EUGens: Efficient, Unified, and General Dense Layers

Sang Min Kim*, Byeongchan Kim*, Arijit Sehanobish*, Somnath Basu Roy Chowdhury*, Rahul Kidambi*, Dongseok Shim, Avinava Dubey*, Snigdha Chaturvedi, Min-hwan Oh, Krzysztof Choromanski*

** Equal Contribution*

Correspondence to arijit.sehanobish1@gmail.com

Random Fourier Feature Maps for Approximating Feed Forward Layers

- Feed Forward Layer (FFL)

$$\mathbf{x} \mapsto f(\mathbf{W}\mathbf{x} + \mathbf{b}),\ x \in \mathbb{R}^d,\ \mathbf{W} \in \mathbb{R}^{l \times d},\ \mathbf{b} \in \mathbb{R}^l(\text{bias}),\ f : \mathbb{R} \to \mathbb{R}\,(\text{ activation function})$$

Random Fourier Feature Maps for Approximating Feed Forward Layers

- Feed Forward Layer (FFL)

  $\mathbf{x} \mapsto f(\mathbf{W}\mathbf{x} + \mathbf{b}),\ x \in \mathbb{R}^d,\ \mathbf{W} \in \mathbb{R}^{l \times d},\ \mathbf{b} \in \mathbb{R}^l(\text{bias}),\ f : \mathbb{R} \to \mathbb{R}(\text{ activation function})$

- EUGen Layers to approximate FFLs :

$$\mathbf{EUGen}^k(\mathbf{W}, \mathbf{x}) = \left\langle \Psi(\mathbf{concat}(\prod_{j=1}^{i} \mathbf{G}_j^i \mathbf{W}^+)_{i=0,\ldots,k}), \Phi(\mathbf{concat}(\prod_{j=1}^{i} \mathbf{G}_j^i \mathbf{x}^+)_{i=0,\ldots,k}) \right\rangle$$

Random Fourier Feature Maps for Approximating Feed Forward Layers

- Feed Forward Layer (FFL)

  $\mathbf{x} \mapsto f(\mathbf{W}\mathbf{x}+\mathbf{b}),\ x \in \mathbb{R}^d,\ \mathbf{W} \in \mathbb{R}^{l \times d},\ \mathbf{b} \in \mathbb{R}^l(\text{bias}),\ f : \mathbb{R} \to \mathbb{R}(\text{ activation function})$

- EUGen Layers to approximate FFLs :

$$\mathrm{EUGen}^k(\mathbf{W},\mathbf{x}) = \left\langle \Psi(\mathrm{concat}(\prod_{j=1}^{i} \mathbf{G}_j^i \mathbf{W}^+)_{i=0,\dots,k}),\ \Phi(\mathrm{concat}(\prod_{j=1}^{i} \mathbf{G}_j^i \mathbf{x}^+)_{i=0,\dots,k}) \right\rangle$$

- $\mathbf{\Phi}, \mathbf{\Psi} : \mathbb{C} \to \mathbb{R}$, **G** : random Gaussian matrices
- $\mathbf{W}^+ := \mathrm{concat}(\mathbf{W}, \mathbb{1}),\ \ \mathbf{x}^+ := \mathrm{concat}(\mathbf{x}, ||\mathbf{x}||_2)$

Random Fourier Feature Maps for Approximating Feed Forward Layers

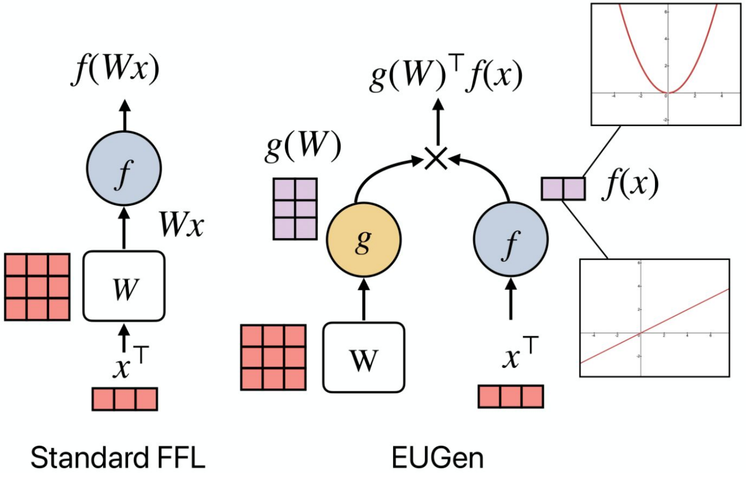
- Feed Forward Layer (FFL)

