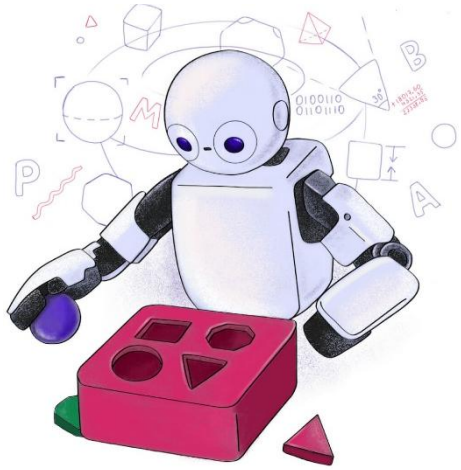
TP558 - Tópicos avançados em Machine Learning:
Learning to Resize Images for Computer Vision Tasks



Inatel

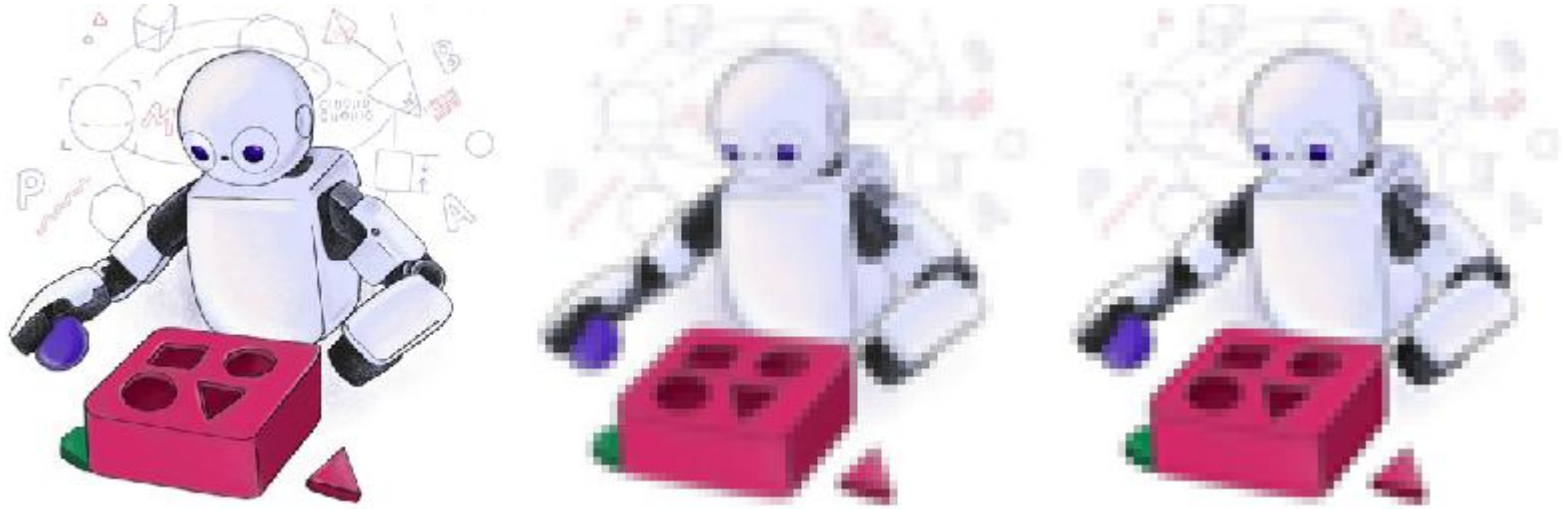
LUIS ENRIQUE CARDOZO RAMIREZ

Introdução

- Já pensaram como **reduzir** uma imagem de **alta resolução**, **sem perder detalhes cruciais**, como a textura da água ou das rochas?
- Imagine uma fotografia que precisa ser exibida em uma tela pequena ou usá-la em um aplicativo.

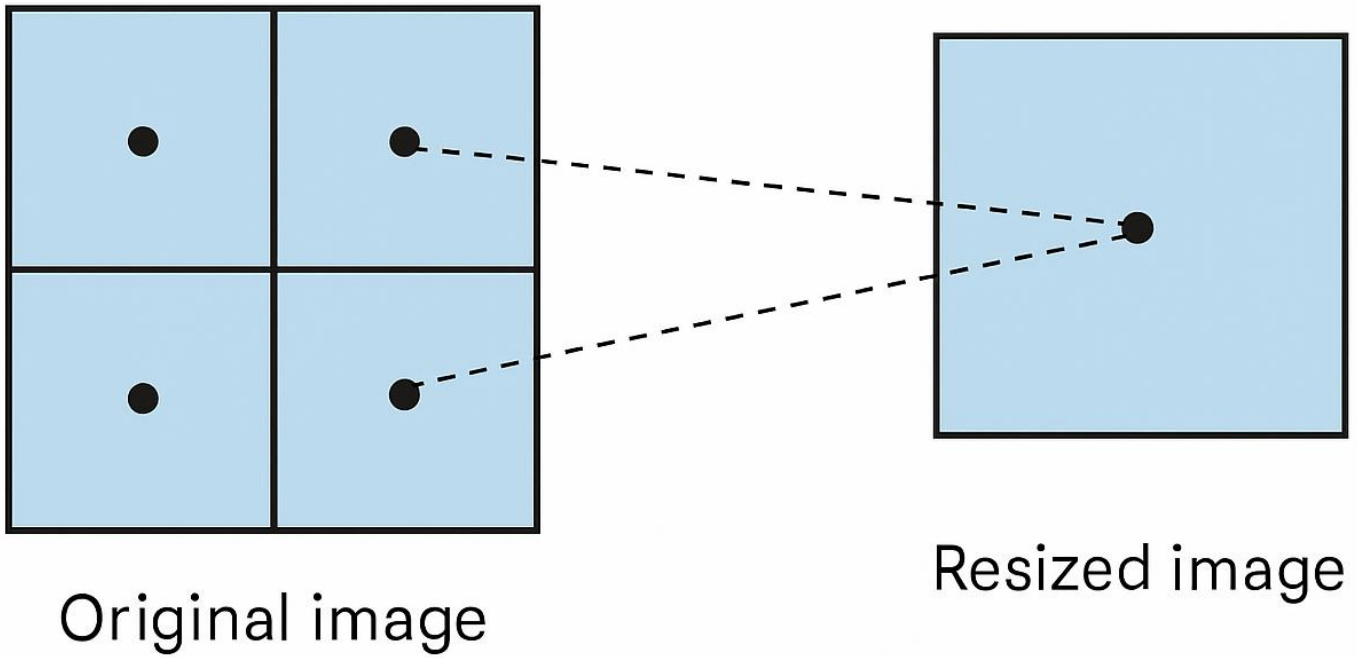
Introdução

- Você precisa torná-la menor. Esse processo é chamado de **redimensionamento**.
- Tradicionalmente, os programas de computador utilizam **métodos padronizados** para isso, baseados em fórmulas matemáticas pré-definidas.



