マルチモーダル対照学習を応用した

五感から感情を推測、制御する手法の提案

5年B組　岡本　晃朋
指導教員　守本　寛治

１．要約

マルチモーダル空間における対照学習を応用することで、五感が感情にどのような影響を与えるかを推測し、特定の感情状態を再現することを目的としている。五感情報を離散的なデータとして扱い、表情から抽出した感情との対照学習を行うことで実現する。

キーワード
Constructive Learning　Matrix　Approximate nearest neighbor search　Multimodal

２．研究の背景と目的

私は他人と話をする時、知識に関わらず話の内容が伝わらない、理解されないことがある。この原因は私と他人が話に対して持つ感情(≒感覚)の違いにあると考えた。そこで、他人に話に対して同じ感情を持たせることができれば会話が成立するのではと考えた。感情に最も影響を与えるものは五感であるという仮説を立て、五感が得る情報と感情の関係性をマルチモーダル空間における対照学習を用いて学習・推論することで、特定環境下での感情のシミュレーション及び特定の感情を作り出す状況の算出を目的としている。

３．予備実験

仮説検証のために同学年120名に対してアンケートを行い、回答のあった20人のデータを示す。

３．１　五感と感情の関係性

感情を五感から感じたことがあるかについて二択調査を行った。Figure1 に示す通り、全ての感覚において75%以上の人が五感は感情に影響を与えていると回答した。

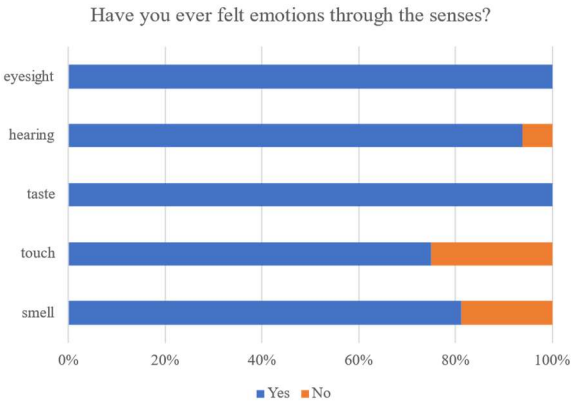

Figure 1: 感覚と感情の関連性

３．２　視覚情報と感情の関連性

視覚情報について、物体の色や形などがどのような影響を与えるかについても調査を行った。

３．２．１　物体の形状における影響

Figure 2 に示す画像についてそれぞれの形状自体が感情に影響を与えるかについて実験を行った。結果を Figure 3 に示す。

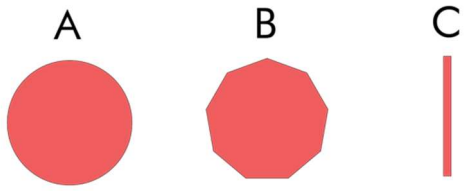

Figure 2: 形状による感情変化

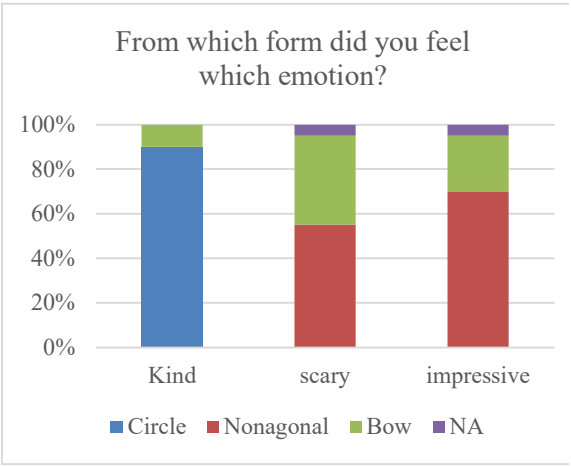

Figure 3: Percentage of feelings toward the form

結果として、90%以上の人が円について”Kind”と感じた。しかし、他の二つの形状にはあまり特徴が見られなかった。

３．２．２ 物体の色における影響

Figure 4 に示す画像について、それぞれのグラデーションに対してどのような感情を持つかについて実験を行った。

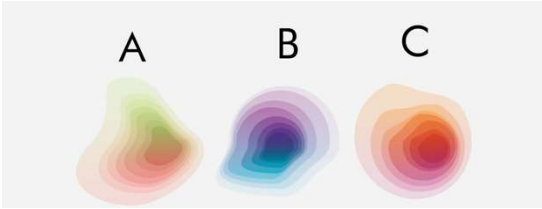

Figure 4: 色による感情変化

左から A(淡い赤-緑), B(紫-青), C(オレンジ-赤)