  $$\mathbf{x} \mapsto f(\mathbf{W}\mathbf{x} + \mathbf{b}),\ x \in \mathbb{R}^d,\ \mathbf{W} \in \mathbb{R}^{l \times d},\ \mathbf{b} \in \mathbb{R}^l (\text{bias}),\ f : \mathbb{R} \to \mathbb{R} (\text{ activation function})$$

- EUGen Layers to approximate FFLs :

  $$\mathbf{Eug}(\mathbf{W}, \mathbf{x}) := \langle \mathbf{\Psi}(\mathbf{W}), \mathbf{\Phi}(\mathbf{x}) \rangle$$

- **Main idea** : nonlinear calculations are replaced with the simple linear (dot-product) kernel, but via the nonlinear maps $\mathbf{\Psi}$ and $\mathbf{\Phi}$.



Some Theoretical Properties of EUGens (Unbiased Estimation)

Theorem : EUGens can unbiasedly approximate FFLs with polynomial activations *f*.

Let *f* be $f(x) := \sum_{i=0}^{k} a_i x^i$, then choosing $\mathbf{G}_j^i(\cdot,\cdot) \sim \frac{1}{\sigma_{i,j} m^{\frac{1}{2i}} |a_i|^{\frac{1}{2i}} \xi_i(2i)} \mathcal{D}_j^i$, where

- $\mathcal{D}$ is a zero-mean distribution
- $\xi_i^t(t) = \mathrm{sgn}(a_i)$.
- Φ(x) = Ψ(x) = x are identity mappings.

Then $\mathbb{E}\left[\mathrm{EUGen}(\mathbf{W}, \mathbf{x})\right] = f(\mathbf{W}\mathbf{x})$

Variance of our Estimator

Theorem : The variance of the estimator $Z = \widehat{\mathrm{FFL}}(\mathbf{W}, \mathbf{x})[u]$ of FFL(**W**, **x**)[u] for u th row is given by

$$\mathrm{Var}(Z) = \frac{1}{m} \sum_{i=0}^{k} ((2(\mathbf{w}(u)^\top \mathbf{x})^2 + \|\mathbf{w}(u)\|_2^2 \|\mathbf{x}\|_2^2 + (\tau_{i,j} - 3) \sum_{s=1}^{d} \mathbf{w}(u)_s^2 x_s^2)^i - \rho_i) a_i^2$$

Where

- **w**(u) := u th row of **w**
- $\varrho_i := (\mathbf{w}(u)^\top \mathbf{x})^{2i}$
- $\boldsymbol{\tau}_{i,j}$ is the 4th moment of the distribution $\mathcal{D}^i_j$

Variance of our Estimator

Theorem : The variance of the estimator $Z = \widehat{\mathrm{FFL}}(\mathbf{W}, \mathbf{x})[u]$ of FFL(**W**, **x**)[u] for u th row is given by

$$\mathrm{Var}(Z) = \frac{1}{m}\sum_{i=0}^{k}((2(\mathbf{w}(u)^\top \mathbf{x})^2 + \|\mathbf{w}(u)\|_2^2\|\mathbf{x}\|_2^2 + (\tau_{i,j} - 3)\sum_{s=1}^{d}\mathbf{w}(u)_s^2 x_s^2)^i - \rho_i)a_i^2$$

Where

- **w**(u) := u th row of **w**
- ϱ_i := $(\mathbf{w}(u)^{\top}\mathbf{x})^{2i}$
- $\boldsymbol{\tau}_{i,j}$ is the 4th moment of the distribution $\mathcal{D}^i_j$

