Context-Aware Meta-Learning

Christopher Fifty[1], **Dennis Duan**[1,2], **Ronald G. Junkins**[1],
Ehsan Amid[3], **Jure Leskovec**[1], **Christopher Ré**[1], **Sebastian Thrun**[1]
[1]Stanford University, [2]Google, [3]Google DeepMind
fifty@cs.stanford.com

Abstract

Large Language Models like ChatGPT demonstrate a remarkable capacity to learn new concepts during inference without any fine-tuning. However, visual models trained to detect new objects during inference have been unable to replicate this ability, and instead either perform poorly or require meta-training and/or fine-tuning on similar objects. In this work, we propose a meta-learning algorithm that emulates Large Language Models by learning new visual concepts during inference without fine-tuning. Our approach leverages a frozen pre-trained feature extractor, and analogous to in-context learning, recasts meta-learning as sequence modeling over datapoints with known labels and a test datapoint with an unknown label. On 8 out of 11 meta-learning benchmarks, our approach—without meta-training or fine-tuning—exceeds or matches the state-of-the-art algorithm, P>M>F, which is meta-trained on these benchmarks.

1 Introduction

Meta-learning algorithms for image classification aim to classify a set of unlabeled images from only several labeled examples. The labeled examples are termed the *support set* and the set of unknown images is called the *query set*. In an n-way-k-shot meta-learning paradigm, the support set spans n different classes, each class contains k labeled images, and the meta-learning algorithm predicts the class of each unlabled image in the query set from the n classes in the support set.

Nearly all meta-learning algorithms ascribe to a common pattern of pre-training, meta-training, and/or fine-tuning [Hu et al., 2022]. Pre-training initializes the meta-learner's feature extractor with a pre-trained vision model; meta-training trains the model's parameters to learn how to classify new visual concepts during inference by training the model on a series of n-way, k-shot classification tasks; and fine-tuning updates the model's parameters on the support set at inference.

While meta-training excels in learning new classes during inference that are similar to those seen during meta-training, it often fails to generalize to new classification paradigms. For example, models meta-trained on coarse-grained object detection often fail to generalize to fine-grained image classification. Fine-tuning on the support set during inference can rescue an otherwise poor performing model; however, training a model during inference is often impractical and prohibitive of many real-time applications. In this regard, visual meta-learning algorithms lag behind recent advancements in natural language where Large Language Models (LLMs) exhibit a remarkable capacity to learn new concepts during inference without fine-tuning [Brown et al., 2020].

In this work, we develop a meta-learning algorithm that emulates LLMs by learning new visual concepts during inference without fine-tuning. Drawing inspiration from in-context learning in LLMs, we reformulate n-way-k-shot image classification as sequence modeling over the support set and an unknown query image. Due to its capacity to learn visual information "in-context", we term our approach *Context-Aware Meta-Learning* (CAML).

Workshop on Distribution Shifts, 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

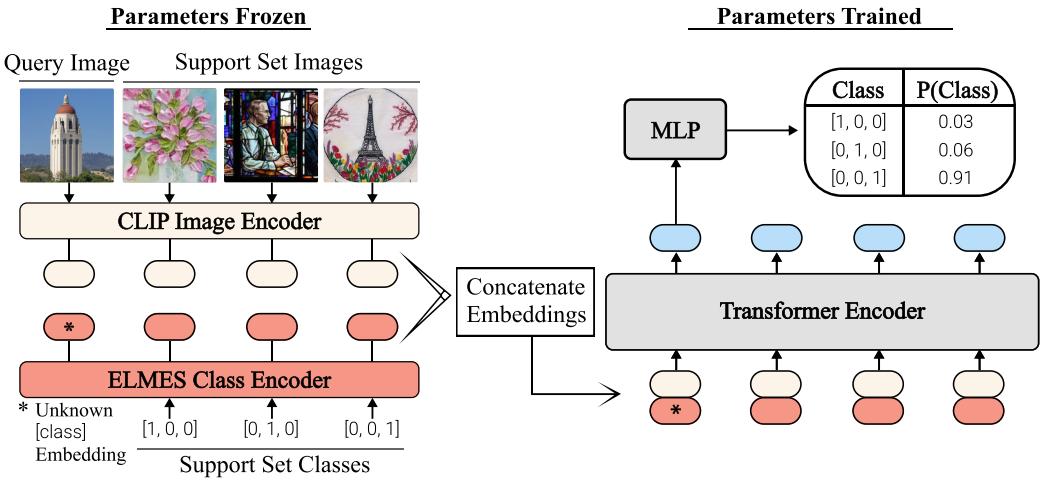


Figure 1: Overview of CAML. Query and support set images are encoded with a CLIP feature extractor and then concatenated with their corresponding ELMES label embeddings. We feed the resulting sequence of concatenated vectors into a Transformer encoder and extract the transformed query vector from the output sequence to predict its class.

2 Approach

We adapt the ideas underpinning in-context learning in LLMs—namely learning to classify a query from a context of support set demonstrations in a single forward pass—to image classification. A similar concept has recently been explored by Fifty et al. [2023] for few-shot molecular property prediction. Dissimilar from this work, we avoid meta-training and instead focus on universal image classification: learning to detect new visual classes during inference without meta-training on related classes or fine-tuning on the support set.

Architecture. CAML consists of three components: (1) a frozen CLIP image encoder, (2) a fixed ELMES class encoder, and (3) a Transformer encoder sequence model. CAML first encodes query and support set images using a frozen CLIP feature extractor. Crucially, the CLIP embedding space distills images into low-dimensional representations so that images with similar visual characteristics and semantic meanings have similar embeddings. We encode the classes of the support set with an ELMES class encoder. In Appendix A, we prove that an ELMES encoding of mutually exclusive labels allows the Transformer encoder sequence model to maximally identify classes within the support set. As the class of the query is unknown, it uses a special learnable "unknown token" embedding.

