

研究ノート

意味空間の画像表現について

Image Representation of a Semantic space

近山 英輔^{*}、中田 豊久^{*}Eisuke Chikayama^{*}, Toyohisa Nakada^{*}

概要

深層学習の技術発展により生まれた大規模言語モデル (LLM) は人間社会を変革する発明であると期待されている。画像処理分野でも様々な人工知能 (AI) タスクが高精度で達成されつつある。マルチモーダル AI では画像の輝度値パターン類似性が用いられているが、画像の意味解釈は輝度値パターンだけで簡単に計算できない。今回 MS-COCO キャプションデータセットと、LLM の falcon-7b-instruct、および対称差集合の手法を用いて画像キャプション間関係性を定量する問題にアプローチした予備段階的結果について述べる。5000 画像のキャプションに対する MS-COCO の 7254 トークンと、それらのキャプションから LLM で連想された 20113 トークンについて画像に関連付け、トークン集合間対称差集合要素数を画像間距離と解釈して分析結果を得た。Falcon-7b-instruct (MS-COCO) に関して対称差距離の最小値は 69 (7)、最大値は 425 (95)、共通部分要素数の最大値は 84 (22)、最小値は 9 (0)、和集合要素数の最小値は 105 (21)、最大値は 467 (101) であった。

キーワード: 意味空間、大規模言語モデル、LLM、MS-COCO

1 Introduction

伝統的な自然言語処理分野の流れから、近年の人工知能 (AI) 分野における深層学習の技術発展により生まれた大規模言語モデル (Large Language Model, LLM) (山田ら, 2023; Naveed, 2023) は、人間知能解明への一歩としてだけでなく、人間社会を変革する発明であると期待されている。しかし自然言語処理に関しての全貌が解明されたことは未だなく、例えば最重要課題の 1 つとして意味とは何かというものがある。意味という単語を国語辞書で調べると言葉が示す内容であると出てくる。しかし意味とは何か、つまり意味の意味に関して明確な答えは未だないとされる (岡崎ら, 2022)。文書レベルでは BoW (Bag of Words) などの単語出現頻度で文書を表現する手法や TF-IDF (Term Frequency-Inverse Document Frequency) で重み付けする手法などが伝統的に自然言語処理手法で用いられている。単語レベルでは分散表現というものがあり、単語はその文脈に現れる単語によって特徴づけられるという分布仮説に依拠して周囲の文脈単語を考慮して構築した単語の意味空間のようなものを数学的な座標空間の 1 点であるとみなし、1 単語を 1 ベクトルとみなして単語に関わる計算に用いるものがある。通常近い単語は近い距離に配置され、ニューラルネットを用いて文脈単語を考慮して単語埋め込みを行ったモデルとして Word2Vec (Mikolov, 2013) はよく知られている。

画像処理分野では AI の最新技術が浸透してきており、通常、画像の類似性は輝度値パターンの

^{*} 新潟国際情報大学経営情報学部情報システム学科

類似性で計算される。それをを用いて様々な AI タスクが高精度で達成されており、輝度値パターンを用いた分野では大きな技術的問題は解決されつつあるといえる。さらに最近では画像とその他の AI タスクを結合したマルチモーダル AI へと研究対象が移行しており、画像と言語を結合した AI タスクの精度が向上してきている。例えば CLIP (Radford, 2021) では text to image タスクを高い精度で実現でき、自然言語を介した画像の結合が可能になっている。これは対照学習により画像空間と text 空間を同様に並び変えたような空間を構成する。しかし、輝度値パターンの類似性が強く出てしまうアルゴリズムであることは否めない。

画像には意味があり、人間はそこから辿ることのできる多種多様な思考でその近さを判断する。そのような近さは輝度値パターンだけで簡単に計算できるものではない。実際、輝度値パターンの限界は自明であるため、例えば実用化されている類似画像検索などは、本質的には、輝度値パターンによるほとんど同じ画像の検索、あるいはキーワードとの照合による検索のようなものに留まっているともいえ、これらのことは人間が連想や推論により辿る軌跡が脳内の視覚信号とどのように関わっているかという科学的問題の解決に関わりを持つようにも思える。