結果として全ての感情に 50%以上の人が同じ図形を選択したが、色からの判別は難しいと考える。

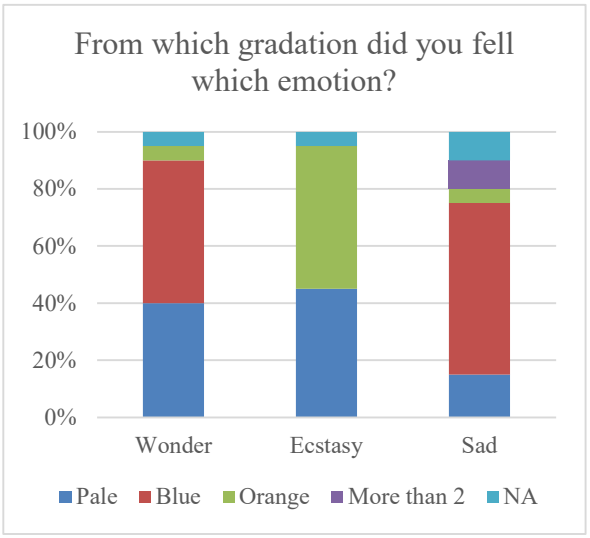

Figure 5: Percentage of feelings toward the color

４． 関連研究

感情の算出などが可能な参考研究を挙げる。

４．１ 感情極性辞書

言語情報から感情を判断するものであり、それぞれの言葉に”Positive” と ”Negative” のラベルをアノテーションすることで作成される。NLP などで利用されるが、単語からスコアを算出するため、感情の多様性や一般性の欠如が問題である。

４．２ Word2vec

単語を固定端のベクトルで表現することを目指したものであり、CBOW や skip-gram を採用したモデルである。「単語の意味は周囲の単語によって形成される」という分布仮説に基づいて構成されており、単語の類似性などの発見が可能だが、多義語や固有名詞への対応が難しい。

４．３ BERT

Transformer を利用したモデルで、単語を単位としてではなく単語の接合部や単語

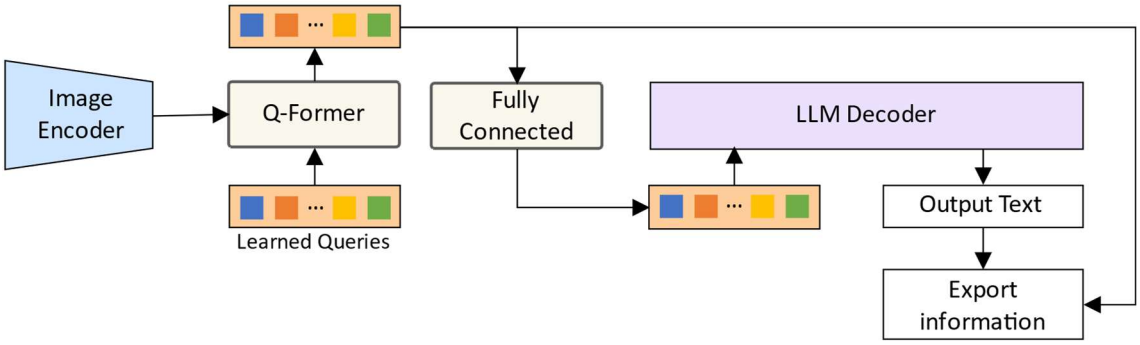

Figure 6: Vision-to-Language Generative Learning の概要
全結合層を用いて出力クエリを LLM のテキスト埋め込みと同じ次元に線形放射することで、Q-Former の解釈を LLM に伝えている。本研究では Q-Former からの出力を含めるようにし、学習に使用している。

の一部分など、文中の token を同時に処理することができる。“POSITIVE”, “NEUTRAL”, “NEGATIVE”の三種類の観点からのスコアを出すことが可能だが、そこから感情を推定するのは難しい。

５． 研究内容

予備実験３．１にて五感の情報が感情に影響を与えることがわかった。しかし、予備実験３．２から、全ての場合において実際に見えているものの形状や色彩が感情に影響を与えているとは考えにくい。これは、色や形は人が物体を認識するのに必要な要素に過ぎないからだと考えられる。そこで、五感の情報を物体の識別子として捉えることで五感の捉える２次情報の再現を試みた。

５．１　感覚刺激の再現

五感の受ける感覚刺激を再現するために視覚や聴覚などの情報を離散的なデータと見做し、識別子の作成を行なった。

５．１．１　視覚識別子の作成

５．１．１．１　Q-Former による特徴推論

視覚情報から視覚識別子を作成するために、BLIP2[7]を応用したエンコーダーを作成した。Figure 6 に示すように、Q-Former からの出力を含めるようにすることで画像特徴量を同時に扱えるようにした。結果を Table 1 に示す。

a tall wooden tower with lights on top of it in the middle of a city street at night.