Key Takeaway : Variance scales as O(1/*m*) so variance converges to 0 as m → ∞

Concentration Results for EUGens

Theorem : Under some mild assumptions,

$$\mathbb{P}[|\widehat{\mathbf{FFL}}(\mathbf{W},\mathbf{x})[u] - \mathbf{FFL}(\mathbf{W},\mathbf{x})[u]| \geq \epsilon] \leq h(\epsilon) \stackrel{\text{def}}{=} 2\exp\left(-\frac{m\epsilon^2}{2(\eta_1+\eta_2)^2k^2}\right)$$

Where

- η_1 and η_2 are explicit constants.
- k := degree of polynomial activation

Concentration Results for EUGens

Theorem : Under some mild assumptions,

$$\mathbb{P}[|\widehat{\mathbf{FFL}}(\mathbf{W},\mathbf{x})[u] - \mathbf{FFL}(\mathbf{W},\mathbf{x})[u]| \geq \epsilon] \leq h(\epsilon) \stackrel{\text{def}}{=} 2\exp\left(-\frac{m\epsilon^2}{2(\eta_1+\eta_2)^2k^2}\right)$$

Where

- η_1 and η_2 are explicit constants.
- k := degree of polynomial activation

Key Takeaway : Probability that EUGen is larger than the approximating FFL can be arbitrary small if *m* is large.

Extending our results for general activation functions

Theorem : For any *k*, there exists a *k*-order EUGen layer, such that

$$\mathbb{P}(|\widehat{\mathbf{FFL}}(\mathbf{W},\mathbf{x})[u] - \mathbf{FFL}(\mathbf{W},\mathbf{x})[u]| \geq \epsilon + C\omega(\frac{1}{\sqrt{k}})) \leq \min(\frac{\mathbf{Var}\left(\widehat{\mathbf{FFL}}(\mathbf{W},\mathbf{x})[u]\right)}{\epsilon^2}, h(\epsilon))$$

Where

- $\omega(x) = \sup_{|t_1 - t_2| \leq x} |f(t_1) - f(t_2)|$
- C is some constant.

Extending our results for general activation functions

Theorem : For any *k*, there exists a k-order EUGen layer, such that

$$\mathbb{P}(|\widehat{\mathbf{FFL}}(\mathbf{W},\mathbf{x})[u] - \mathbf{FFL}(\mathbf{W},\mathbf{x})[u]| \geq \epsilon + C\omega(\frac{1}{\sqrt{k}})) \leq \min(\frac{\mathrm{Var}\left(\widehat{\mathbf{FFL}}(\mathbf{W},\mathbf{x})[u]\right)}{\epsilon^2},\ h(\epsilon))$$

Where

- $\omega(x) = \sup_{|t_1 - t_2| \leq x} |f(t_1) - f(t_2)|$
- *C* is some constant.

Key Takeaway : Since Var scales as O(1/*m*), and h(ϵ) exponentially small in *m*, choosing *m >> 1/ε^2 , can make RHS arbitrarily small.*

Summarizing our Theoretical Contributions

- We have proven the first unbiasedness results for any polynomial activation functions.
- Previous results are known for polynomials with positive coefficients.
- Variance of estimator decreases as the number of random features increase.
- Leveraging the fact that polynomials can approximate any continuous functions, we show that EUGens can be used to approximate arbitrary FFLs.
- Finally non-linearity in EUGen is introduced by Φ or Ψ or $k > 1$.