LLM では人間であるかのように、連想や推論をこなすことが知られている。従って、単純なキーワード照合を代替する可能性を秘めている。画像キャプション生成タスク (例えば Wang, 2022) も可能になってきていることから、image to text が可能になり、画像 A と画像 B が与えられたとき、image A to text A、image B to text B を行い、LLM での text A と text B の関係性を探る可能性が考えられる。今後画像キャプション生成タスクはますます高精度になると期待できるため、LLM で text 間関係性を定量的に計測することが求められるのではないかと考える。そのような指標はある意味 LLM 内の埋め込み空間に構築されているとも考えられるが、LLM 内の構造はネットワークのアーキテクチャや学習データなどに大幅に依存するため、理論的に明瞭なアプローチの検討も必要になるものと考えられる。

今回、LLM を用いて、明瞭な計測手法により画像キャプション間関係性を定量する問題にアプローチした。この手法は単純化しすぎているため、文法を無視するなどの明らかな欠点もあるが、今回の結果は画像と言語をどのように情報空間として体系化するかという問題への予備段階的考察である。画像 A、B のキャプションに対し、それぞれ語彙集合を構築して集合間の対称差を用いた集合間距離で計測する試みを行った。これにより連想を考慮した意味空間を構成できるため、それらの各点に参照画像が紐付いていることから、直接的には輝度値パターンによらないある種の画像空間を構築したことになる。従来 AI 分野における画像空間は位置座標と色空間からの輝度値による空間として用いられることが多かった。今後高度な AI を構築するためには、今回のような研究結果を蓄積し、意味空間を画像で表現した空間はどのような情報空間であるべきかについて解決してゆく必要があるものと考えられる。

2 Materials and Methods

MS-COCO データセット (Lin, 2014) は物体検出、セグメンテーション、人物キーポイント、画像キャプション生成などの AI タスク学習に使用できるアイコンックでない (1 画像に 1 オブジェクトではなく多様な文脈を持つ日常の写真などが多い) 画像のアノテーション付き機械学習用画像データセットである。本家サイト (<https://cocodataset.org/>) では 33 万枚の画像を収録している。今回 Python から容易に利用できる Hugging Face の `from datasets import load_dataset; dtst=load_dataset('HuggingFaceM4/COCO');` `dtst['test']` を MS-COCO キャプションデータとして使用した。ここで test とは訓練結果テスト用データのことになる。そこには 5000 画像のデー

タと 25010 のキャプション (1 画像あたり 5 または 6 キャプション) が収録されていた。キャプションとは画像の説明文である。今回 1 画像に対しこれらの 5 個または 6 個のキャプションを 1 文字列に結合して当該画像に割り当てた。これを全 5000 画像について行った。

Falcon モデルはアラブ首長国連邦技術イノベーション研究所が公開したオープンソースの LLM

表 1 MS-COCO キャプションデータの概要

	画像	キャプション	キャプション文字数	トークン数	全トークン数
計	5000	25010	1310066	138627	7254
平均		5.0	262.0	27.7	

である。Hugging Face の LLM 性能比較サイト Open LLM Leaderboard (https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard) で事前学習とインストラクションチューニングモデルの中では 1800 億パラメーター (GPT3 と同程度の数) を持つ falcon-180B モデルが高い HellaSwag スコア、Winogrande スコアを持つ。ただし他のベンチマークも含む平均スコアではより高いスコアを持つモデルも多くある。180B モデルは重量級であるため、性能の低い 70 億パラメーターのインストラクションチューニング型 7B モデル (falcon-7b-instruct) を採用した。Python では、`import transformers; from transformers import AutoTokenizer; tknzs = AutoTokenizer.from_pretrained('tiiuae/falcon-7b-instruct');`
`pipeline = transformers.pipeline("text-generation", model='tiiuae/falcon-7b-instruct', tokenizer= tknzs, torch_dtype=torch.bfloat16, device=0)` で GPU 上で実行させた。1 画像につき入力文字列を以下のように作成して割り当てた。'What do you associate with the following sentences? Describe the answer in one sentence. :¥n' の後に当該画像の 5 または 6 あるキャプションの 1 つを結合し、falcon-7b-instruct でテキスト生成した結果から指示文を削除した文字列を、再度指示文の前に結合し、テキスト再生成を行い、これを 5 回繰り返すことで、最終的に 1 画像の 1 COCO キャプションに対する 5 回の LLM による連想が結合された連想文を得た。それを