The core idea underpinning CAML is to cast meta-learning as sequence modeling over the support set and query points. We instantiate the sequence model as a Transformer encoder, and during large-scale pre-training, train the model to predict the class of the query from an input sequence composed of the support set and query embedded vectors. Specifically, the input to the Transformer encoder is a sequence of support set and query vectors embedded in the joint image-label embedding space. From the output sequence of the Transformer encoder, we select the element at the same position as the query in the input sequence, and pass this vector through a shallow MLP to predict the label of the query. A visual depiction of CAML is shown in Figure 1.

Large-Scale Pre-Training. As our focus is universal meta-learning—and CAML may encounter any new visual concept during inference—we pre-train CAML's Transformer encoder on few-shot image classification tasks from ImageNet-1k [Deng et al., 2009], Fungi [Schroeder and Cui, 2018], MSCOCO [Lin et al., 2014], and WikiArt [Saleh and Elgammal, 2015]. We chose these datasets because they span generic object recognition (ImageNet-1k, MSCOCO), fine-grained image classification (Fungi), and unnatural image classification (WikiArt). To avoid distorting the CLIP embedding space, we freeze the CLIP feature extractor and only update the Transformer encoder during pretraining. Similarly, since an ELMES minimizes the entropy of detecting classes within the support set, the label encoder is also frozen. In the context of pre-training, meta-training, and

Table 1: **MiniImageNet & CIFAR-fs** mean accuracy and standard error across 10,000 test epochs.

Method (Backbone)	CIFAR-fs		MiniImageNet	
	5w-1s	5w-5s	5w-1s	5w-5s
In-Domain [Meta-training]				
P>M>F (ViT-base) Hu et al. [2022]	84.3	92.2	95.3	98.4
Universal Meta-Learning; No Meta-Training or Finetuning				
MetaQDA (ViT-base) [Zhang et al., 2021]	60.4±.2	83.2±.1	88.2±.2	97.4±.0
CAML (ViT-base)	**70.8**±.2	**85.5**±.1	**96.2**±.1	**98.6**±.0

Table 2: **Pascal & Paintings** mean accuracy and standard error across 10,000 test epochs.

Method (Backbone)	Pascal + Paintings		Paintings		Pascal	
	5w-1s	5w-5s	5w-1s	5w-5s	5w-1s	5w-5s
In-Domain [Meta-Training]						
P>M>F (ViT-base)	60.7	74.4	53.2	65.8	72.2	84.4
Universal Meta-Learning						
MetaQDA (ViT-base)	53.8±.2	74.1±.1	49.4±.2	**66.6**±.1	73.5±.2	85.2±.2
CAML (ViT-base)	**63.8**±.2	**78.3**±.1	**51.1**±.2	65.2±.1	**82.6**±.2	**89.7**±.1

fine-tuning, CAML only requires pre-training and avoids meta-training on the train/validation splits of meta-learning benchmarks or fine-tuning on the support set during inference.

3 Experiments

To quantify universal image classification performance, we evaluate a diverse set of 11 meta-learning benchmarks divided across 4 different categories:

1. Generic Object Recognition: mini-ImageNet [Vinyals et al., 2016], tiered-ImageNet [Ren et al., 2018], CIFAR-fs [Bertinetto et al., 2018], and Pascal VOC [Everingham et al.]
2. Fine-Grained Image Classification: CUB [Wah et al., 2011], Aircraft [Maji et al., 2013], meta-iNat [Wertheimer and Hariharan, 2019], and tiered meta-iNat [Wertheimer and Hariharan, 2019]
3. Unnatural Image Classification: ChestX [Guo et al., 2020] and Paintings [Crowley and Zisserman, 2015]
4. Inter-Domain Image Classification: Pascal+Paintings [Everingham et al., Crowley and Zisserman, 2015].

Generic object recognition, fine-grained image classification, and unnatural image classification are standard benchmarking tasks in meta-learning literature [Chen et al., 2020, Hu et al., 2022, Wertheimer et al., 2020, Guo et al., 2020]. Beyond this, we compose a challenging new *inter-domain* category by combining Pascal VOC with Paintings so that each class is composed of both natural images and paintings. This allows us to evaluate the ability of meta-learning algorithms to generalize across domains within the same class. For example, the support image for the class "tower" may be Van Gogh's *The Starry Night*, while the query may be a picture of the Eiffel Tower. Humans have the ability to generalize visual concepts between such domains; however, meta-learning algorithms struggle with this formulation.

3.1 Baselines

We evaluate CAML and MetaQDA [Zhang et al., 2021] in a universal meta-learning setting by pre-training them with the same CLIP feature extractor over the same image corpus as CAML and similarly freezing their weights at inference time. We select MetaQDA as this algorithm benefits from pre-trained feature extractors and was designed for cross-domain meta-learning [Zhang et al., 2021, Hu et al., 2022].

Table 3: **meta-iNat & tiered meta-iNat & ChestX** mean accuracy and standard error across 10,000 test epochs.

Method (Backbone)	meta-iNat		tiered meta-iNat		ChestX	
	5w-1s	5w-5s	5w-1s	5w-5s	5w-1s	5w-5s
In-Domain [Meta-Training]						
P>M>F (ViT-base)	91.2	96.1	74.8	89.9	27.0	32.1
Universal Meta-Learning						
MetaQDA (ViT-base)	86.3±.2	95.9±.1	76.0±.2	**92.4±.1**	**22.6±.1**	**27.0±.1**
CAML (ViT-base)	**91.2±.2**	**96.3±.1**	**81.9±.2**	91.6±.1	21.5±.1	22.2±.1

Table 4: **CUB & tiered-ImageNet & Aircraft** mean accuracy and standard error across 10,000 test epochs.