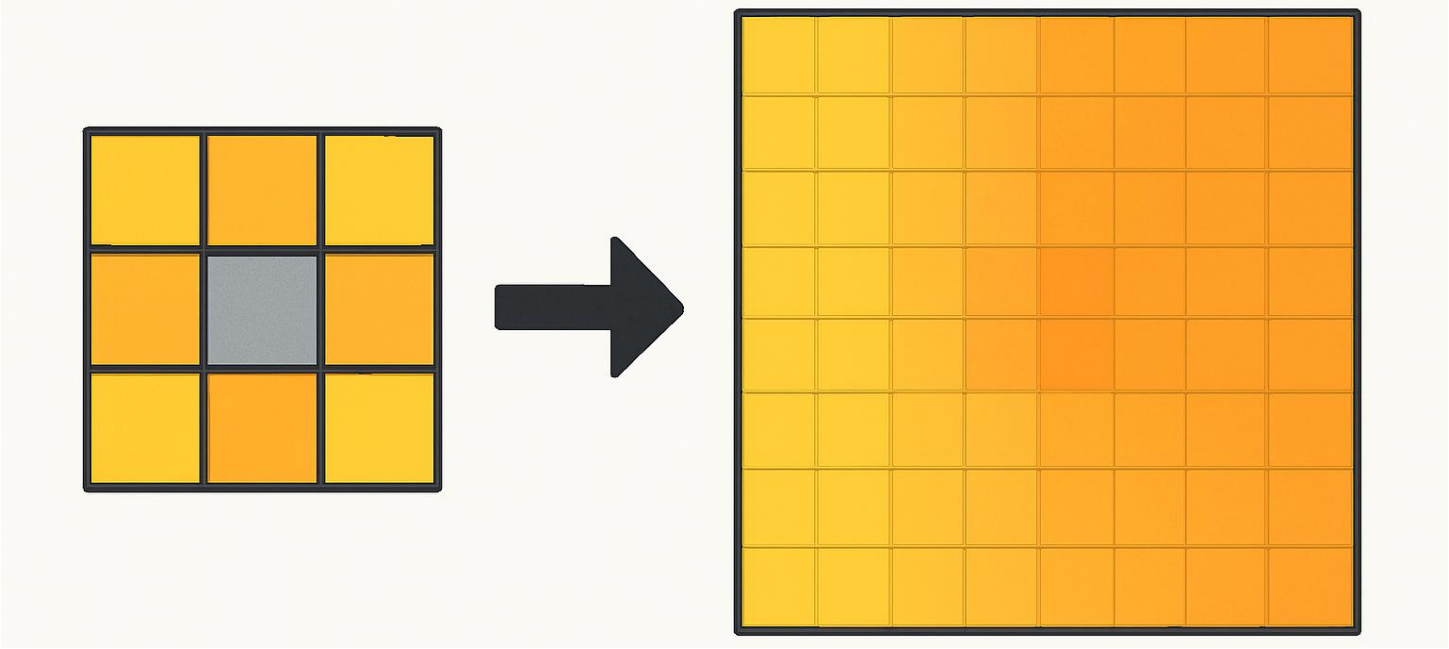
Introdução

- Um método bastante utilizado é a **Interpolação Bilinear**, que considera os quatro pixels vizinhos mais próximos na imagem original e calcula uma média ponderada para gerar um novo pixel na imagem redimensionada.



Introdução

- Outro método bastante utilizado é a **Interpolação Bicubica**, Técnica avançada de redimensionamento de imagens que calcula novos pixels com base nos 16 vizinhos mais próximos, gerando transições suaves e maior qualidade visual.



Introdução

- O problema central reside no fato de que esses resizers aplicam a mesma **lógica de escalonamento em toda a imagem**, sem levar em conta quais detalhes são cruciais para o reconhecimento
- Ele não distingue, por exemplo, entre o gramado ao fundo. Ao calcular a média das cores, ao mesmo tempo em que **preserva sem dificuldade áreas de menor relevância**, como uma parede uniforme.

Introdução

Solução Inteligente:

Um Redimensionador que "**Adapta**" a Imagem para **Máquinas**

> E se, em vez de usar um **método fixo** como o redimensionamento bilineal ou bicubico, tivéssemos um "**especialista digital**" que analisa a imagem e a redimensiona de forma a **destacar** as características mais importantes para uma tarefa específica de **visão computacional**?

Introdução

Solução Inteligente:

Um Redimensionador que "**Adapta**" a Imagem para **Máquinas**

Learning to Resize Images for Computer Vision Tasks

Hossein Talebi and Peyman Milanfar

Introdução

Proposed learned image resizer

- O método utiliza uma **rede neural convolucional (CNN)** que **aprende, junto com o modelo** de visão, a redimensionar imagens de maneira otimizada para a tarefa-alvo.
- O treinamento é conjunto, utilizando a **mesma função de perda** (entropia cruzada para classificação).

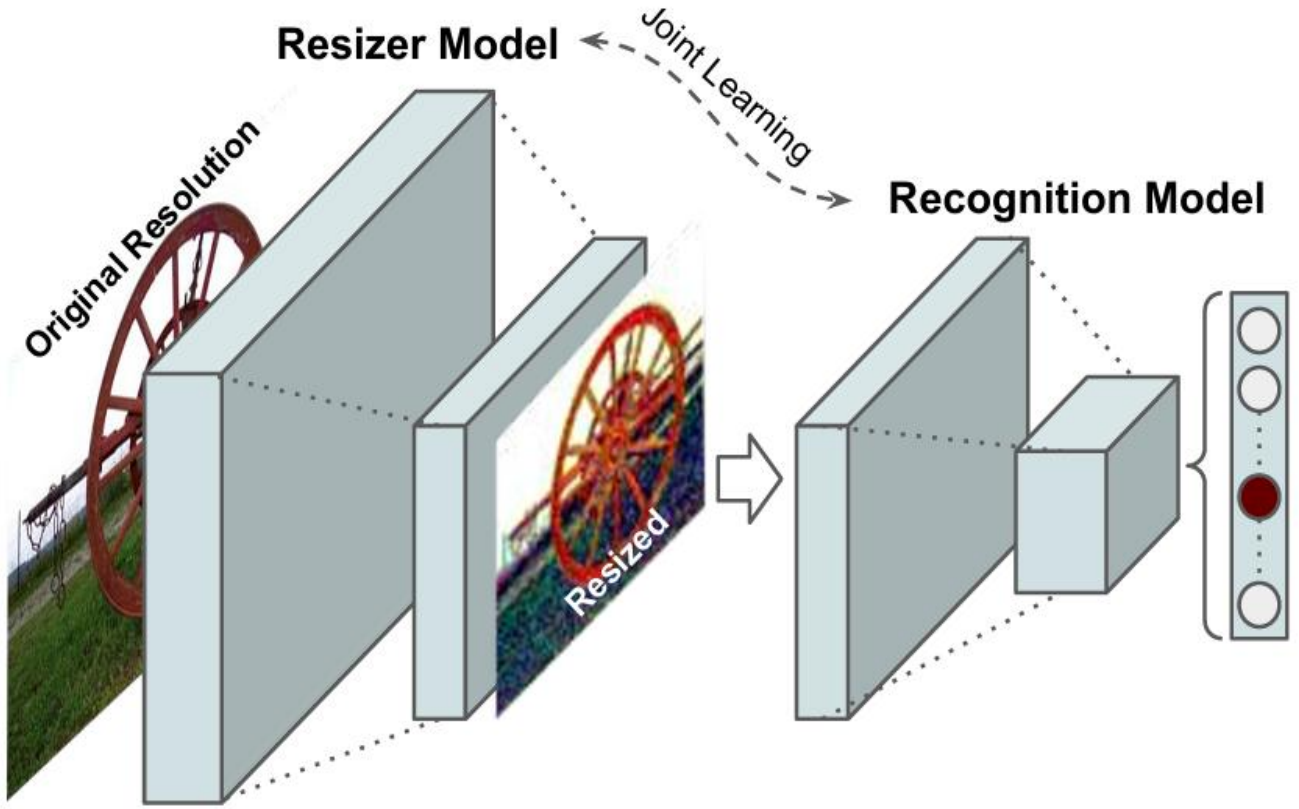


Figure 1. Our proposed framework for joint learning of the image resizer and recognition models.