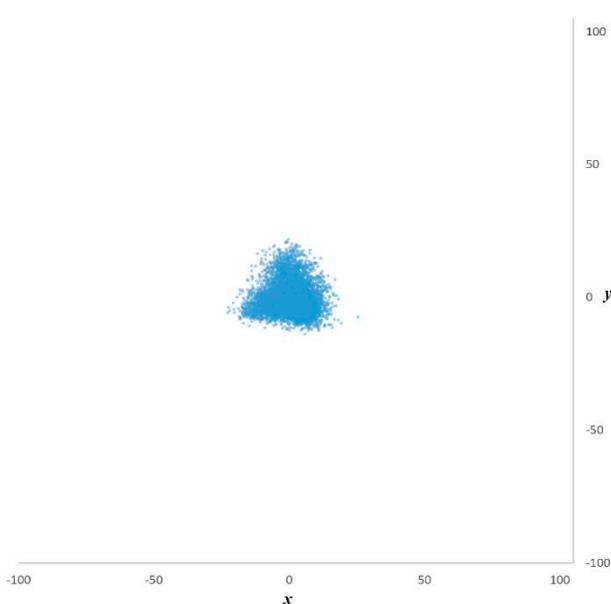


図 1 MS-COCO キャプションデータセット (5000 画像) をユークリッド空間に配置し、2 次元平面へ射影

1 画像の 5 または 6 の各々のキャプションで行い、最終的に全ての連想文を 1 文字列に結合し、これを今回 1 画像に対する 1 文字列として割り当てた。これは 1 画像から LLM が (5 または 6 キャプション) × 5 回の連想を経て得た長文とみなされる。これを全 5000 画像について行った。

画像間の距離を画像説明文間の距離として計算した。文章と文章の距離として、集合 A 、 B 間の対称差集合 $A \Delta B = A \cup B - A \cap B$ の要素数を用いた。これは対称差距離と呼ばれ、距離の公理を満たすことが知られている (Deza, 2012)。

1 文章を 1 集合にするために、MS-COCO データの場合は予め含まれていた、キャプションをトークン単位で分割した 1 次元配列を用い、それをトークン集合 (語彙集合の類) に変換した。Falcon-7b-instruct の場合は 1 画像の連想文を得た後、falcon-7b-instruct 用のトークナイザーで分割して

得た 1 次元配列を用いた。これらのトークンの配列はおおよそ英単語単位で分割されていた。

MS-COCO の 5000 画像の全てについて、前述の方法で 1 画像につき 1 トークン集合が割り当てられたので、画像間距離を計算するために、5000 画像の全ての対の組み合わせに対し対称差距離を計算した。対称差距離から 5000 行 5000 列の距離行列を作成し、計量多次元尺度法でユークリッド空間へ 5000 の画像を配置した。falcon-7b-instruct のトークン集合に対しても同じ計算を行った。計算は Python 3.10.6 でプログラムを稼働させた。集合は set 型を用い、固有値分解は `from import numpy as np; scipy import linalg; linalg.eigh` (2 次元 ndarray 型距離行列) を使用した。

全ての計算プログラムは NVIDIA DGX Station A100 で稼働させ、falcon-7b-instruct を用いるときは、Pytorch 2.0.1 の transformers 4.36.2 モジュールで 1 GPU を使用した。GPU への転送量はおよそ 15 GB 未満、転送時間は数秒であった。コードが最適化されていないこともあり、1 回のテキスト文生成の推論には数秒かかっていたようであった。図 3 の作成には R 4.0.3 を用いた。