Table 1: Q-Former を用いて生成したテキスト
Figure 6 にて示したテキストには画像の特徴が多く反映されていると考えられる。しかし、文章形式であるため学習に流用するのは難しい。

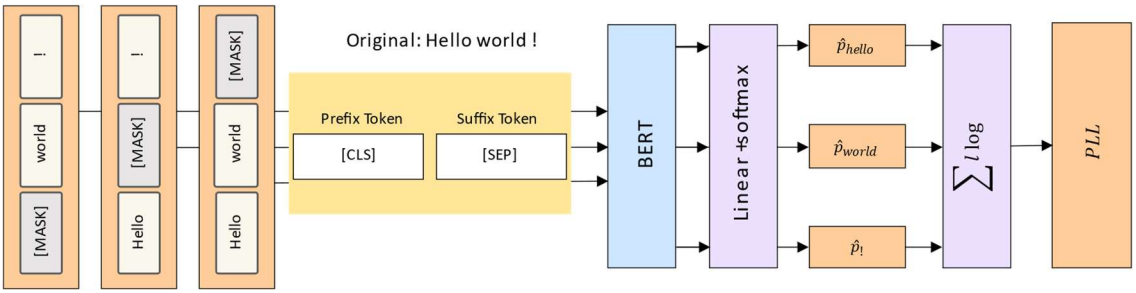

Figure 7: pseudo-log-likelihood scores によるスコアリング

各単語W_nに対して MASK をかけ、その位置にくる単語を予測し、実際の単語の予測順位から$\hat{p}$を計算し、それらの対数尤度の和を文章のスコアとする。[CLS]トークンの後には事前学習させた MLM の PLL によって線形マップを教師あり学習を行っている。

５．１．１．２　PLL による文章補完

PLL[10]を用いて Q-Former を用いて生成した文章における代名詞などを補完する。(Figure 7)

PLL では、単語の列

$$W_{\setminus t} := (\omega_1, \ldots, \omega_{t-1}, \omega_{t+1}, \ldots, \omega_{|W|})$$

に対し、単語を Mask で隠して予測した時の条件付き確率の対数尤度の和

$$\mathrm{PLL}(W) := \sum_{t=1}^{|W|} \log P_{MLM}\left(w_t | W_{\setminus t}\right)$$

をスコアとして用いる。実際の計算スコアを Table 2 に示す。

５．１．１．４　CLIP による識別子の最小単位の決定

PLL にて補完したテキスト情報から、最も意味が強く影響する最小単位へ変換する。CLIP[12]を用いて画像とテキストでの Zero-Shot Image Classification を行い、分割前と分割後の image-text similarity score を比較し、値の大きい方を使用する。この処理は文章全体に対して行い、すべての要素から原文が復号できるようにしている。

	Pseudo-log-likelihood: Evaluation Score
tower	-28.120616793632507
lights	-29.078521132469177
top	-31.276191599667072
tall	-31.34084489569068
street	-31.35562378168106
wooden	-33.037727903574705
middle	-33.079763650894165
night	-36.80495597422123

Table 2: PLL によるスコアリング

Table 1 にて生成した文章中の it に入る単語を推測している。Figure 7 で示した手法のように、it に文中にある単語を入れたスコアを計算している。この場合、tower が正解であり、結果も tower が最も高いスコアを示している。また、スコアが負の値なのは対数尤度をとっているからである。