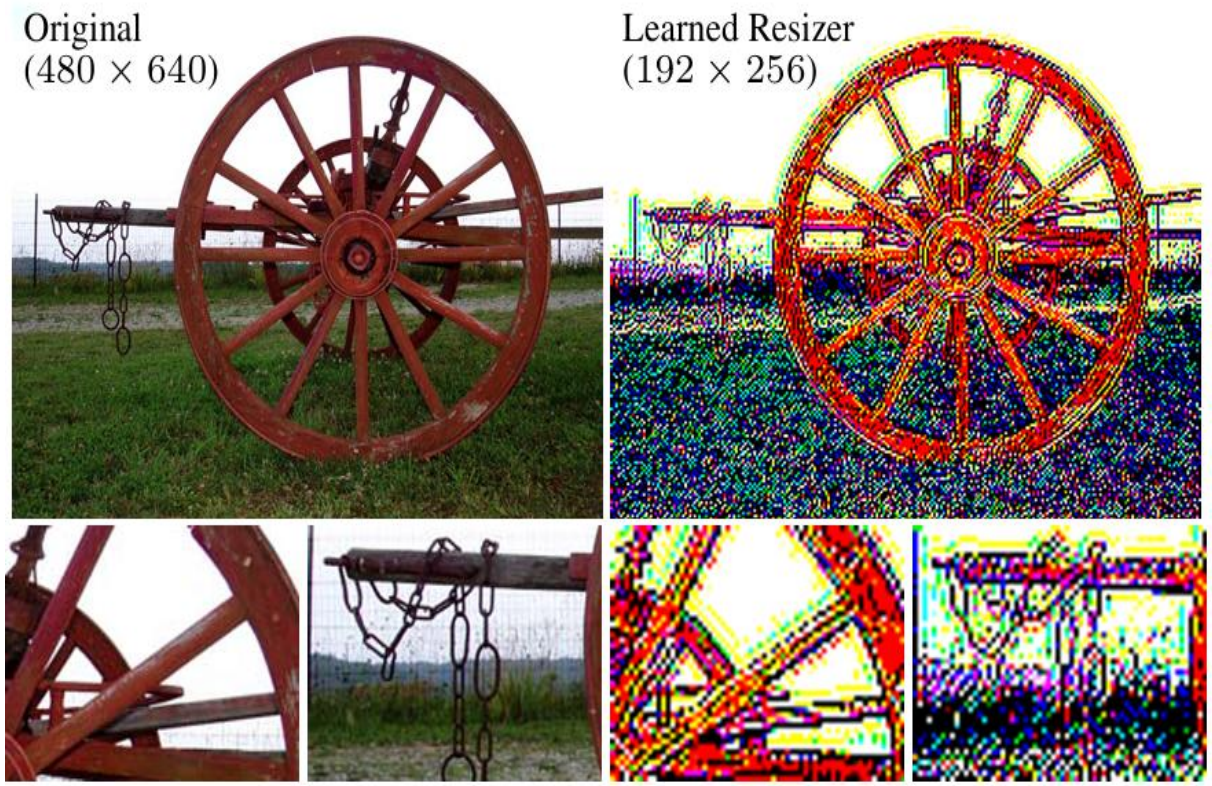
Introdução

Proposed learned image resizer

Característica	Descrição
Integração adaptativa	Adapta-se a diferentes modelos, melhorando consistentemente os resultados.
Independência de perdas perceptuais	Gera efeitos visuais otimizados para máquinas, não para humanos.
Flexibilidade de escala	Permite redimensionamento em fatores arbitrários, buscando a resolução ideal.
Expansão para IQA	Adapta-se com sucesso à avaliação de qualidade de imagem.
Aplicação em inferência remota	Reduz latência em sistemas cliente-servidor, mantendo precisão.

Introdução

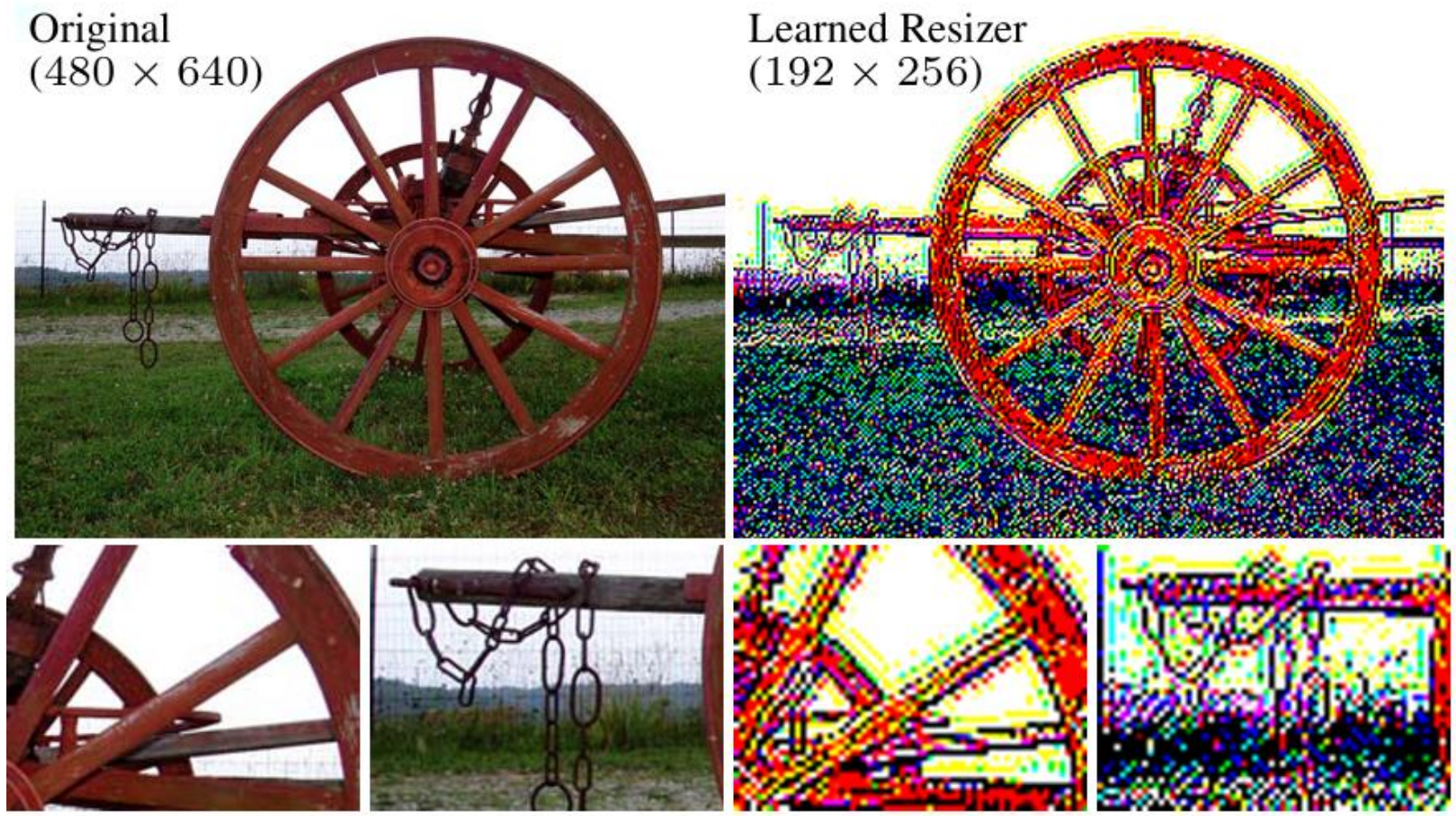
- **Redimensionamento Otimizado para Máquinas**
- A imagem original é redimensionada por um modelo aprendido em conjunto com o **Inception-v2**, com uma redução de 480 × 640 para 192 × 256 pixel.
- Diferente dos métodos tradicionais, a **imagem resultante não foca na qualidade visual humana**, mas sim em realçar características "amigáveis para máquinas", como detalhes de alta frequência.