表 2 MS-COCO キャプションデータセットにおいて画像 ID 100 に距離が最も近かった画像の例

画像	画像 ID100 との対称差距離; 画像 ID; COCO ID; COCO キャプション
	0; 100; 473237; 'A young girl eating a slice of pizza.', 'A young girl with a green jacket eating a piece of pepperoni pizza.', 'A girl takes a bite of her pepperoni pizza.', 'A little girl eating a slice of pizza in a room.', 'Small girl in green shirt holding a slice of pizza to her face.'
	23; 82; 507065; 'A little boy sitting on a wooden bench eating half a sandwich.', 'a small boy with a green shirt is eating a sandwich', 'A young boy sitting on a bench with a sandwich.', 'A little boy holding half a sandwich in each hand.', 'A boy is holding a sandwich up to his face.'
	24; 2353; 381051; 'A little girl sits on a chair holding a teddy bear.', 'A small girl holding a teddy bear on a chair', 'Small girl sitting in a chair with a teddy bear.', 'A young child sits holding a stuffed bear.', 'A little girl in a rocking chair holding a teddy bear.'
	25; 1725; 27570; 'A woman blow drying her hair, in he mirror.', 'a woman stands in a bathroom blow drying her hair ', 'A woman blow drying her hair in a room with a window.', 'A young woman blow drying her hair in a bathroom.', 'A girl in a bathroom blow-drying her hair.'
	25; 4065; 286981; 'A red hair woman holding an open box of pizza.', 'A young woman holding a pizza in a box.', 'a woman is holding a box of pizza', 'A woman is posing with an open pizza box.', 'A woman holds an open box of pizza.'
	25; 1876; 357096; 'A young child plunging a toilet with a plunger.', 'A little girl trying to plunge a toilet bowl', 'A little girl uses a plunger in the toilet.', 'a small child with a plunger in a toilet', 'A small child plunging a toilet in a bathroom.'

3 Results and Discussion

3.1 MS-COCO キャプションデータセットを用いた解析

MS-COCO キャプションデータを解析した結果、1 画像についての平均トークン数は 27.7 であった (表 1)。また、5000 画像の全トークン集合の和集合の要素数は 7254 トークンであった。

全 5000 画像の対の組み合わせに対し対称差距離を計算し距離行列を作成した。画像空間可視化のために距離行列を計量多次元尺度法でユークリッド空間に配置しようと試みたが、距離行列の

正定値性が満たされなかったため、5000 の固有値のうち正の値は 1266 個になった。0 以下の固有値に属する次元は 0 とし、最大固有値の 2 つの次元を XY 平面として可視化した (図 1)。1 画像 1 点になる。点は数十の範囲に広がり、これは画像間距離が数十程度になることを意味している。

距離行列の要素に注目した例について述べる。今回の予備段階的報告では、最初に詳細を確認

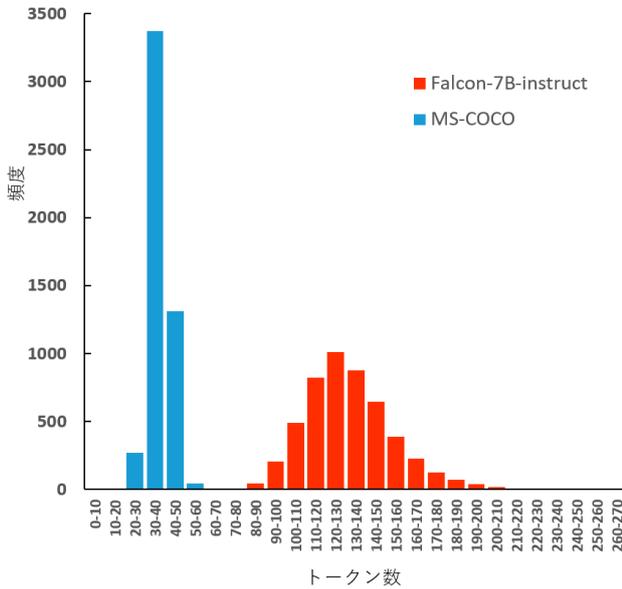


図 2 MS-COCO キャプションデータセット (5000 画像) と falcon-7B-instruct 連想 (5000 画像) における 1 画像当たりのトークン数の分布

した画像 ID 100 の例について述べる (表 2)。画像 ID 100 と距離が 25 以下であった 5 画像についての結果を抽出した。画像 ID 100 (表 2 の 1 行目) に最も近かった画像は距離 23 (2 行目) であった。距離 23 は互いの文章間で異なるトークンが 23 個あったことを意味する。対称差距離は意味が理解しやすい利点がある。画像 ID 100 は女子がピザを食べる画像であるが、サンドイッチを食べる子供という概念的に近い画像が最近接であると認識されている。

表 3 LLM 連想結果の概要

	画像	キャプション	キャプション文字数	トークン数	全トークン数
計	5000	25010	6895429	611131	20113
平均		5.0	1379.1	122.2	