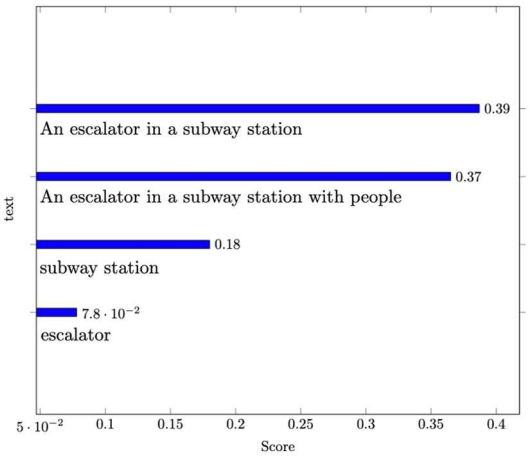

Figure 8: CLIP によるスコアリング

生成した文章を分割し、CLIP を用いてスコアリングを行った。この場合、状況を最も強く表す文章として、" An escalator in a subway station"が最も強い。

５．１．３　CLAP による聴覚識別子の作成

聴覚識別子の作成に CLAP[8]: (Contrastive Language-Audio Pretraining)を利用した。音声エンコーダーには transformer ベースの HTSAT[12]を採用し、テキストエンコーダーには BERT[13]を採用した。CLAP では損失関数

$$\frac{1}{2N}\sum_{i=1}^{N}\left(\log\frac{\exp\left(\frac{E_i^a \cdot E_j^t}{\tau}\right)}{\sum_{j=1}^{N}\exp\left(\frac{E_i^a \cdot E_j^t}{\tau}\right)} + \log\frac{\exp\left(\frac{E_i^t \cdot E_j^a}{\tau}\right)}{\sum_{j=1}^{N}\exp\left(\frac{E_i^t \cdot E_j^a}{\tau}\right)}\right)$$

E^a: audio embedding

E^t: text embedding

τ: learnable temperature parameter

が最小になるように学習している。つまり、audio 埋め込みE_p^aに対して最も近いテキストE_q^tを M までのテキスト間$E^t = \{E_1^t, \ldots, E_M^t\}$からコサイン類似度関数を用いて見つけることができる。こちらも CLIP と同様に Zero-Shot Classification を利用した。(Figure 9)

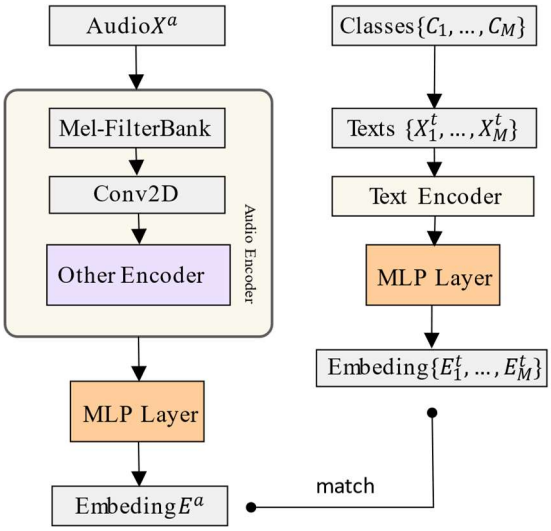

Figure 9: CLAP における Zero-Shot Classification

MLP:多層パーセプトロンを用いてE^aとE^tの同じ次元を得ている。損失関数の値が最小になるように学習を行っているので、E^aとE^tのペアでコサイン類似度が 1 になるのが全く同じ意味を持つことになる。

５．２　EmoCo による感情推定

特定状況における表情から感情を抽出する。画像をエンコードして得た特徴量ベクトルに対し Softmax(Sm)

$$\mathrm{Sm}(\mathbf{z}) = \frac{e^{\mathbf{z}_i}}{\sum_i e^{\mathbf{z}_i}}$$

SparseMax(Sp)

$$\mathrm{Sp}(\mathbf{z}) = \underset{\mathbf{p}\in\Delta^d}{\operatorname{argmin}}\|\mathbf{p}-\mathbf{z}\|^2 \mathrm{softmax}(\mathbf{z})$$

を利用して注意領域と不注意領域を生成する。EmoCo[1]では Sp は神経注意メカニズム、Sm は顔から注意情報を得るのに使用している。(Figure 10) を施している。