Some Practical Benefits of EUGens

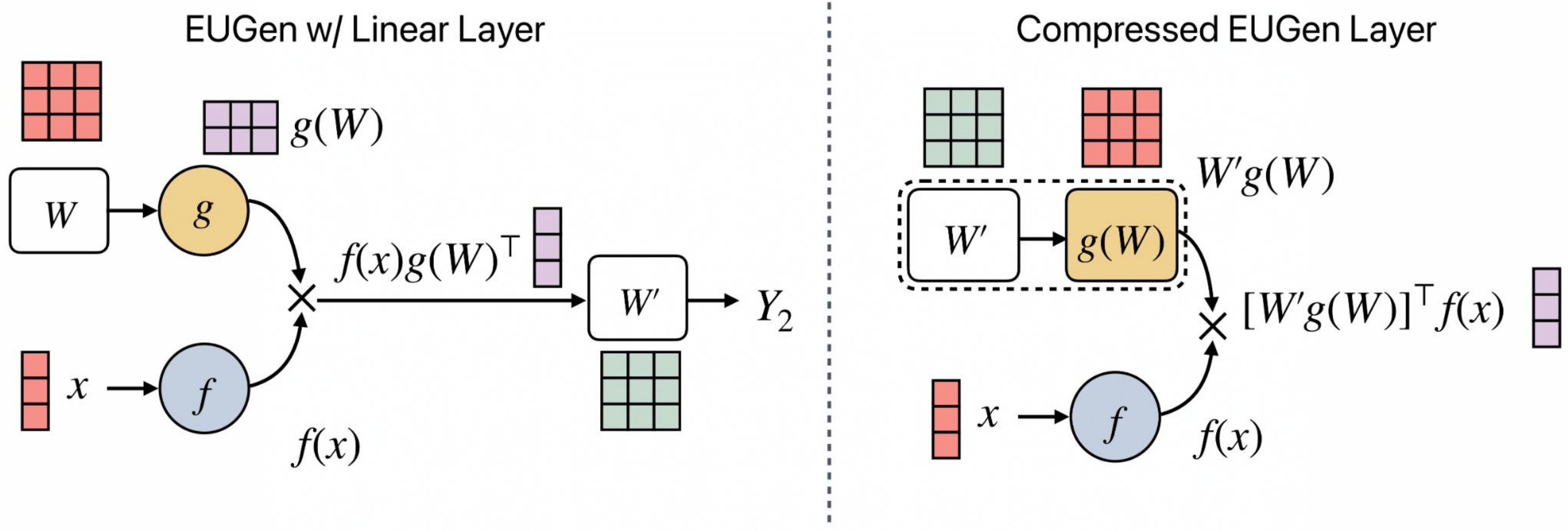
- **Network Compression :** Instead of transforming layer parameters with Ψ, one can directly learn vectors $\Psi(w)$. Then the number of trainable parameters becomes **$O(m\ell)$** rather than $O(\ell d)$ for $m << d$, reducing the parameter count.

- **Computational Savings :** the overall time complexity (given pre-computed embeddings $\Psi(w)$) is **sub-quadratic** in layers' dimensionalities.

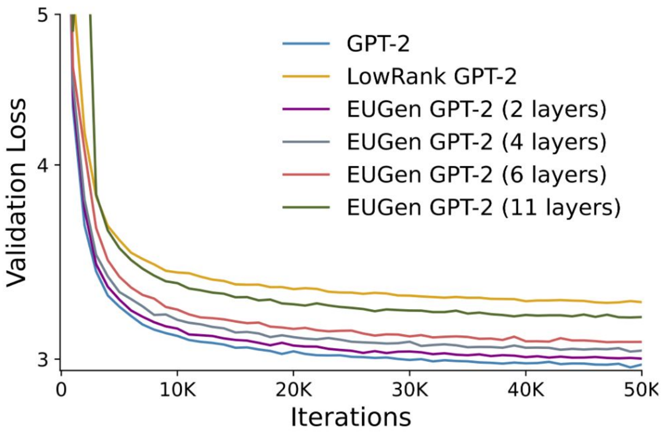
- **Compression during inference :** a two-tower representation can be used iteratively to compactify multiple FFLs of NNs.

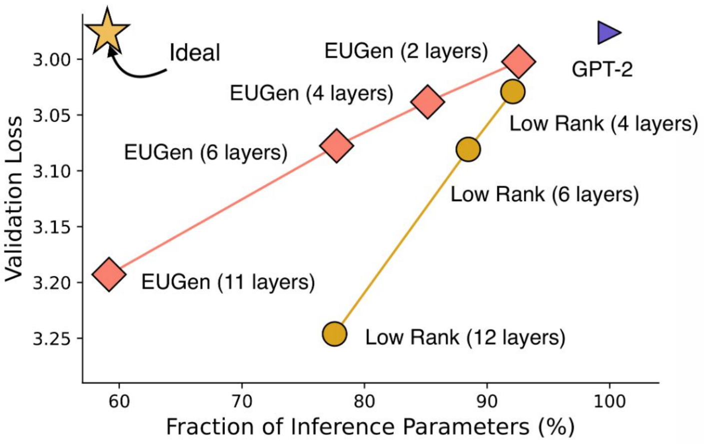
- **Training without Backpropagation :** Possible for certain loss functions and can be used for distillation.

