* $c = 1$	33.2	42.1	37.9
3. $\langle \cdot, \cdot \rangle_{\mathcal{L}}$ *in contrastive*	32.6	42.3	37.3
MERU ViT-L/16	32.6	41.9	38.8
1. *no entailment loss*	32.7	42.2	33.8
2. *fixed* $c = 1$	0.9	0.9	0.7
3. $\langle \cdot, \cdot \rangle_{\mathcal{L}}$ *in contrastive*	–	*did not converge*	–

$$d_{\mathcal{L}}(\mathbf{x}, \mathbf{y}) = \sqrt{1/c} \cdot \cosh^{-1}(-c \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}})$$

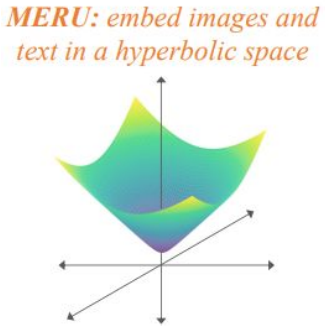
Визуально-семантическая иерархия

Изображения являются листьями, а тексты — промежуточными вершинами. Корень представляет собой самый универсальный элемент.

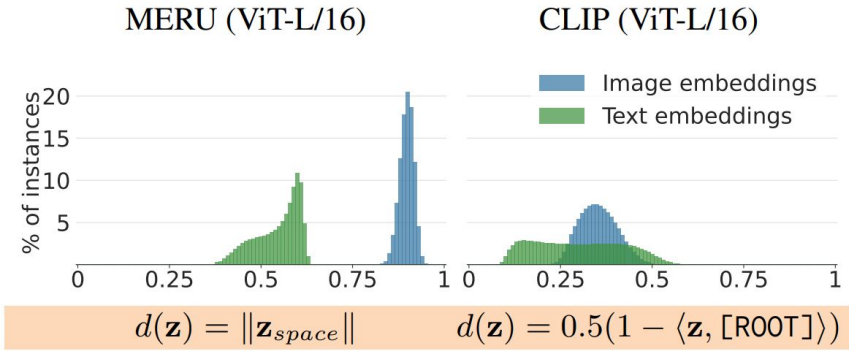
Распределение дистанций до корня показывает, что иерархическая структура функционирует эффективно.



Корень соответствует среднему значению вложений обучающего набора данных.



Корень представляет собой точку на гиперболоиде.



Примеры

MERU	CLIP
squirrel up on the snow covered tree	*squirrel up on the snow covered tree*
squirrel	*squirrel*
wildlife	↓
fluffy	↓
[ROOT]	[ROOT]

MERU	CLIP
seagull	*seagull*
bird	*bird*
air	↓
coast	↓
day	↓
[ROOT]	[ROOT]

MERU	CLIP
cute pug sitting on floor in white kitchen	*cute pug sitting on floor in white kitchen*
pug	↓
domestic	↓
little	↓
[ROOT]	[ROOT]

MERU	CLIP
three zebras	*three zebras*
zebras	*wild animals*
safari	↓
animal photography	↓
wild	↓
[ROOT]	[ROOT]

MERU	CLIP
bread and coffee for breakfast	*bread and coffee for breakfast*
pastry	↓
art	↓
[ROOT]	[ROOT]

MERU	CLIP
grilled cheese	*grilled cheese*
lunch	↓
delicious	↓
classic	↓
[ROOT]	[ROOT]

Для воссоздания промежуточных узлов применяется метод линейной интерполяции, который осуществляется между изображением и корневой точкой.

Плюсы и минусы

1. Интерпретируемые эмбеддинги, поддерживающие семантическую иерархию
2. Разработаны с использованием CLIP, обученного на общедоступных данных с ограниченными использованием GPU.
3. Улучшенная эффективность в вычислениях для гиперболических преобразований.
4. Параметр c (кривизна гиперболоида) подлежит обучению.

1. Тестирование системы ограничено несколькими задачами zero-shot.
2. Качество вложений в R^n оказалось ниже, чем у CLIP, что было подтверждено с использованием linear probe.
3. Entailment Loss, обеспечивающий интерпретируемость, может негативно влиять на качество модели.