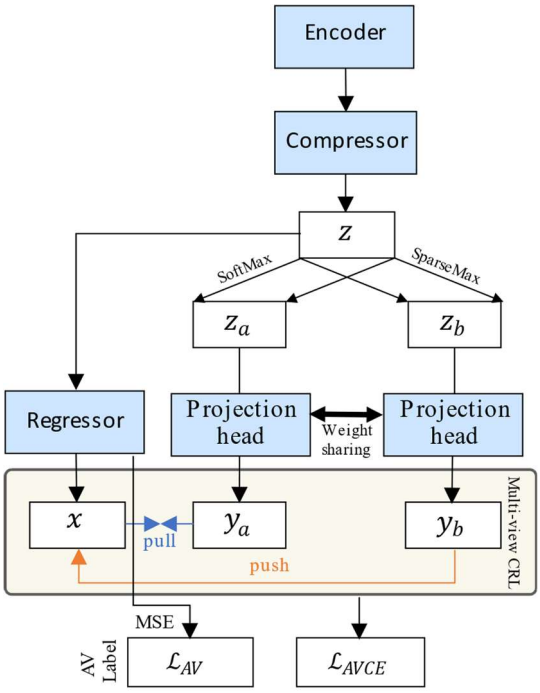

Figure 10: EmoCo による感情推定

MLP:多層パーセプトロンを用いてE^aとE^tの同じ次元を得ている。損失関数の値が最小になるように学習を行っているので、E^aとE^tのペアでコサイン類似度が 1 になるのが全く同じ意味を持つことになる。

５．２．１ SoftMax と SparseMax の比較

SoftMax は

$$\mathrm{Softmax}(\mathbf{z}) = \frac{e^{\mathbf{z}_i}}{\sum_i e^{\mathbf{z}_i}}$$

と先ほど説明したが、argmax を使用して

$$\mathrm{softmax}(\mathbf{z}) \coloneqq \max_{\mathbf{p}\in\Delta^d} \mathbf{p}^{\mathrm{T}}\mathbf{z} + \mathrm{H}^{\mathrm{S}}(\mathbf{p})$$

$$\text{where } H^s(p) = \sum_j p_j \log p_j$$

と表すことができる。この表記から、Softmax は argmax にシャノンのエントロピーが加わっているのが分かる。

同様に SparseMax は

$$\mathrm{sparsemax}(\mathbf{z}) \coloneqq \max_{\mathbf{p}\in\Delta^d} \mathbf{p}^{\mathrm{T}}\mathbf{z} + \mathrm{H}^{\mathrm{G}}(\mathbf{p})$$

$$\text{where } H^G(p) = \frac{1}{2}\sum_j p_j(1-p_j)$$

と表せる。どちらも右辺に argmax が入っているが、SoftMax では値が極大、極小出会っても出力が 0 にならないのに対して、SparseMax は$-1 \le t \le 1$以外の値をすべて 0 または 1 に切り捨てる。(Figure 11)

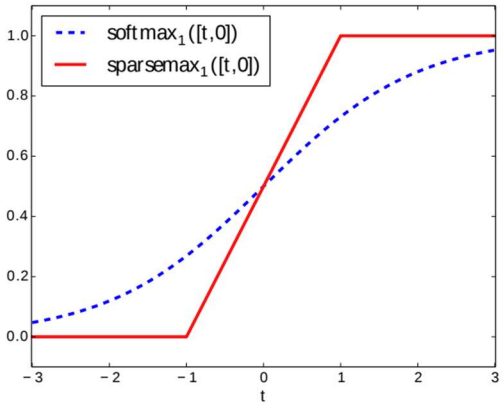

Figure 11: Softmax と Sparsemax の比較

2 次元における Softmax と Sparsemax を表す。Softmax が全体に値を与えているのに対して、sparsemax は$-1 \le t \le 1$以外での値が切り捨てられているのがわかる。また、2 次元における Softmax はシグモイド関数と同値である。(A sparse model of attention and multi-label classification. International conference on machine learning[9], p3 Figure1 より引用)