Method (Backbone)	CUB		tiered-ImageNet		Aircraft	
	5w-1s	5w-5s	5w-1s	5w-5s	5w-1s	5w-5s
In-Domain [Meta-Training]						
P>M>F (ViT-base)	92.3	97.0	93.5	97.3	79.8	89.3
Universal Meta-Learning						
MetaQDA (ViT-base)	88.3±.2	**97.4±.1**	89.4±.2	97.0±.1	**63.6±.3**	**83.0±.2**
CAML (ViT-base)	**91.8±.2**	**97.1±.1**	**95.4±.1**	**98.1±.1**	**63.3±.3**	79.1±.2

To assess the performance gap between universal meta-learning and the typical meta-training approach, we also benchmark the performance of the current state-of-the-art meta-learning algorithm, P>M>F [Hu et al., 2022], which is meta-trained on each dataset. While previous cross-domain approaches often involve fine-tuning on the support set at inference time, we forgo this step as fine-tuning is incompatible with universal meta-learning and developing real-time meta-learning applications.

When pre-training all models in the universal meta-learning setting, we set the learning rate to a fixed 1×10^{-5} and do not perform any hyperparameter tuning in order to match the practices used by P>M>F. We use early stopping with a window size of 10 epochs during pre-training and the code release of Hu et al. [2022] to benchmark P>M>F with the training settings and hyperparameters described in their work.

3.2 Results

Our findings are summarized in Table 1, Table 2, Table 3, and Table 4 and indicate that CAML significantly outperforms MetaQDA on 15 of 22 evaluation settings. For 5 of the remaining 7 evaluation settings, CAML matches—or nearly matches—MetaQDA. Remarkably, CAML also performs competitively with P>M>F on 8 out of 11 meta-learning benchmarks, even though P>M>F meta-trains on the training set of each benchmark. This result suggests that the amount of new visual information learned during inference when using only foundational models and novel meta-learning techniques without fine-tuning is comparable to the amount learned when directly meta-training on in-domain data. This capacity may unlock new applications in the visual space, just as the emergence of in-context learning in LLMs has enabled many new applications in natural language.

4 Conclusion

In this work, we develop CAML, a meta-learning algorithm that emulates in-context learning in LLMs by learning new visual concepts during inference without fine-tuning. Our empirical findings show that CAML—without meta-training or fine-tuning—exceeds or matches the performance of the current state-of-the-art meta-learning algorithm on 8 out of 11 benchmarks. This result indicates visual meta-learning models are ready for deployment in a manner similar to LLMs, and we hope this work recalibrates our sense of limitations for the universal meta-learning paradigm.

References

Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification, 2020.

Elliot J Crowley and Andrew Zisserman. In search of art. In *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I 13*, pages 54–70. Springer, 2015.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Matthew Fickus, John Jasper, Emily J King, and Dustin G Mixon. Equiangular tight frames that contain regular simplices. *Linear Algebra and its applications*, 555:98–138, 2018.

Christopher Fifty, Jure Leskovec, and Sebastian Thrun. In context learning for few-shot molecular property prediction. 2023.

Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 124–141. Springer, 2020.

Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M. Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015.

Brigit Schroeder and Yin Cui. FGVCx fungi classification challenge 2018. `github.com/visipedia/fgvcx_fungi_comp`, 2018.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Manfred K Warmuth and Dima Kuzmin. Randomized online pca algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, 9(Oct):2287–2320, 2008.

Lloyd Welch. Lower bounds on the maximum cross correlation of signals (corresp.). *IEEE Transactions on Information theory*, 20(3):397–399, 1974.

Davis Wertheimer and Bharath Hariharan. Few-shot learning with localization in realistic settings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6558–6567, 2019.

Davis Wertheimer, Luming Tang, and Bharath Hariharan. Few-shot classification with feature map reconstruction networks, 2020.

Xueting Zhang, Debin Meng, Henry Gouk, and Timothy Hospedales. Shallow bayesian meta learning for real-world few-shot recognition, 2021.

Appendix

A Theoretical Analysis

In this section, we explore the symmetries inherent in CAML. These symmetries allow us to formulate the problem of learning support set class representations as an entropy minimization problem with a closed-form solution. We prove that this solution is an ELMES. Later, we show it maintains permutation invariance, a vital property of meta-learning algorithms that conveys consistent predictions irrespective of the ordering of elements within the sequence.

A.1 Equiangular Tight Frames

Papyan et al. [2020] coin the term Simplex Equianguar Tight Frame to describe a set of vectors $\{\phi_j\}_{j=1}^d$ such that the minimum angle between any two pairs of vectors is maximized and all vectors have equal norm. Formally,

Definition 1. *Let $\mathbb{R}^d$ be a $d-$dimensional inner product space over $\mathbb{R}$ with the Euclidean inner product. A **Simplex ETF** is a set of d vectors $\{\phi_j\}_{j=1}^d$, $\phi_j \in \mathbb{R}^d$, specified by the columns of $\sqrt{\frac{d}{d-1}}(I_d - \frac{1}{d}\mathbb{1}\mathbb{1}^T)$*

where $I_d \in \mathbb{R}^{d\times d}$ is the identity matrix and $\mathbb{1} \in \mathbb{R}^{d\times 1}$ is the ones vector. Somewhat contradictory, a Simplex Equiangular Tight Frame is not an Equiangular Tight Frame [Welch, 1974] as this set of vectors does not form a tight frame in $\mathbb{R}^d$.