Introdução

Solução Inteligente: Redimensionamento **Otimizado para Máquinas**

- A Figura 2 ilustra um exemplo do redimensionador aprendido, treinado para a classificação no conjunto de dados ImageNet.



Fundamentação teórica

1. Interpolação Bilinear (Métodos Tradicionais de Redimensionamento)

A interpolação Bilinear é usada como baseline no artigo (Tabela 1, Figura 3), redimensionando imagens de forma suave para uniformidade em mini-batches. No entanto, não é otimizada para percepção de máquinas, causando perda uniforme de detalhes ao contrário do redimensionador aprendido.

O valor do píxel interpolado Q(x, y) é calculado como:

$$Q(x, y) = (1 - a)(1 - b)P_{11} + a(1 - b)P_{21} + (1 - a)bP_{12} + abP_{22}$$

Fundamentação teórica

Uma Receita Adaptável

Este método é guiado por uma **formulação que otimiza o redimensionador** e o modelo base juntos, como descrito na Figura 1. A base teórica é o aprendizado conjunto, onde a imagem redimensionada é otimizada para **maximizar o desempenho da tarefa, não a qualidade visual humana.**

Formulação Matemática

O objetivo é encontrar os parâmetros ótimos do redimensionador (θ) e do modelo base (ϕ) que **minimizem o erro da tarefa final**, conforme a equação:

$$(\theta^*, \phi^*) = \arg\min_{\theta,\phi} E_{(x,y)\sim D}\left[\mathcal{L}\left(g_\phi\left(f_\theta(x)\right), y\right)\right]$$

Fundamentação teórica

$$(\theta^*, \phi^*) = \arg\min_{\theta,\phi} E_{(x,y)\sim D} \left[\mathcal{L} \left(g_\phi \left(f_\theta(x) \right), y \right) \right]$$

Análise dos Componentes

- $f_\theta(x)$: O redimensionador (CNN, Figura 3), que transforma a imagem de entrada x (e.g., 480×640) em uma versão redimensionada (e.g., 192×256). θ são seus pesos (11k–93k parâmetros, Tabela 2).
- $g_\phi(...)$: O modelo base (e.g., Inception-v2, ResNet-50) que processa a imagem redimensionada para a tarefa. ϕ são seus pesos.
- $\mathcal{L}(..., y)$: Função de perda da tarefa: entropia cruzada com *label-smoothing* para classificação (ImageNet, Equação 1) ou Earth Mover's Distance (EMD) para IQA (AVA, Equação 2).
- $E_{(x,y)\sim D}$: Média sobre o dataset D (e.g., pares imagem-etiqueta do ImageNet ou AVA).
- $\arg\min_{\theta,\phi}$: Otimização conjunta via gradiente descendente (e.g., *momentum optimizer*).

Fundamentação teórica

Entropia Cruzada con Label Smoothing

A função de perda é definida como

$$L = -\sum_{k=1}^{K} q'_k \log(p_k)$$

Entropia Cruzada com Label Smoothing ajusta os parâmetros do modelo de classificação (ϕ) para **minimizar o erro de previsão**, ao mesmo tempo que evita que o modelo se torne excessivamente confiante em suas predições.

Fundamentação teórica

Entropia Cruzada con Label Smoothing

$$L = -\sum_{k=1}^{K} q'_k \log(p_k)$$

Análise dos Componentes

- $\sum_{k=1}^{K}$: A soma é realizada sobre todas as classes K, garantindo que o modelo seja avaliado em relação a todas as possíveis classes.

- K: Número total de classes (e.g., $K = 1000$ para ImageNet). Essa suavização distribui parte da probabilidade entre todas as classes, reduzindo a confiança excessiva do modelo na classe correta.

- q'_k: As etiquetas suavizadas são calculadas como $q'_k = (1 - \epsilon)\delta_{k,y} + \frac{\epsilon}{K}$, onde:

- $\delta_{k,y}$: É 1 se $k = y$ (classe verdadeira) e 0 caso contrário.

- ϵ: Parâmetro de suavização ($\epsilon = 0.1$).

- $\log(p_k)$: O logaritmo das probabilidades preditas, que penaliza fortemente predições incorretas ou incertas.

Fundamentação teórica

Earth Mover’s Distance (EMD) (para IQA)

O objetivo da (Earth Mover's Distance - EMD) é medir a **diferença entre duas distribuições de probabilidades**, especialmente no contexto de avaliação de qualidade de imagens (Image Quality Assessment - IQA). A fórmula da **EMD** é definida como:

$$L_{\text{quality}} = \left(\frac{1}{K} \sum_{k=1}^{K} |\text{CDF}(p_k) - \text{CDF}(q_k)|^d \right)^{1/d}$$

EMD permite uma comparação mais precisa e robusta. No artigo, essa **métrica foi usada com sucesso para treinar modelos de avaliação de qualidade** em conjuntos de dados como o AVA dataset AVA que proporciona histogramas de calificaciones (de 1 a 10) para cada imagen, basadas en opiniones de múltiples evaluadores humanos.

Fundamentação teórica

Earth Mover's Distance (EMD) (para IQA)

$$L_{\text{quality}} = \left(\frac{1}{K} \sum_{k=1}^{K} |\text{CDF}(p_k) - \text{CDF}(q_k)|^d \right)^{1/d}$$

L_{quality}: A Perda Total (Função de Perda)

K: Número de Classes ou Bins no Histograma

$\sum_{k=1}^{K}$: Soma sobre os Bins

$|\text{CDF}(p_k) - \text{CDF}(q_k)|$: Diferença Absoluta nas Funções de Distribuição Cumulativa

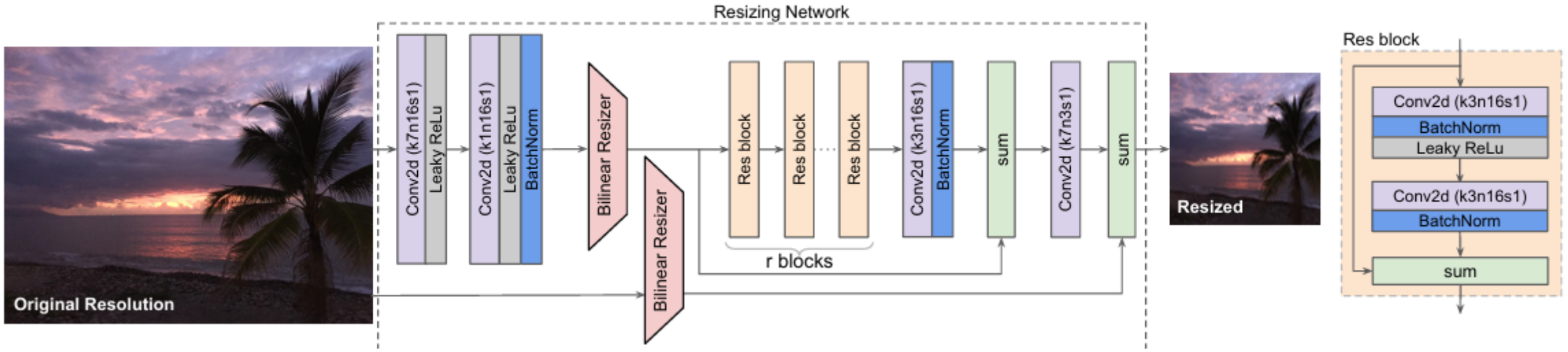
- p_k: Probabilidade predita pelo modelo para o bin k (após Softmax nos 10 logits).
- q_k: Probabilidade real do conjunto de dados para o bin k (do histograma de votos humanos normalizado).