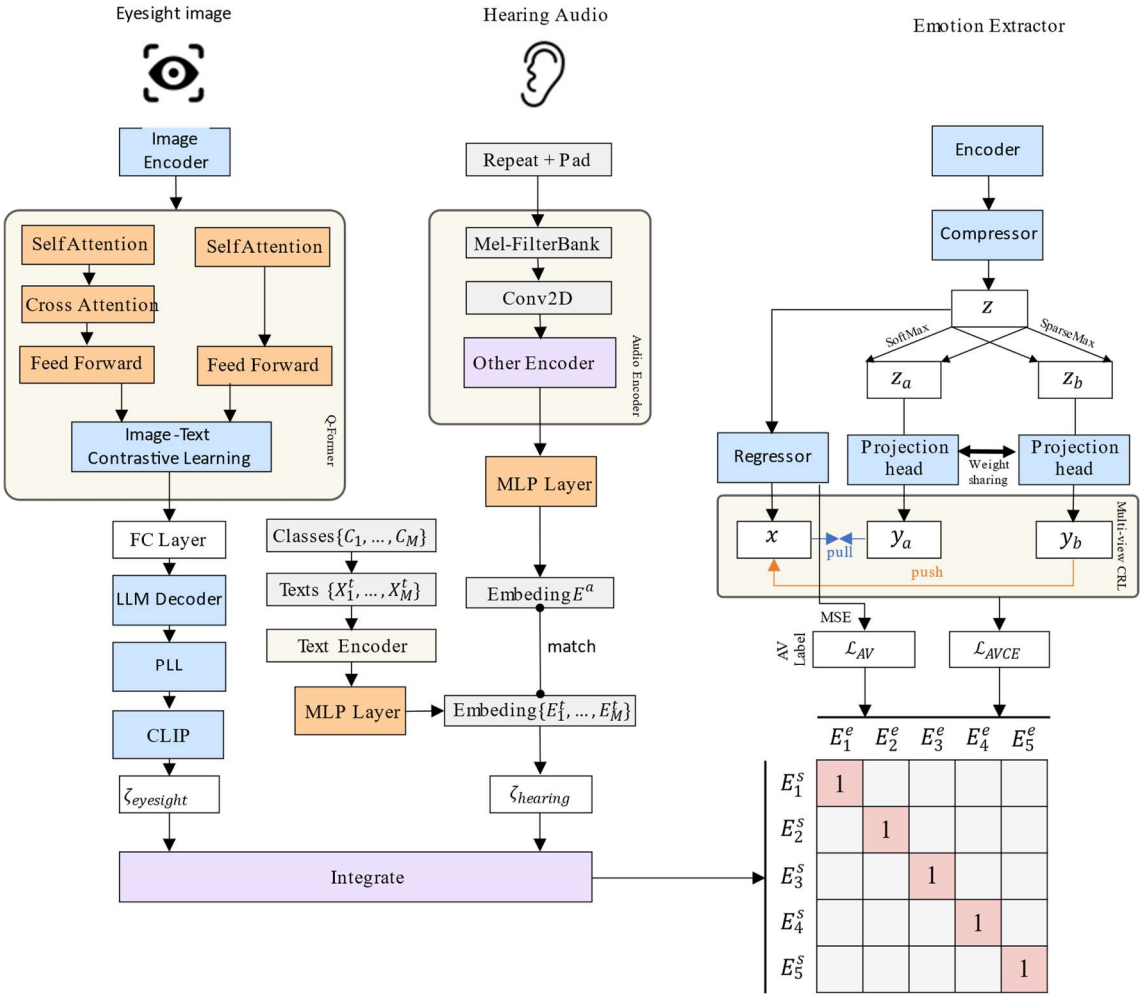

Figure 12: EmoSence-MMC のアーキテクチャ

5.1 で作成した識別子と 5.2 の EmoCo を用いて作成した感情についてコサイン類似度が 1 になるよう

に学習を行う。損失関数は $\frac{1}{2N}\sum_{i=1}^{N}(\log\frac{\exp(\frac{E_i^s \cdot E_j^e}{\tau})}{\sum_{j=1}^{N}\exp(\frac{E_i^s \cdot E_j^e}{\tau})} + \log\frac{\exp(\frac{E_i^e \cdot E_j^s}{\tau})}{\sum_{j=1}^{N}\exp(\frac{E_i^e \cdot E_j^s}{\tau})})$

５．３　モデルアーキテクチャの概要

これまで作成してきたものから、感情と感覚のマルチモーダルな対照学習を行う。Figure12 に示すように、視覚識別子と聴覚識別子の作成と EmoCo が組み込まれている。また、学習に用いるデータには画像、テキストの両方を使用しており、マルチモーダル空間での処理を行うことでより感覚刺激を学習時に再現しようと試みている。

６．問題点

本手法の問題点として、マルチモーダル間での対照学習において、データセットの自作が大きな問題となっている。利用できるデータセットがないため、現在自作しているが、規模の大きさや一貫性の問題でまだ完成に至っていない。そのため、まだ本手法の検証が完了していなく、学習方法の変更などを考えている。