しかし距離 24 (3 行目)、25 (4-6 行目) の中で概念的な類似画像は 5 行目のみであったといえるのかもしれない。

3.2 大規模言語モデル Falcon-7B を用いた解析

LLM の falcon-7b-instruct の結果では、1 画像についての平均トークン数は 122.2 であり (表 3)、MS-COCO キャプションのみの結果よりも語彙が 5 回の連想により拡大していることがわかる。全 5000 画像に出現した全トークン数は 20113 個で LLM による連想により 2.8 倍に拡大した。このことはそれらの分布の変化からも分かる (図 2)。この分布の積算は 5000 画像になる。

対称差集合 $A \Delta B$ の要素数を距離としているため、相殺してしまう語彙の共通部分についても可視化した (図 3)。この散布図の赤の 1 点は 1 画像対を示す。つまり、12497500 点が散布されている。この図は画像を文章として表現した時の文章間の語彙の離れ具合と類似具合をトークン数を軸の単位として表現している。Falcon-7b-instruct (MS-COCO) に関して対称差距離の最小値は 69 (7)、最大値は 425 (95)、共通部分要素数の最大値は 84 (22)、最小値は 9 (0)、和集合要素数の最小値は 105 (21)、最大値は 467 (101) であった。Falcon と COCO の両方で共通部分よりも対称差集合のほうが広がっているため同じ画像ペアの分離能が高い可能性がある。Falcon で共通部分が 0 の場合がないことは連想により語彙が拡大し 2 画像間を結合した可能性があり、COCO が 0 であることは意味空間を埋めるために 2 画像が遠すぎる解釈もあり得る。対称差と共通部分の falcon-7b-instruct (MS-COCO) に対する相関係数は 0.217 (-0.201) であった。これは語彙の不一致と一致の程度に相関がないことを意味している。従って距離をどちらと考えるかで意味空間

の意味が変わってくる考えられる。

Falcon の結果に対し図 1 と同様の計算を行った結果を示す (図 4)。5000 の固有値のうち正の値は 1904 個になった。図 3 の横軸の広がりからも分かるように MS-COCO の配置に比べユークリッド

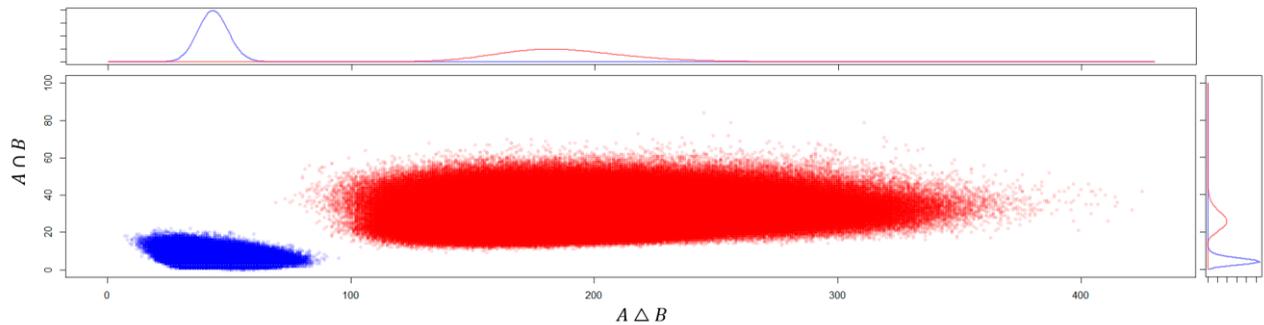


図 3 MS-COCO キャプションデータセット (5000 画像) と falcon-7B-instruct 連想 (5000 画像) における 2 画像間対称差集合の要素数 (2 画像間距離)、2 画像間共通部分集合要素数の散布図、および周辺分布