Definition 2. *Let $\mathbb{R}$ be a $d-$dimensional space over $\mathbb{R}$ with the Euclidean inner product. An **Equiangular Tight Frame (ETF)** is a set of non-zero, equal norm vectors $\{\phi_j\}_{j=1}^n$, $n \geq d$, that achieves the Welch lower bound:*

$$\max_{j\neq j'} \frac{|\langle \phi_j \ , \ \phi_{j'}\rangle|}{\|\phi_j\|\|\phi_{j'}\|} = \sqrt{\frac{n-d}{d(n-1)}}$$

It is well-known that a set of non-zero equal-norm vectors satisfies the Welch lower bound if and only if that set of vectors is equiangular and also a tight frame for $\mathbb{R}^d$ [Fickus et al., 2018].

Definition 3. *A set of non-zero, equal norm vectors $\{\phi_j\}_{j=1}^n$ is **equiangular** if $\forall j \neq j'$, $|\langle \phi_j \ , \ \phi_{j'}\rangle| = c$ for some $c \in \mathbb{R}$, $c > 0$.*

Definition 4. *$\{\phi_j\}_{j=1}^n$ is a **tight frame** for $\mathbb{R}^d$ if, $\forall v \in \mathbb{R}^d$, $\exists A > 0$ such that $A\|v\|^2 = \sum_{j=1}^n |\langle \phi_j \ , \ v\rangle|^2$.*

Remark 1. *A Simplex Equiangular Tight Frame is not a tight frame.*

Proof. Observe that for any finite d, for $\{\phi_j\}_{j=1}^d$ equal to the columns of $\sqrt{\frac{d}{d-1}}(I_d - \frac{1}{d}\mathbb{1}\mathbb{1}^T)$, it is the case that $\sum_{j=1}^{d-1} \phi_j = -1 * \phi_d$. So $\{\phi_j\}_{j=1}^n$ do not span $\mathbb{R}^d$, and therefore, cannot be a tight frame. □

Similarly, a Simplex ETF is not a $d-$simplex.

Remark 2. *A Simplex Equiangular Tight Frame is not a simplex.*

Proof. A simplex in $\mathbb{R}^n$ requires $n+1$ points. □

To align terminology with properties, we generalize a Simplex ETF to an ELMES in **??**: a set of d vectors in a $(d+k)$-dimensional ambient space with $k \geq 0$. We define an Equal Length and Maximally Equiangular Set (ELMES):

Definition 5. *An **Equal Length and Maximally Equiangular Set (ELMES)** is a set of non-zero vectors $\{\phi_j\}_{j=1}^d$, $\phi_j \in \mathbb{R}^{d+k}$ for some $k \geq 0$, such that $\forall j \neq j'$, $\|\phi_j\| = \|\phi_{j'}\|$ and $\langle \phi_j \ , \ \phi_{j'}\rangle = \frac{-1}{d-1}$. Simply, all vectors in this set are equal length and maximally equiangular.*

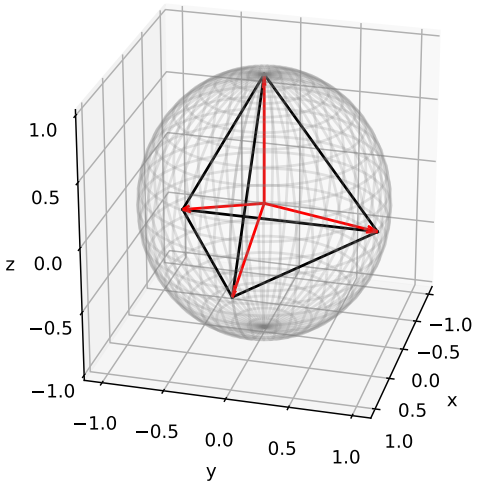


Figure 2: A visualization of a $d = 4$ ELMES in $\mathbb{R}^3$. Observe the endpoints of the vectors of an ELMES lie on the vertices of a centered regular tetrahedron.

Observe that a regular simplex is a special type of ETF in which the number of vectors in the set is one more than the dimension of the space that they span [Fickus et al., 2018]. Building on this observation, an intuitive view of ELMES is a regular $d-$simplex immersed in $\mathbb{R}^{d+k}$.

Remark 3. *Consider a centered d-dimensional regular simplex with vertices* $\{\phi_j\}_{j=1}^{d+1}$, $\phi_j \in \mathbb{R}^{d+1}$. *Let* $\imath_{can}$ *be the canonical inclusion map:* $\mathbb{R}^d \to \mathbb{R}^{d+1}$, $\imath_{can}(x_1, x_2, ..., x_d) = (x_1, x_2, ..., x_d, 0_{d+1})$, *then* $\{\imath_{can}(\phi_j)\}_{j=1}^{d+1}$ *is an ELMES.*

Proof. The two criteria of an ELMES are maximally equiangular and equal length. As all vertices of a centered regular $d-$simplex are equal length from the origin, $\{\phi_j\}_{j=1}^{d+1}$ are equal length and therefore $\{\imath_{can}(\phi_j)\}_{j=1}^{d+1}$ must also have equal length.

Similarly, from Lemma 10 of Papyan et al. [2020], we know the cosine of the angle between any two vectors in a $(d+1)-$dimensional ELMES is $\frac{-1}{d}$. It is known that for a $d-$dimensional regular simplex in $\mathbb{R}^d$ centered at the origin, the angle subtended by any two verticies through the origin is $\cos(\theta) = \frac{-1}{d}$. Immersing $\{\phi_j\}_{j=1}^{d+1}$, $\phi_j \in \mathbb{R}^d$, into $\mathbb{R}^{d+1}$ via the canonical inclusion operator $\imath_{can}$ does not change the pairwise angle between vectors in this set: $\langle \phi_j \ , \ \phi_{j'} \rangle = \langle \imath_{can}(\phi_j) \ , \ \imath_{can}(\phi_{j'}) \rangle$. As $\{\imath_{can}(\phi_j)\}_{j=1}^{d+1}$ are equal length and maximally equiangular, it forms an ELMES. □

We now show that an ELMES immersed in a higher dimension remains an ELMES. Taken with Remark 3, we can view a high-dimensional ELMES in $\mathbb{R}^d$ composed of $n+1$ vectors $\{\phi_j\}_{j=1}^{n+1}$, $d >> n+1$, as simply a $n-$simplex immersed in $\mathbb{R}^d$ via the canonical inclusion operator.