d: Expoente para a Norma

$(...)^{1/d}$: Raiz d-ésima (Normalização Final)

Arquitetura e funcionamento

A Figura 3 apresenta a **arquitetura da CNN** usada como redimensionador aprendido, que transforma imagens em resoluções otimizadas para tarefas como classificação e IQA. Seus componentes são:



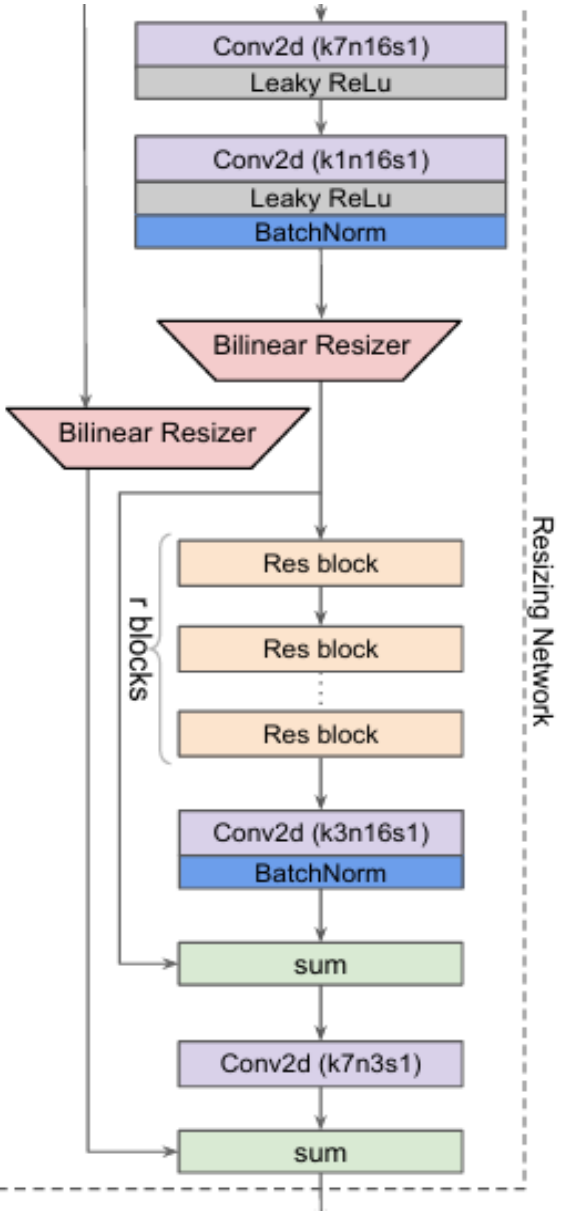
***Entrada**: Imagem em resolução original (e.g., 480 × 640). ***Convolução inicial**: Kernel 7×7 extrai características iniciais.

***Blocos residuais** (*r*=1 ou 2, *n*=16 filtros): Convoluções 3×3, normalização de batch e LeakyReLU.

***Redimensionamento bilinear**: Ajusta características para a resolução desejada (e.g., 192 × 256).

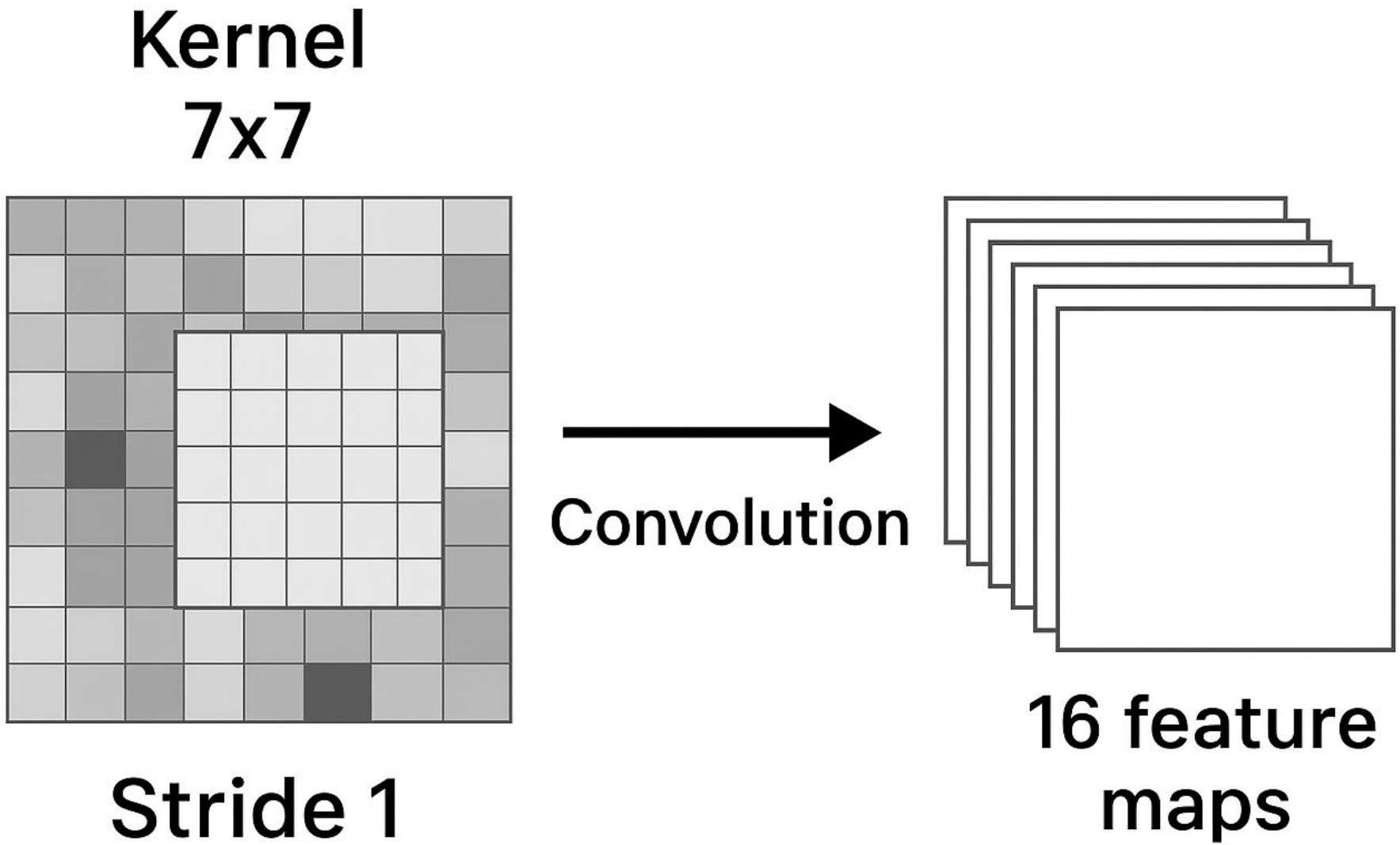
***Conexão de salto**: Soma a imagem bilinear à saída da CNN. ***Convolução final**: Kernel 7×7 gera a imagem redimensionada.