７． 今後の展望

７．１　entmax の導入及び応用

5.2.1 で述べた Softmax と Sparsemax の代わりに、entmax を導入する。先ほどの表記と同じように argmax を用いて

$$\text{entmax}(\mathbf{z}) \coloneqq \max_{\mathbf{p}\in\Delta^d} \mathbf{p}^\top \mathbf{z} + H_\alpha^T(\mathbf{p})$$

$$where$$

$$H_\alpha^T \coloneqq \begin{cases} \dfrac{1}{\alpha(\alpha-1)} \sum_j \left(p_j - p_j^\alpha\right), & \text{if } \alpha \neq 0 \\ H^s(p), & \text{if } \alpha = 0 \end{cases}$$

となる。ここで、αを温度パラメータとして最適化するように学習することができれば、より実際の感覚に近い活性化関数にできると考えている。

７．２　物体による視覚識別子の追加

現在、3.2 の結果から物体の名前などから視覚識別子を作成している。これは実在論という名詞に対応した者自体が存在するという考え方と一致する。それに相反する形而上学的唯名論の観点から、物体の色や形から未知の物体などの補完を行うことで、より正確に視覚に映るものを表現することが可能だと考える。U3HS フレームワーク[4]では、既知のオブジェクトと未知を事前知識なしで区分することが可能であるため、未知に対するキャプションの生成を行えると考える。

７．３　識別子の合成

現在作成した識別子はテキストであり、他の感覚のテキストと結合しているだけの状態である。そこで、テキストの tokenize を行い、ベクトルとしてそれらを合成することを考える。五感すべての情報があるとき、ベクトルを複素数平面に拡張すると考えると、

$$\zeta = \exp\left(\frac{2\pi i}{5}\right)$$

を $0 \leq |\zeta| \leq 1$ の範囲で定義する。このとき、ζ^2やζ^3はオイラーの公式よりθが$\frac{2\pi}{5}$動いた場所に位置する。(Figure 13)

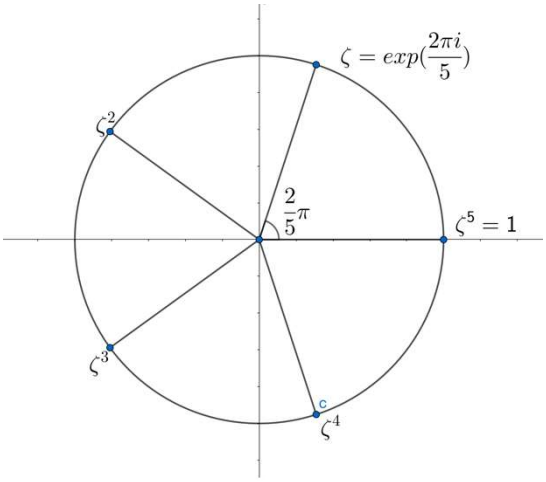

Figure 13: 単位円上のζの位置

このζベクトル上に変換したベクトルを適用することで、異なる部位から得た情報を傾きと表しながら合成することができる。(Figure 14)

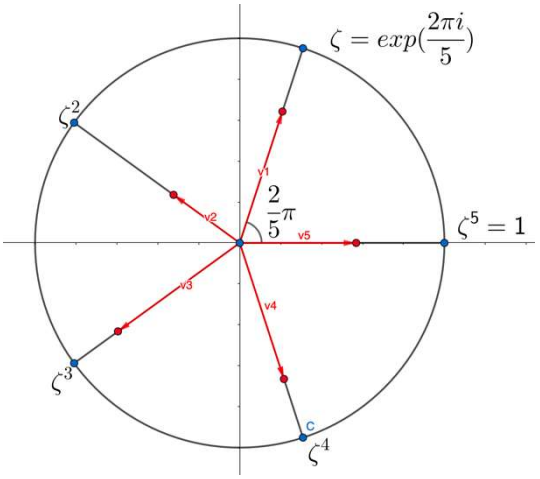

Figure 14: ベクトルをζ上に展開する

また、すべての$|\zeta|$の値が等しいとき、ζ と ζ^4, ζ^2 と ζ^3は複素共役になり、

$$\zeta + \zeta^2 + \zeta^3 + \zeta^4 + \zeta^5 = 0$$

となる。

次に、それぞれのベクトルの加法によって得られた原点Oからのベクトル$\overrightarrow{O\xi}=\vec{V}$について考える。(Figure 15)