Lemma 1. *Let* $\imath_{can} : \mathbb{R}^d \to \mathbb{R}^{d+k}$. *If* $\{\phi_j\}_{j=1}^{n}$ *is an ELMES , then* $\{\imath_{can}(\phi_j)\}_{j=1}^{d}$ *is an ELMES.*

Proof. This reduces to proving that the maximum angle between a set of d equiangular points in $\mathbb{R}^d$ is the maximum angle between a set of d equiangular points in $\mathbb{R}^{d+k}$. Let $\{\phi_j\}_{j=1}^{d}$ be an ELMES such that $\phi_j \in \mathbb{R}^d$ and $\{\psi_j\}_{j=1}^{d}$ be an ELMES such that $\psi_j \in \mathbb{R}^{d+k}$. Then $\{\psi_j\}_{j=1}^{d}$ lie in a $d-$dimensional subspace of $\mathbb{R}^{d+k}$: $\exists \gamma_1, ..., \gamma_d$ and basis vectors $e_1, ..., e_d$ such that $\forall \psi_j \in \{\psi_j\}_{j=1}^{d}$, $\psi_j = \sum_{i=1}^{d} \gamma_i e_i$. Therefore, $\forall j \neq j'$, $\langle \psi_j \ , \ \psi_{j'} \rangle \leq \langle \phi_j \ , \ \phi_{j'} \rangle$ as $\{\phi_j\}_{j=1}^{d}$ are an ELMES for $\mathbb{R}^d$. □

A.2 ELMES Rotational Symmetry

There are infinitely many ELMES by rotating one such set of vectors about the origin.

Remark 4. *Let $\{\phi_j\}_{j=1}^d$ be an ELMES in $\mathbb{R}^{d+k}$ for some $k \geq 0$. Let $o : \mathbb{R}^{d+k} \to \mathbb{R}^{d+k}$ be an operator from the special orthogonal group $SO(d+k)$. Then $\{o(\phi_j)\}_{j=1}^d$ is also an ELMES .*

Proof. Length is preserved as operations in $SO(d+k)$ have determinant 1 and angles are similarly preserved as operations in $SO(d+k)$ are unitary (i.e. preserving inner product). □

A.3 A Set of Orthonormal Basis Vectors Is Not an ELMES

A final remark relates to the common misconception that a set of orthonormal basis vectors $\{\psi_j\}_{j=1}^d$ is an ELMES. While $\{\psi_j\}_{j=1}^d$ is an ETF in $\mathbb{R}^d$ since this set realizes the Welch lower-bound in Definition 2, these vectors are not maximally equiangular: $\langle \psi_j \ , \ \psi_{j'} \rangle = 0 > \frac{-1}{d-1}$.

A.4 Label Symmetry

Symmetry in the assignment of support classes to numeric labels is an important property of meta-learning algorithms. For example, if we have the support set classes {tower, bear, tree}, the mapping of {bear -> 1, tower -> 2, tree -> 3 } should produce the same prediction for a query point as a different mapping {bear -> 2, tower -> 3, tree -> 1}. To explore this symmetry, we examine how class embeddings are being used by the model.

From our formulation in Section 2, we represent a demonstration vector as a concatenation of an image embedding ρ and a label embedding ϕ: $[\rho \mid \phi]$. This vector is directly fed into the self-attention mechanism, where we matrix multiply with key, query, and value self-attention heads. Taking only one of these matrices for simplicity with head-dimension k:

$$[\rho \mid \phi] \left[\begin{array}{ccc} \Gamma_1 & ... & \Gamma_k \\ \hline \psi_1 & ... & \psi_k \end{array}\right] = [\langle \rho \ , \ \Gamma_1 \rangle \quad ... \quad \langle \rho \ , \ \Gamma_k \rangle] + [\langle \phi \ , \ \psi_1 \rangle \quad ... \quad \langle \phi \ , \ \psi_k \rangle] \tag{1}$$

The output of this transformation will be the sum of two vectors: one composed of the inner products between the image embedding and the learnable $\{\Gamma_i\}_{i=1}^k$s and the other composed of the class embedding and the learnable $\{\psi_i\}_{i=1}^k$.

We postulate a capacity to distinguish among the classes of demonstration vectors is necessary for the model to predict the class of the query vector. Conversely, if a meta-learning algorithm predicts among d classes, and all classes maintain the same embedding $\phi_j = \phi_i \ \forall i \in \{1, ..., d\}$, the model will be unable to identify the class of the query vector as all demonstration vectors appear to have the same class identity. Such an embedding would maximize the Shannon entropy for any learnable ψ_i

$$H_i(X) := -\sum_{x \in \mathcal{X}} p_i(x) \ln(p_i(x))$$

where we define $\mathcal{X} = \{1, 2, ..., d\}$ to be the different classes, X to be a random variable which takes on values in $\mathcal{X}$, and $p_i(X = j) = \frac{e^{\langle \psi_i \ , \ \phi_j \rangle}}{\sum_{\ell \in \mathcal{X}} e^{\langle \psi_i \ , \ \phi_\ell \rangle}}$ as the softmax probability of class j given that ψ_i is learned to detect class i (i.e. maximize $p_i(X = i)$ and minimize $H_i(X)$).

Contrary to the above example, we assume a capacity to learn a ψ_i that maximally detects a given class j will be beneficial to minimizing the loss for meta-learning paradigms. As we use the softmax of the inner product to determine class probabilities, maximizing $p_i(X = j)$ is equivalent to minimizing $p_i(X = \ell)$ for all $\ell \neq j$.