Compression During Inference



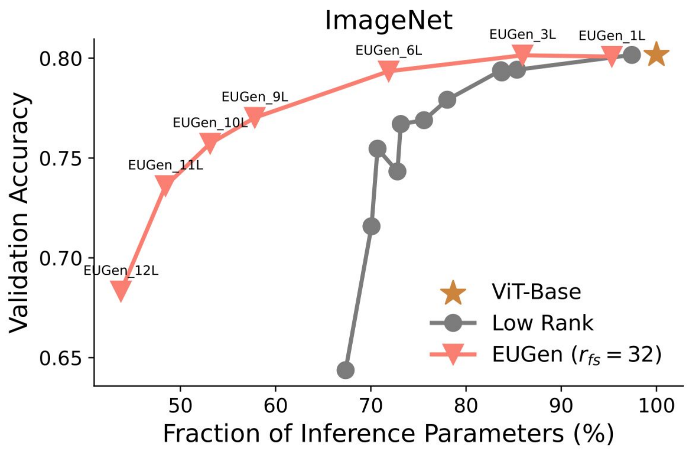
Experiments : Language Modeling (Pretraining)

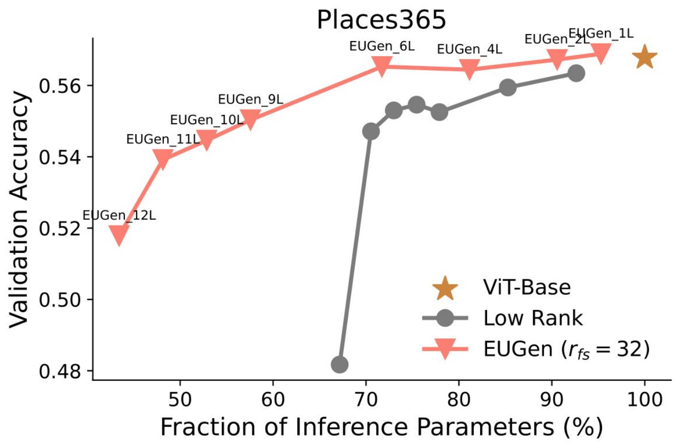




- Pre-train GPT-2 architecture on **36.8B** tokens and replace the FFLs with EUGen layers.
- Compare against GPT-2, low-rank GPT-2, and GPT-2 with varying EUGen layers.
- Observe that EUGens are good approximations of FFLs, achieving similar validation loss to GPT-2.
- More EUGen layers slightly raise loss due to error accumulation.
- Other NLP experiments using BERT and DistilBERT are in our paper.