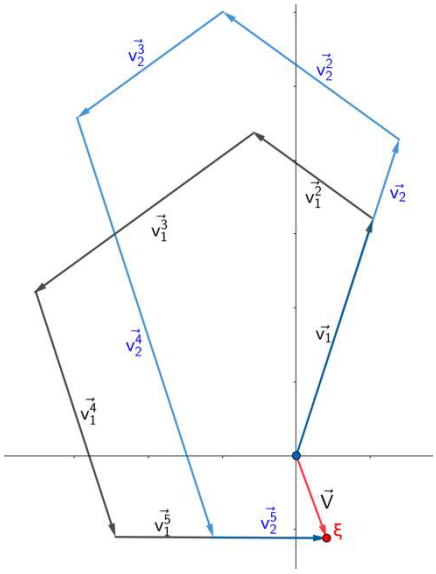

Figure 15: ベクトルの加法

Figure 15 に示すように、

$$\begin{aligned}&\overrightarrow{v_1}+\overrightarrow{v_1^2}+\overrightarrow{v_1^3}+\overrightarrow{v_1^4}+\overrightarrow{v_1^5}\\&=\overrightarrow{v_2}+\overrightarrow{v_2^2}+\overrightarrow{v_2^3}+\overrightarrow{v_2^4}+\overrightarrow{v_2^5}\\&=\vec{V}\end{aligned}$$

のように、各ベクトルの長さを調節することで、特定の感情$\overrightarrow{O\xi}$を表す感覚値を作り出すことが可能になり、感情の逆算がより簡単に実現できると考える。

７．４　他感覚の追加

現在、視覚と聴覚を用いた推論を行なっており、他三つの感覚の手法を模索中である。

７．４．１　嗅覚・味覚

これらの仕組みについて調べてみると、化学的なプロセスが多く、再現は難しいと考える。五味やヘニングの嗅覚プリズムなどを用いるとなると、匂いや味を事前に調査する必要があり、本研究には向いていない。そこで、これらの情報を視覚から抽出できないかと考えている。

７．４．２　触覚

触覚は嗅覚・味覚とは違い、明確に分かれているため、比較的再現が簡単だと考える。温覚・冷覚などは外気温などから流用し、触覚・痛覚を視覚から抽出できないかと考えている

７．５　Faiss を用いた個人差の調整

近傍探索ライブラリである Faiss[6]を用いてそれぞれの結果に対し調整を行う。事前学習として個人の index を作成し、後ほど重みつき結合を行うことで結果を個人の感覚へと寄せることが可能だと考える。

８．　参考文献

1. Daeha Kim and Byung Cheol Song. "Emotion-aware multi-view contrastive learning for4 facial emotion recognition." In Lecture Notes in Computer Science, Lecture notes in computer science, pages 178-195. Springer Nature Switzerland, Cham, 2022, doi: 10.1007/978-3-031-19778-9_11.
2. Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor, "AffectNet: A New Database for Facial Expression, Valence, and Arousal Computation in the Wild," IEEE Transactions on Affective Computing, 2017.
3. D. Kollias, S. Zafeiriou: "Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface." In: 30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019. https://bmvc2019.org/wp-content/uploads/papers/0399-paper.pdf.
4. Stefano Gasperini, Alvaro Marcos-Ramiro, Michael Schmidt, Nassir Navab, Benjamin Busam, Federico Tombari: "Segmenting Known

Objects and Unseen Unknowns without Prior Knowledge."
https://arxiv.org/abs/2209.05407, 2022.

5. "A Neural Algorithm of Artistic Style" by Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge.
https://arxiv.org/abs/1508.06576, 2015.
6. The Faiss library, Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy , Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, Hervé Jégou.
https://arxiv.org/abs/2401.08281, 2024
7. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi.
https://arxiv.org/abs/2301.12597, 2023.
8. CLAP: Learning Audio Concepts From Natural Language Supervision Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, Huaming Wang. https://arxiv.org/abs/2206.04769, 2022.
9. Martins, Andre, and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. International conference on machine learning. PMLR, 2016.
https://arxiv.org/abs/1602.02068
10. Masked Language Model Scoring, Julian Salazar, Davis Liang, Toan Q. Nguyen, Katrin Kirchhoff, ACL 2020.,
https://arxiv.org/abs/1910.14659
11. Learning Transferable Visual Models From Natural Language Supervision, Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever,
https://arxiv.org/abs/2103.00020, 2021
12. HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection. Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, Shlomo Dubnov.
https://arxiv.org/abs/2202.00874, 2022
13. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova.
https://arxiv.org/abs/1810.04805, 2019

９． 謝辞

本研究において仮説検証のアンケートに答えて頂いた 20 名の方、大会の応募などを手伝っていただいた守本寛治先生に感謝をこの場で申し上げます。