By symmetry in the assignment of class embeddings to support classes, we can assume that the number of ψ_i learned to detect class i is similar to the number of ψ_j learned to detect class j for all pairs (i, j). Then $p_i(X = i)$ for all $1 \leq i \leq d$ is jointly maximized $\iff$ the d-class embeddings $\{\phi_j\}_{j=1}^d$ is an ELMES . Before we prove this result, we leverage symmetry in the assignment of labels to classes to make the following assumptions:

Assumption 1. *Suppose $\{\psi_i\}_{i=1}^k$ are learnable class detectors of unit norm with at least one ψ_i detecting each class $1 \leq i \leq d$. The probability $p_j(X = j) = p_i(X = i)$ for $1 \leq i, j \leq d$.*

Justification of Assumption 1. This property is implied by symmetry in the assignment of class embeddings to support classes. As the assignment is arbitrary, all learnable ψ_i class detectors should have equal probability of detecting their respective class. □

Assumption 2. *Suppose $\{\psi_i\}_{i=1}^k$ are learnable class detectors of unit norm with at least one ψ_i detecting each class $1 \leq i \leq d$. Define $p_i(X = i) \backslash \{\phi_l\}_{l=(m+1)}^d$ as the probability of ψ_i detecting ϕ_i from the set of vectors $\{\phi_j\}_{j=1}^m$, $m < d$. Then the probability $p_j(X = j) \backslash \{\phi_l\}_{l=(m+1)}^d = p_i(X = i) \backslash \{\phi_l\}_{l=(m+1)}^d$ for $1 \leq i, j \leq m$ and $m \geq 2$.*

Justification of Assumption 2. Informally, this property states that, for any m-subset of classes $\{\phi_j\}_{j=1}^m$, the probability of ψ_j detecting class j is equal to the probability of ψ_i detecting class i. This is again implied by symmetry in the assignment of class embeddings to support classes as meta-learning algorithms may predict among a subset of m classes in the support set rather than the maximum number of classes d. □

Assumption 3. *When $\psi_i = \frac{\phi_i}{\|\phi_i\|}$, $p_i(X = i)$ is maximized.*

Justification of Assumption 3. Recall in $\mathbb{R}^d$, $\langle \psi \ , \ \phi \rangle = \|\psi\| \|\phi\| \cos(\theta)$ where θ is the angle between ψ_i and ϕ_i. Then this assumptions constrains our set $\{\phi_j\}_{j=1}^d$ so that relative norm of ϕ_i with respect to ϕ_j is lower bounded by $\cos(\theta_{i,j})$: $\frac{\|\phi_i\|}{\|\phi_j\|} > \cos(\theta_{i,j})$.

Informally, the $\{\phi_j\}_{j=1}^d$ are sufficiently spread out in the ambient space so that the learnable ψ_i that maximizes $p_i(X = i)$ is ϕ_i itself: $\psi_i = \frac{\phi_i}{\|\phi_i\|}$. This constraint helps us avoid degenerative cases similar to the $\{\phi_j\}_{j=1}^d$ all equal maximum entropy case described earlier. For example, $\phi_j = \alpha \phi_i$, $i \neq j$ with $\alpha > 0$ is one such degenerative case where one class embedding vector is stacked on a different class embedding, but with higher norm. □

When Assumption 1, Assumption 2, and Assumption 3 hold, the set of class embeddings that maximize the probability of a learnable ψ_i detecting class i is necessarily an ELMES.

Theorem 1. *The set of class embeddings $\{\phi_j\}_{j=1}^d \ \forall j, \ 1 \leq j \leq d$ that maximizes $p_j(X = j)$ is necessarily an ELMES.*

Proof of Theorem 1. Taken with Assumption 1, Assumption 2, and Assumption 3, it suffices to show Theorem 2 and Lemma 4 to prove Theorem 1. □

Theorem 2. *$p_1(X = 1) = p_2(X = 2) = ... = p_d(X = d) \iff \{\phi_j\}_{j=1}^d$ are equiangular and equal norm.*

To show the forward ($\Rightarrow$) direction, it suffices to first show $p_1(X = 1) = p_2(X = 2) = ... = p_d(X = d) \Rightarrow \{\phi_j\}_{j=1}^d$ are equal norm and then show $p_1(X = 1) = p_2(X = 2) = ... = p_d(X = d) \Rightarrow \{\phi_j\}_{j=1}^d$ are equiangular.

Lemma 2. *$p_1(X = 1) = p_2(X = 2) = ... = p_d(X = d) \Rightarrow \{\phi_j\}_{j=1}^d$ are equal norm.*

Proof. This implication holds when $d = 2$:

$$
\begin{aligned}
p_1(X = 1) = \frac{e^{\|\phi_1\|}}{e^{\|\phi_1\|} + e^{\|\phi_2\| \cos(\theta_{1,2})}} &= \frac{e^{\|\phi_2\|}}{e^{\|\phi_2\|} + e^{\|\phi_1\| \cos(\theta_{1,2})}} = p_2(X = 2) \\
e^{\|\phi_1\|}(e^{\|\phi_2\|} + e^{\|\phi_1\| \cos(\theta_{1,2})}) &= e^{\|\phi_2\|}(e^{\|\phi_1\|} + e^{\|\phi_2\| \cos(\theta_{1,2})}) \\
e^{\|\phi_1\| + \|\phi_1\| \cos(\theta_{1,2})} &= e^{\|\phi_2\| + \|\phi_2\| \cos(\theta_{1,2})} \\
\|\phi_1\|(1 + \cos(\theta_{1,2})) &= \|\phi_2\|(1 + \cos(\theta_{1,2})) \\
\|\phi_1\| &= \|\phi_2\|
\end{aligned}
$$