Arquitetura e funcionamento

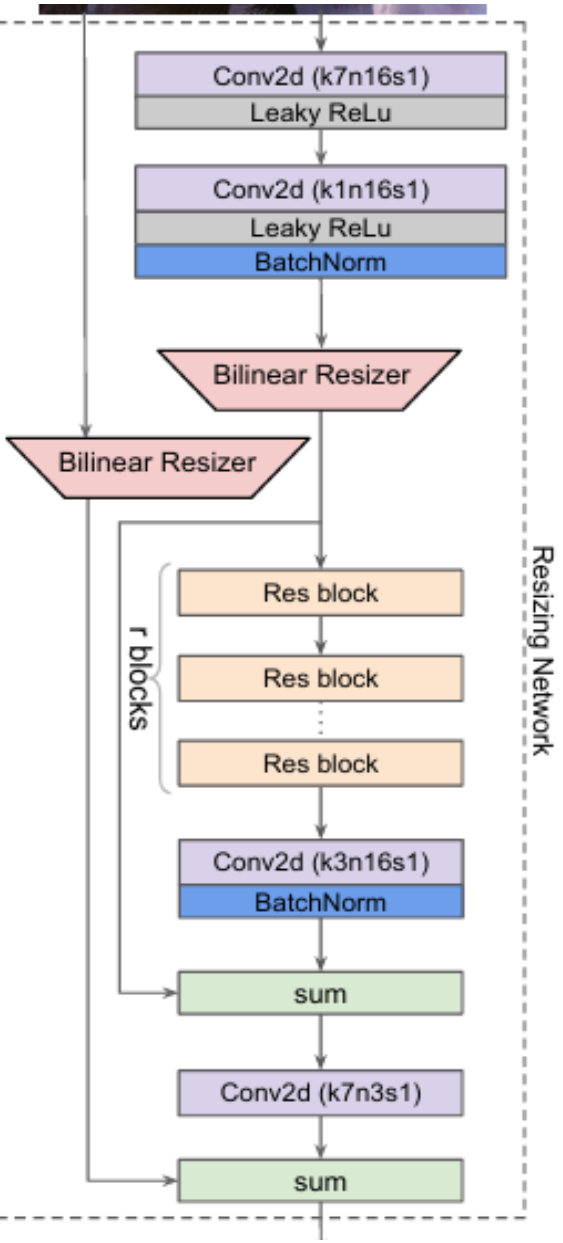


Componente	Descrição	Propósito
Imagem Original	Entrada inicial ao modelo.	Fornecer os dados de entrada para o modelo.
Conv2d(k7n16s1)	Camada convolucional inicial.	Extrair características iniciais da imagem.
Bilinear Resizer 1	Redimensionador bilinear após BatchNorm.	Facilitar o fluxo de informações diretamente para Sum1.
Conexão Direta 1	Conexão paralela do Bilinear Resizer 1 ao Sum1.	Combinar características aprendidas com a imagem redimensionada bilinearmente.
Bloques Residuais (r blocks)	Blocos residuais para aprender características complexas.	Aprender características específicas para a tarefa.
Sum1	Combinação de características aprendidas e bilineares.	Integrar informações de diferentes fontes.
Conv2d(k7n3s1)	Camada convolucional final.	Gerar a imagem redimensionada final.
Bilinear Resizer 2	Redimensionador bilinear da imagem original.	Proporcionar uma base sólida para o redimensionamento.
Conexão Direta 2	Conexão paralela do Bilinear Resizer 2 ao Sum2.	Garantir que não se perca informação importante da imagem original.
Sum2	Combinação final das características.	Produzir a saída final do modelo.

Arquitetura e funcionamento

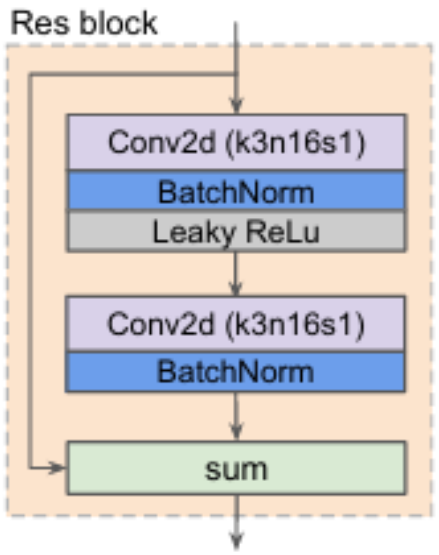


Arquitetura e funcionamento



Componente	Descrição	Propósito
Imagem Original	Entrada inicial ao modelo.	Fornecer os dados de entrada para o modelo.
Conv2d(k7n16s1)	Camada convolucional inicial.	Extrair características iniciais da imagem.
Bilinear Resizer 1	Redimensionador bilinear após BatchNorm.	Facilitar o fluxo de informações diretamente para Sum1.
Conexão Direta 1	Conexão paralela do Bilinear Resizer 1 ao Sum1.	Combinar características aprendidas com a imagem redimensionada bilinearmente.
Bloques Residuais (r blocks)	Blocos residuais para aprender características complexas.	Aprender características específicas para a tarefa.
Sum1	Combinação de características aprendidas e bilineares.	Integrar informações de diferentes fontes.
Conv2d(k7n3s1)	Camada convolucional final.	Gerar a imagem redimensionada final.
Bilinear Resizer 2	Redimensionador bilinear da imagem original.	Proporcionar uma base sólida para o redimensionamento.
Conexão Direta 2	Conexão paralela do Bilinear Resizer 2 ao Sum2.	Garantir que não se perca informação importante da imagem original.
Sum2	Combinação final das características.	Produzir a saída final do modelo.

Arquitetura e funcionamento



Componente	Descrição	Impacto no Modelo
Filtros (n)	Número de filtros nas camadas convolucionais.	Quanto maior o número de filtros (n), maior a capacidade de aprender características complexas, mas também aumenta o número total de parâmetros.
Blocos Residuais (r)	Número de blocos residuais no modelo.	Cada bloco residual contém duas camadas convolucionais. Quanto maior o número de blocos, maior a profundidade e capacidade do modelo, mas também aumenta a complexidade.
Configurações Chave	$n = 16, r = 1$: Modelo mais leve. $n = 32, r = 2$: Modelo mais pesado.	Ideal para tarefas com limitações de memória. Melhora o desempenho em troca de maior consumo computacional.
Número de Parâmetros	Varia entre 11.87K e 93.37K.	O modelo é significativamente mais leve que arquiteturas como ResNet-50 (23M parâmetros).

- O modelo proposto é relativamente **leve** e pode ser **facilmente integrado** em diferentes tarefas de visão computacional.

Arquitetura e funcionamento

A **Tabela 2** apresenta o número de **parâmetros treináveis (em milhares) do redimensionador**, variando o número de blocos residuais (r) e filtros (n):

Filters \ Blocks	$r = 1$	$r = 2$	$r = 3$	$r = 4$
n=16	11.87	16.48	21.08	25.69
n=32	38.08	56.51	74.94	93.37

RESULTADOS

- A **Tabela 1** apresenta as **métricas de avaliação** de desempenho e qualidade para **quatro modelos** de redes neurais amplamente utilizados em **visão computacional**.

Métrica	Resumo
Top-1 Error	Percentual de vezes que a classe correta não é a 1ª predição. (Menor é melhor ↓)
IQA	Tarefa de avaliar a qualidade da imagem comparando com julgamentos humanos.
PLCC	Correlação linear entre predições do modelo e avaliações humanas. (Maior é melhor ↑)

RESULTADOS

- A **Tabela 1** apresenta as **métricas de avaliação** de desempenho e qualidade para **quatro modelos** de redes neurais amplamente utilizados em **visão computacional**.