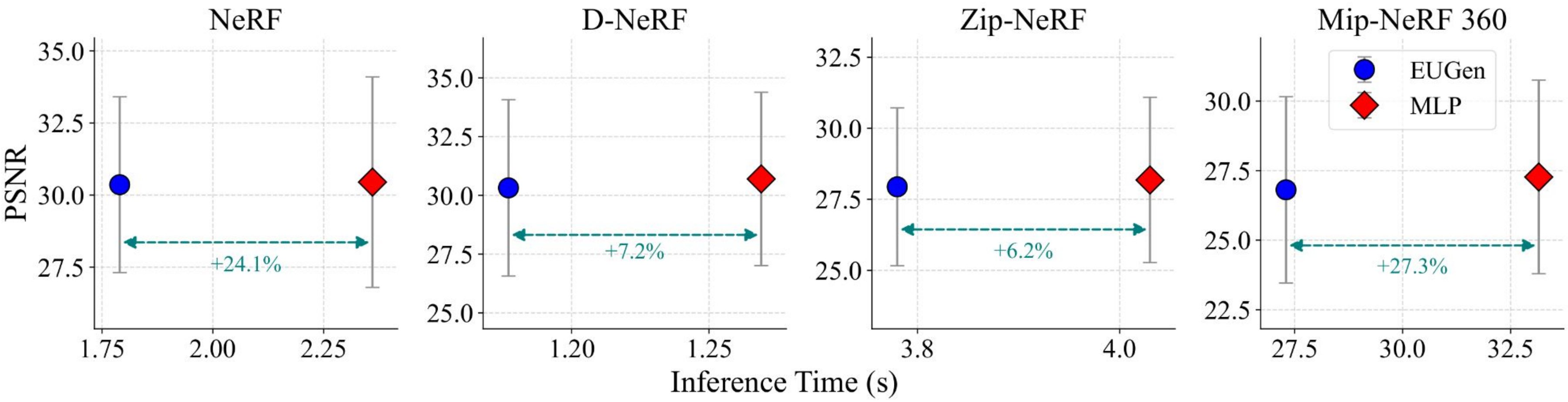
Experiments : Image Classification





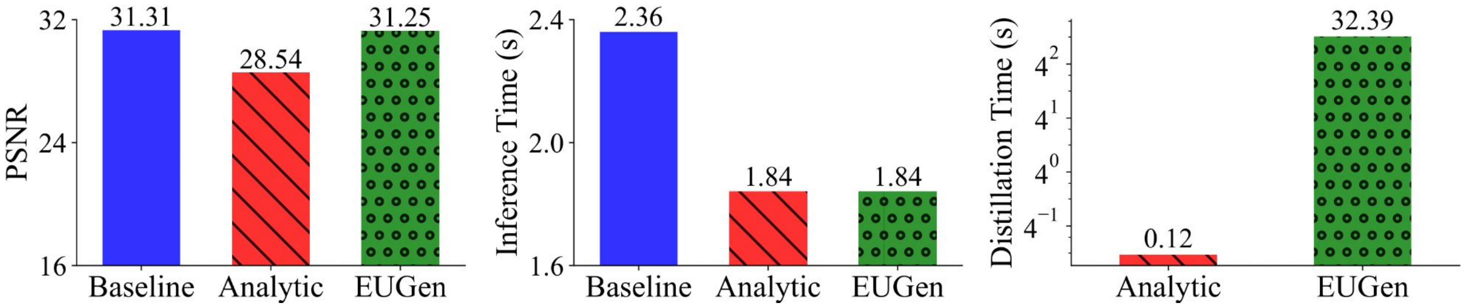
- Systematically replace all the FFLs in ViT with EUGen and evaluate it on the ImageNet and Places365.
- EUGens achieve the best trade-off, matching ViT performance with **30%** fewer inference parameters.
- More results on diverse datasets, EfficientViT, and larger models like ViT-L are in our paper.

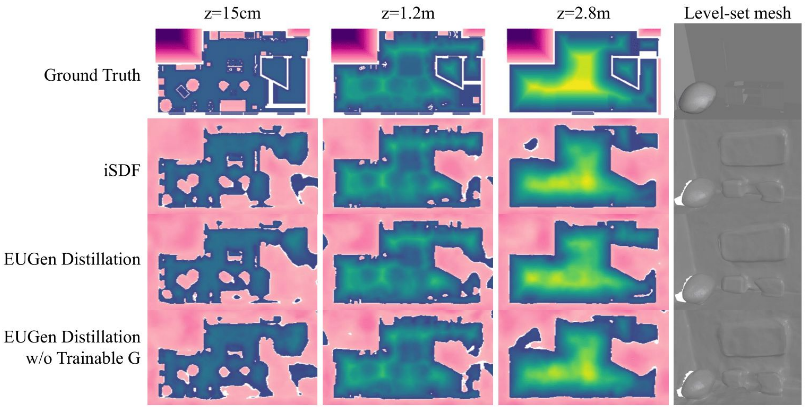
Experiments : 3D Reconstruction



- Seamlessly inject EUGens into various neural 3D scene reconstruction methods including NeRF, Mip-NeRF 360, Zip-NeRF, and D-NeRF.
- EUGen significantly reduces inference time with virtually no loss in reconstruction quality.
- Also use EUGens into iSDF, achieving **5%** faster training and **23%** faster inference with similar accuracy.

Experiments : Distillation using EUGens





- Layer-wise Knowledge Distillation (KD) by EUGens by mimicking the outputs of a FFL.
- Easily slotted in pretrained nets
- Analytic (w/o trainable G) variants refers to KD performed without *backpropagation.*
- We match baseline performance while reducing inference compute.

For more results see our paper and come to our poster session

Thank you!