Suppose $d > 2$ and $p_1(X = 1) = ... = p_d(X = d)$. By Assumption 2, all $m-$combinations $\binom{d}{m}$ of $\{p_1(X = 1), ..., p_d(X = d)\}$ are equal. This implies all 2-combinations are equal: $p_i(X = i) = p_j(X = j) \Rightarrow \|\phi_i\| = \|\phi_j\|$. Therefore, $\|\phi_1\| = ... = \|\phi_d\|$. □

Lemma 3. $p_1(X=1) = p_2(X=2) = ... = p_d(X=d) \Rightarrow \{\phi_j\}_{j=1}^d$ *are equiangular.*

Proof. This implication is trivially true when $d = 2$ (see the proof of Lemma 2), and we show it is similarly true when $d = 3$. Following the steps in the proof of Lemma 2, we arrive at the following 3 pairs of equalities:

$$(1)\ e^{\|\phi_1\|(1+\cos(\theta_{1,2}))} + e^{\|\phi_1\|+\|\phi_3\|\cos(\theta_{2,3})} = e^{\|\phi_2\|(1+\cos(\theta_{1,2}))} + e^{\|\phi_2\|+\|\phi_3\|\cos(\theta_{1,3})}$$
$$(2)\ e^{\|\phi_1\|(1+\cos(\theta_{1,3}))} + e^{\|\phi_1\|+\|\phi_2\|\cos(\theta_{2,3})} = e^{\|\phi_3\|(1+\cos(\theta_{1,3}))} + e^{\|\phi_3\|+\|\phi_2\|\cos(\theta_{1,3})}$$
$$(3)\ e^{\|\phi_2\|(1+\cos(\theta_{2,3}))} + e^{\|\phi_2\|+\|\phi_1\|\cos(\theta_{1,3})} = e^{\|\phi_3\|(1+\cos(\theta_{2,3}))} + e^{\|\phi_3\|+\|\phi_1\|\cos(\theta_{1,2})}$$

From Lemma 2, $p_1(X=1) = p_2(X=2) = p_3(X=3) \Rightarrow \|\phi_1\| = \|\phi_2\| = \|\phi_3\|$, so the above pairs of equalities reduce to:

$$(1)\ \cos(\theta_{2,3}) = \cos(\theta_{1,3})$$
$$(2)\ \cos(\theta_{2,3}) = \cos(\theta_{1,3})$$
$$(3)\ \cos(\theta_{1,3}) = \cos(\theta_{1,2})$$

and when $d = 3$, $\{\phi_j\}_{j=1}^3$ are equiangular.

Suppose $d > 3$ and $p_1(X=1) = ... = p_d(X=d)$. By Assumption 2, all $m-$combinations $\binom{d}{m}$ of $\{p_1(X=1), ..., p_d(X=d)\}$ are equal. This implies all 3-combinations are equal: $p_i(X=i) = p_j(X=j) = p_k(X=k) \Rightarrow \theta_{i,j} = \theta_{i,k} = \theta_{j,k}$. Therefore, all angles are equal $\theta_{i,j} = \theta_{l,m}$ for $1 \le i, j, l, m \le d$. □

Proof of Theorem 2. ($\Rightarrow$) Suppose $p_1(X=1) = p_2(X=2) = ... = p_d(X=d)$.

By Lemma 2 and Lemma 3, $p_1(X=1) = p_2(X=2) = ... = p_d(X=d) \Rightarrow \{\phi_j\}_{j=1}^d$ are equiangular and equal norm.

($\Leftarrow$) Suppose $\{\phi_j\}_{j=1}^d$ are equiangular and equal norm. Let $\|\phi\|$ be the norm of any vector in our set and $\cos(\theta)$ be the pairwise angle between any two vectors. Then

$$p_i(X=i) = \frac{e^{\|\phi\|}}{e^{\|\phi\|} + (d-1)e^{\|\phi\|\cos(\theta)}} = p_j(X=j)$$

for any $1 \le i, j \le d$. □

Lemma 4. *For a set of equiangular and equal norm vectors, maximum equiangularity maximizes* $\sum_j p_j(X=j)$.

Proof. The maximum pairwise angle between two vectors in $\mathbb{R}^d$ is π, and from Theorem 2

$$p_i(X=i) = p_j(X=j) = \frac{e^{\|\phi\|}}{e^{\|\phi\|} + (d-1)e^{\|\phi\|\cos(\theta)}}$$

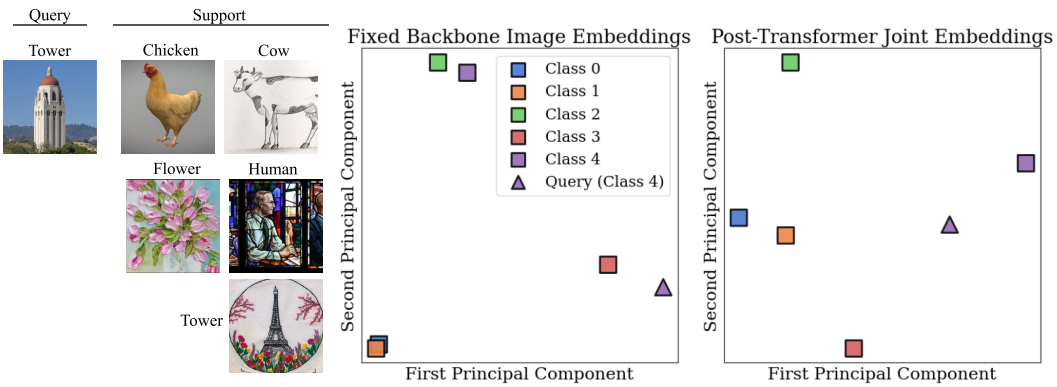
for all $1 \le i, j \le d$. Increasing the angle θ decreases $\cos(\theta)$. Decreasing $\cos(\theta)$ only decreases the denominator, which in turn, increases $p_i(X=i)$. Therefore, maximizing the pairwise angle between all vectors maximizes $p_i(X=i)$ for all $1 \le i \le d$. □