Task	Model	Top-1 Error ↓	
		Bilinear Resizer	Proposed Resizer
Classification	Inception-v2 [34]	26.7%	24.0%
	DenseNet-121 [9]	33.1%	29.8%
	ResNet-50 [8]	24.7%	23.0%
	MobileNet-v2 [28]	29.5%	28.4%
		PLCC ↑	
		Bicubic Resizer	Proposed Resizer
IQA	Inception-v2 [34]	0.662	0.686
	DenseNet-121 [9]	0.662	0.683
	EfficientNet-b0 [38]	0.642	0.671

- Inception-v2 [34]: desenvolvida por Szegedy et al. (2016)
- DenseNet-121 [9]: Proposta por Huang et al. (2017)
- ResNet-50 [8]: Desenvolvida por He et al. (2016),
- MobileNet-v2 [28]: Proposta por Sandler et al. (2018)
- EfficientNet-b0 [38]: Desenvolvida por Tan y Le (2019)

RESULTADOS

- Inception-v2: Conhecida por usar blocos inception, que **capturam características em múltiplas escalas**. Foi testada para classificação e IQA.
- DenseNet-121: Utiliza conexões densas para promover a **reutilização de características e melhorar a eficiência**. Foi testada em ambas as tarefas.
- ResNet-50: Uma rede residual que usa "**conexões de salto**" para facilitar o **treinamento de redes muito profundas**. Foi testada para classificação.
- MobileNet-v2: Uma arquitetura leve, **otimizada para dispositivos móveis**, que emprega convoluções separáveis. Foi testada para classificação.
- EfficientNet-b0: Uma rede que equilibra **profundidade, largura e resolução** de forma eficiente. Foi testada apenas para IQA.

RESULTADOS

- A **Tabela 1** apresenta as **métricas de avaliação** de desempenho e qualidade para **quatro modelos** de redes neurais amplamente utilizados em **visão computacional**.

Task	Model	Top-1 Error ↓	
		Bilinear Resizer	Proposed Resizer
Classification	Inception-v2 [34]	26.7%	24.0%
	DenseNet-121 [9]	33.1%	29.8%
	ResNet-50 [8]	24.7%	23.0%
	MobileNet-v2 [28]	29.5%	28.4%
		PLCC ↑	
		Bicubic Resizer	Proposed Resizer
IQA	Inception-v2 [34]	0.662	0.686
	DenseNet-121 [9]	0.662	0.683
	EfficientNet-b0 [38]	0.642	0.671

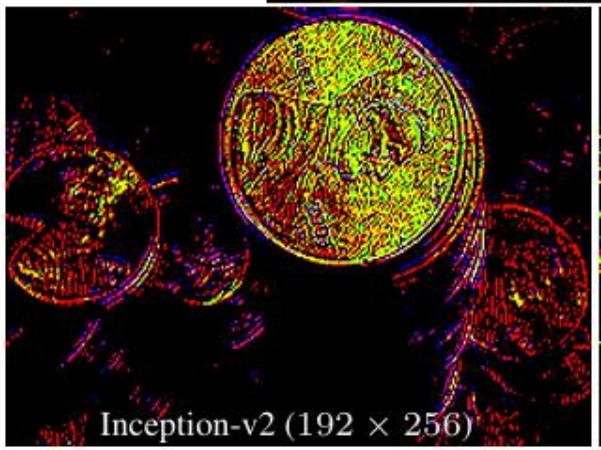
- Inception-v2 [34]: desenvolvida por Szegedy et al. (2016)
- DenseNet-121 [9]: Proposta por Huang et al. (2017)
- ResNet-50 [8]: Desenvolvida por He et al. (2016),
- MobileNet-v2 [28]: Proposta por Sandler et al. (2018)
- EfficientNet-b0 [38]: Desenvolvida por Tan y Le (2019)

RESULTADOS

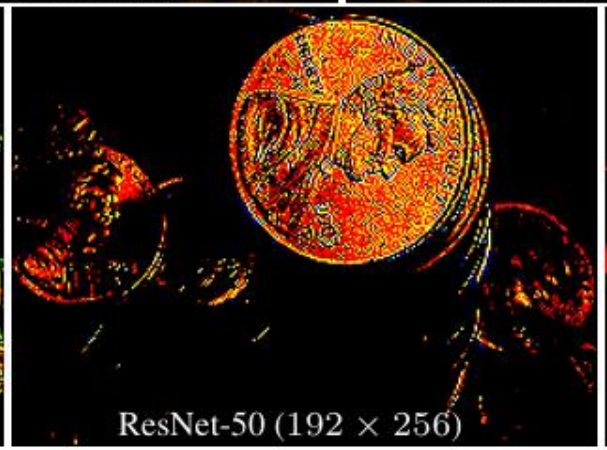

Original (427 × 640)


Bilinear (192 × 256)


Bicubic (192 × 256)

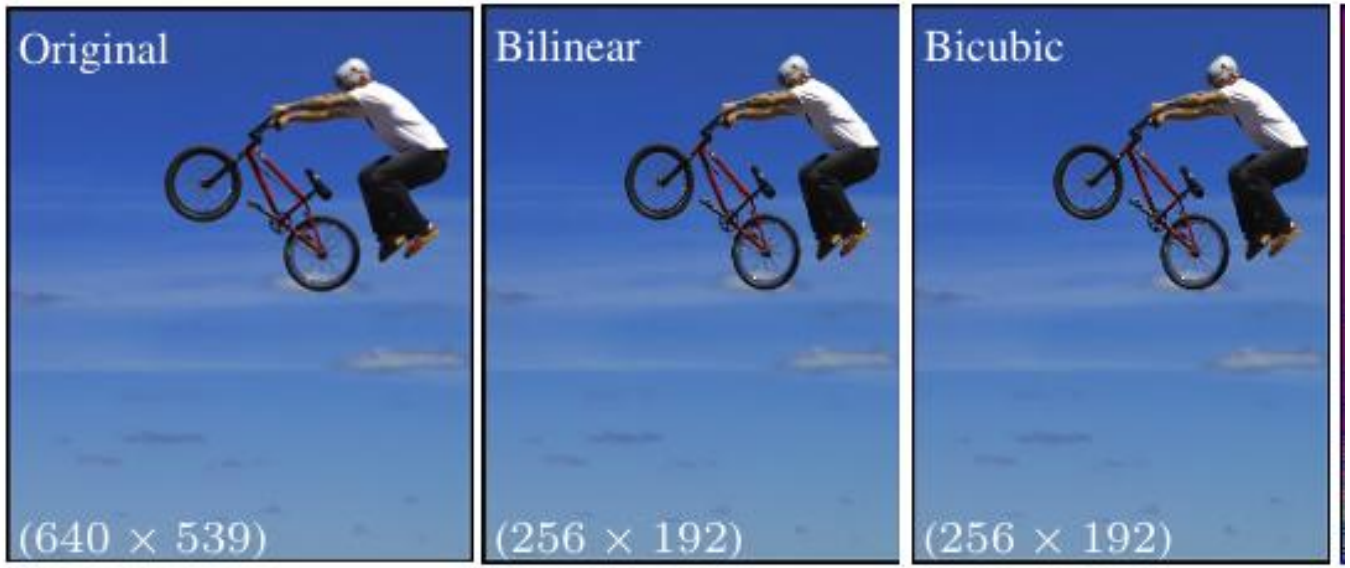

Inception-v2 (192 × 256)