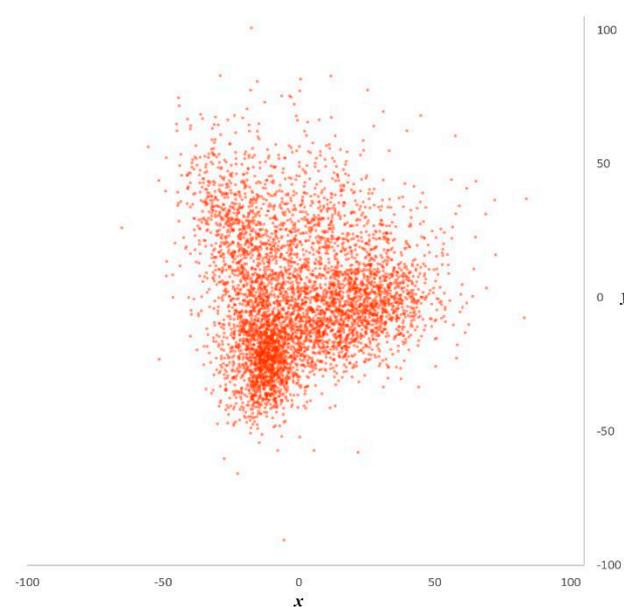


図 4 Falcon-7B-instruct 連想 (5000 画像) をユークリッド空間に配置し、2 次元平面へ射影

空間でも広がっていることが分かる。しかし負の固有値があることから厳密なユークリッド空間とはいえないこと、2 軸の寄与率が Falcon (MS-COCO) で 4% (6%) と低いことからこれらは参考程度の結果になる。

Falcon-7b-instruct による連想文を用いた結果についても、表 2 と同様の抽出を行った (表 4)。画像 ID 100 について距離が 112 以下であった画像は 4 個であった。画像 ID 100 (表 4 の 1 行目) に最も近かった画像は距離 108 (2 行目) であった。距離 108 は互いの文章間で異なるトークンが 108 個あったことを意味し、今回連想の元になる原文が 5 か 6 個あり、それらの各々に連想を 5 回繰り返したため互いの連想文に用いられる単語が拡散してゆくことから距離が大きくなっていることが分かる。ピザが連想されるのは想定内であり、むしろ連想により人物よりもピザが着目されたことになる。4 行目についてもピザであり、5 行目に関してはアップルパイがピザを食べる女子と連想で結合したと考えられる。連想文を読むとキッチンの火から木が焼ける匂いなどかなり飛躍した連想を示しているが、これらは対称差距離の増大に寄与しているであろう。3 行目はなぜサッカーなのかと違和感があるが、連想文を読むとピザが楽しい、とサッカーが楽しい、というような感情表現の文章の不一致性の少なさによるものかもしれない。

謝辞

大規模言語モデル推論タスクを行うための GPU 計算機利用環境を提供していただきました藤田晴啓博士に感謝いたします。

表 4 falcon-7b-instruct (LLM) 連想データセットにおいて画像 ID 100 に距離が最も近かった画像の例

画像

画像 ID100 との対称差距離; 画像 ID; COCO ID; COCO キャプションからの Falcon 連想文



0; 100; 473237;

'A young girl eating a slice of pizza. Pizza is delicious. Pizza is associated with joy, happiness, and comfort. The deliciousness of pizza evokes joy and happiness in people, making it a source of comfort. (Descriptive)', 'A young girl with a green jacket eating a piece of pepperoni pizza. Pepperoni pizza. A young girl enjoying a slice of pizza. A slice of pizza brings joy to young girls. A young girl eating a slice of pizza, experiencing the joy of a slice. Joy.', 'A girl takes a bite of her pepperoni pizza. - A satisfying taste. The girl is happy. - Positive emotion.', 'A little girl eating a slice of pizza in a room. Pizza. The scent of freshly baked pizza. The feeling of satisfaction after enjoying pizza. A child eating pizza. A feeling of joy and relaxation after eating pizza. Pleasure and relaxation. The sentences evoke images of pizza, its delightful scent, and the pleasure and relaxation one experiences after eating it.', 'Small girl in green shirt holding a slice of pizza to her face. - Nostalgic memory of childhood days. A vivid recollection of simpler times. - Nostalgic memory of childhood days. The image evokes nostalgia for simpler times and childhood days.



108; 3658; 132415;

'A white paper plate holding two pieces of pizza. A white paper plate holding two pieces of pizza evokes thoughts of comfort food and relaxation. (Cheese and pepperoni toppings are associated with pizza.) *Pizza is a popular comfort food.', 'A slice and a half of sausage pizza on a paper plate .A delicious, savory meal. The combination of savory and sweet. A slice of pizza with a wide variety of toppings and an even wider range of emotions. I associate the sentence with the feeling of contentment and relaxation.', 'One and a half pieces of pizza on a paper plate. A slice of pizza. A pizza. You associate the sentences with pizza. Pizza slice .', 'an image of a slice of pizza on a paper plate .The slice of pizza on the paper plate represents the concept of food. A slice of pizza symbolizes a meal, nourishment, and the joy of eating. It can also evoke feelings of comfort, satisfaction, and contentment.', 'Pizza is sitting on a plate with a fork. A slice of pizza. -> Food .