A.5 Label Symmetry

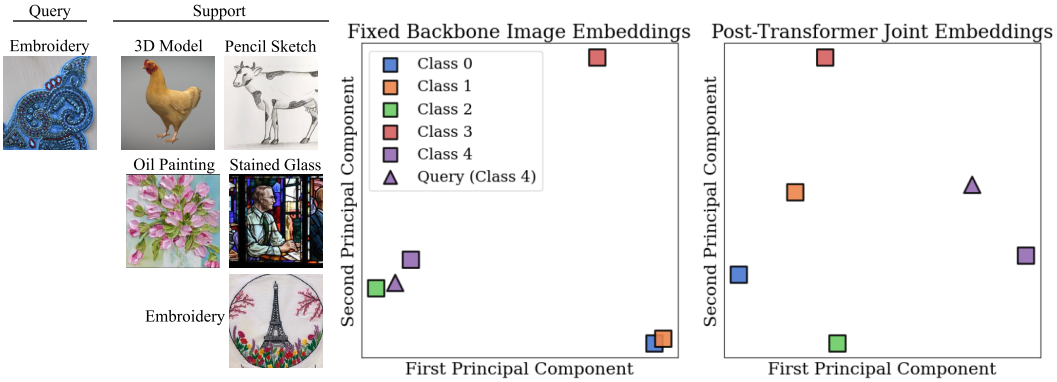
Alternatively when viewed through the lens of information theory, we can reinterpret an ELMES as the class embedding that minimizes the entropy of ψ_i detecting class i. Informally, ELMES causes ψ_i to have the least uncertainty when detecting class i.

Lemma 5. *Let* H_i *be the entropy of* $p_i(X)$. *An ELMES minimizes* H_i.

Proof of Lemma 5. Equal norm and equiangular $\{\phi_j\}_{j=1}^d$ are bounded in norm, and thus, the set of probability distributions we obtain $\{p_1, p_2, ..., p_d\}$ belong to a capped simplex [Warmuth and Kuzmin, 2008] $\Delta_c^d = \{p \in \Delta | \max_i p_i \le c\}$ where $c = \frac{e^{\|\phi\|^2}}{e^{\|\phi\|^2} + (d-1)e^{\|\phi\|^2\cos(\theta)}}$. Clearly, among such probability vectors, the minimum entropy is achieved at the boundary where $\cos(\theta)$ is minimized, i.e., when the $\{\phi_j\}_{j=1}^d$ are maximally equiangular. □



(a) Left: An example task, classifying images by the objects depicted. Center: CLIP image embeddings on this task's images. Right: joint image+label representations after the last CAML attention layer for the same task.



(b) Left: An example task, classifying images by the artistic medium used. Center: CLIP image embeddings on this task's images. Right: joint representations after the last CAML attention layer for the same task.

Figure 3: Two sample tasks over the same support images but utilizing different criteria to define classes. The nature of the query image informs the task being presented, e.g. classification by object (top) vs. classification by texture (bottom). For both, the final-layer attention outputs provide better separation between class representations and groups the query representation with the proper task, even when projected into 2D space by PCA.

A.6 Permutation Invariance.

In addition to label symmetry, it is also desirable for the output prediction of CAML to not depend on the order of demonstrations in the sequence. As Fifty et al. [2023] show that a two-class, non-ELMES version of CAML is invariant to permutations in the input sequence, it suffices to show that the ELMES label encoder is equivariant to permutations in the input sequence.

Lemma 6. *Consider a n-sequence of one-hot labels stacked into a matrix $\mathcal{S} \in \mathbb{R}^{n \times w}$, and an ELMES label encoder denoted by $W \in \mathbb{R}^{w \times d}$ with w denoting "way" and d the dimension of the label embedding. The label embedding $\mathcal{S}W$ is equivariant to permutations.*

B Analysis

To better understand how CAML learns, we conduct empirical analyses on (1) its ability to dynamically update its representations at inference time, and (2) the effect of the fixed ELMES class embedding.

Context-Aware Representations. Dissimilar from other meta-learning algorithms and due to recasting meta-learning as sequence modeling, CAML considers the full context of a query and support set to predict the label of the query. Specifically, the query dynamically influences the representation of support set points, and the support set points dynamically influence the representation of the query as this sequence is passed through the layers of a Transformer encoder. This property

enables universal meta-learning by allowing the model to update the support and query representations based on the context of the task, not only the contents of the images.

An example where the query dynamically influences the support set is visualized in Figure 3. Given only the 5 support examples, the prediction task is ambiguous. However, the nature of the query determines the prediction task. The query image of a tower in Figure 3a reduces the task to generic object recognition: classify the query based on the object portrayed in the image. On the other hand, and as visualized in Figure 3b, the query image of embroidery reduces the prediction task to texture identification: classify the query based on artistic medium.

To analyze how dynamic representations affect CAML, we examine the representations of the support set and query vectors at the input to and output of the Transformer encoder. For both examples visualized in Figure 3a and Figure 3b, the Transformer encoder learns to separate support set vectors by class identity, and moreover, group the query representation with the correct support set example.

We find the frozen CLIP image embeddings are actually antagonistic for the classification-by-texture task visualized in Figure 3b: the query image embedding is closest to the support set example for the second class, "oil painting". Moreover, we find that MetaQDA, which relies on frozen CLIP image embeddings, groups the query with "oil painting" and therefore misclassify this example. On the other hand, as CAML considers the full context of the query and support set, it develops representations of the query in the context of the support set—and the support set in the context of the query—to group the query with the "embroidery" support set image as they share the same texture, thereby correctly classifying this example.