110; 2077; 257624;

'there is a young girl playing with soccer balls in a park ;The girl is enjoying herself in the park. The girl playing with soccer balls in a park is a symbol of joy and happiness. The girl playing with soccer balls in a park is a representation of the innocence and joy of childhood. ', 'A little girl standing in the grass with two soccer balls in front of her. Soccer. A young girl playing with her toy soccer balls. Childhood innocence and joy. The girl and soccer in the grass. Innocence and joy in childhood.', 'A young girl standing on top of a grass covered field. Innocence and joy. Grass, innocence, joy. :)', '"a little girl is standing outside with a soccer ball , smiling. The child's innocence and joy radiates through the image. A feeling of childhood innocence and joy. A little girl, standing outside with a smile, playing with a soccer ball evokes a feeling of nostalgia for simpler times. "', '"The young child is standing beside two soccer balls. The young child is standing beside two soccer balls, smiling. The child's happiness and excitement over the soccer balls. The child's joy and delight over playing the game. The child's pleasure and excitement in discovering a new activity they enjoy."



112; 3533; 53491;

'A kitchen counter with a pizza ready for the oven. Pizza. Pizza. A kitchen counter with a pizza in it. Pizza. Pizza.', 'An uncooked pizza sits on the counter. - Cheesy, tempting. The pizza needs to be cooked to be delicious. The pizza needs to be cooked to be delicious. The pizza needs to be cooked in order to be tasty.', 'A uncooked pizza on a counter in a kitchen. Pizza in a kitchen. Pizza. Cheese. Pizza.', '"A homemade pizza is waiting to be cooked. Homemade. You're the one I love. A feeling of comfort, joy, and contentment. :)', '"A homemade pizza with jarred sauce and cheese .Pizza. Jarred sauce. SauceJarred cheese. CheeseHomemade pizza with jarred sauce and cheese.. Pizza. SauceHomemade pizza with jarred sauce and cheese is delicious! - Jarred pizza sauce.'



112; 2688; 473406;

'A little girl sitting on a kitchen counter next to a plate of food. A feeling of comfort and familiarity. The little girl and the food represent a feeling of contentment and familiarity, evoking a sense of nostalgia. -A feeling of comfort and familiarity. ', 'A child on a counter by a plate of fires. The smell of burning wood. A plate of burning wood evokes the image of a child playing with fire, a dangerous and thrilling experience that is associated with the smell of burning wood. ', 'A little boy sitting on a kitchen counter top next to a plate of sliced apples. Apples. Apple Pie. The boy represents innocence and the pie signifies comfort, a connection between the past and the present. Apples represent childhood memories and the connection between the present and the past, between innocence and comfort. ', 'A young child sitting on a kitchen counter next to some sliced apples. Apple Pie. Apples. A child eating and enjoying apples. Apple Pie. Apple Pie.', 'A young child is sitting on top of the kitchen counter. A feeling of joy and contentment. A memory of innocence and childhood. Joyful childhood memories. : A child, joyful and innocent, sitting atop a kitchen counter.

参考文献

- 岡崎直観、荒瀬由紀ら、*自然言語処理の基礎*、オーム社 (2022)
- 山田育矢、鈴木正敏ら、*大規模言語モデル入門*、技術評論社 (2023)
- Deza, M. M. and Deza, E., *Encyclopedia of Distances 2nd Ed.*, Springer, p49 (2012)
- Lin, T. Y., Maire, M., Belongie, S. et al., “Microsoft COCO: Common Objects in Context” *arXiv* 1405.0312 (2014)
- Mikolov, T., Chen, K., Corrado, G., and Dean, J., “Efficient Estimation of Word Representations in Vector Space”, in *ICLR* (2013)
- Naveed, J., Khan, A. U., Qiu, S., et al., “A Comprehensive Overview of Large Language Models”, *arXiv* 2307.06435 (2023)
- Radford, A., Kim, J. W., Hallacy, C. et al., “Learning Transferable Visual Models From Natural Language Supervision”, *arXiv* 2103.00020 (2021)
- Wang, J., Yang, Z., Hu, X., et al. “GIT: A Generative Image-to-text Transformer for Vision and Language”, *arXiv* 2205.14100 (2022)