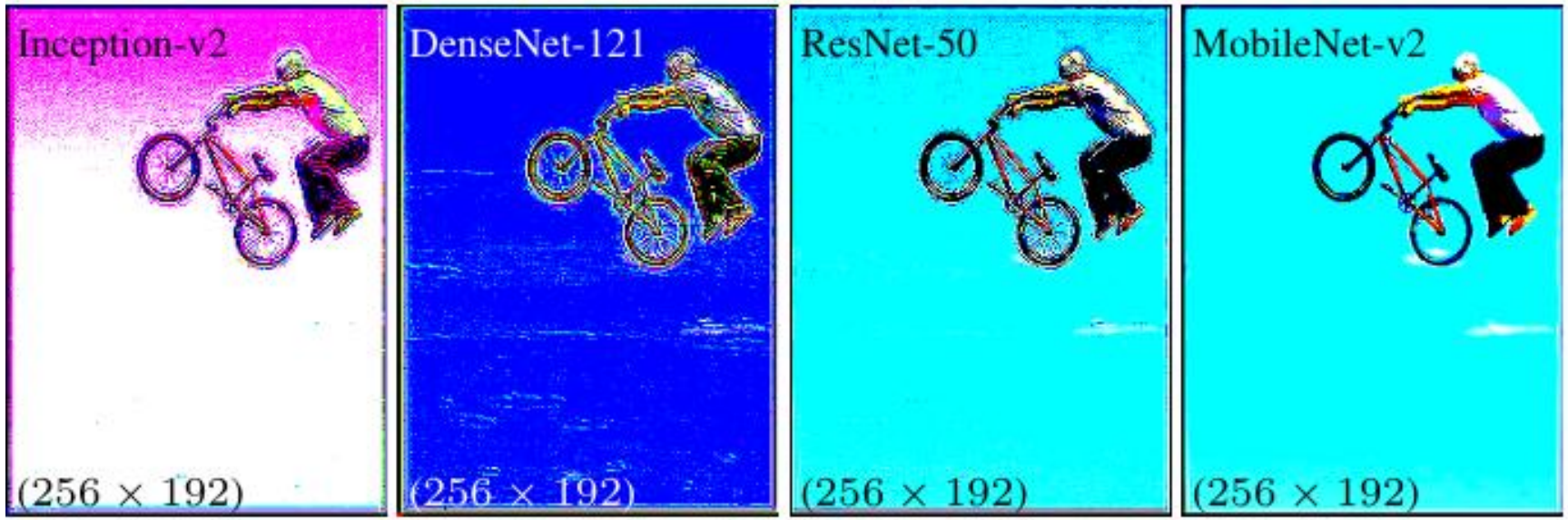

DenseNet-121 (192 × 256)


ResNet-50 (192 × 256)


MobileNet-v2 (192 × 256)

RESULTADOS


Original
(640 × 539)
Bilinear
(256 × 192)
Bicubic
(256 × 192)


Inception-v2
(256 × 192)
DenseNet-121
(256 × 192)
ResNet-50
(256 × 192)
MobileNet-v2
(256 × 192)

RESULTADOS

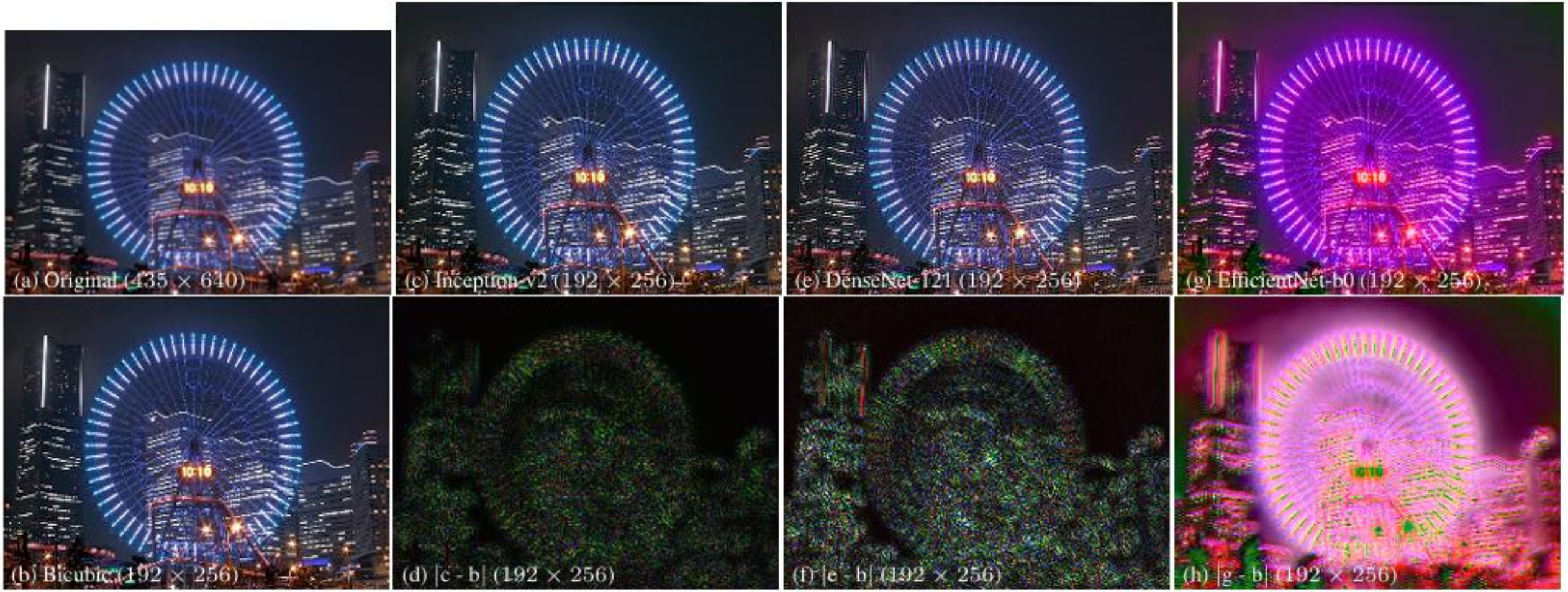


Figure 6. Examples of the proposed learned resizer trained with various IQA models on the AVA dataset [24]. (c), (e), and (f) are results from trained resizers with respective base models. (d), (f), and (h) represent the difference between bicubic and learned resizers.

Vantagens e desvantagens

- O redimensionamento aprendido reflete um avanço em relação aos métodos tradicionais, pois é projetado especificamente para **melhorar o desempenho em tarefas de visão computacional**, em vez de focar na qualidade perceptual da imagem.

Vantagens	Desvantagens
Melhoria no desempenho em tarefas de visão computacional (ex.: redução de erro top-1 de 26,7% para 24,0% em ImageNet).	Complexidade adicional no treinamento debido al entrenamiento conjunto con el modelo de base.
Flexibilidad para redimensionar a cualquier resolución objetivo, ajustándose a diferentes tareas.	Qualidade perceptual no es priorizada, resultando en imágenes menos atractivas visualmente.

Vantagens e desvantagens

- O redimensionamento aprendido reflete um avanço em relação aos métodos tradicionais, pois é projetado especificamente para **melhorar o desempenho em tarefas de visão computacional**, em vez de focar na qualidade perceptual da imagem.

Vantagens	Desvantagens
Arquitectura ligera (11,87 mil parámetros para n=16, r=1) con ganancias significativas en rendimiento.	Dependencia del modelo base, exigiendo entrenamiento específico para cada arquitectura.
Adaptable a diferentes tareas, como clasificación y IQA, con ganancias consistentes.	Aumento en el costo computacional de inferencia (ex.: aumento de FLOPS de 3,88 para 5,07 billones).
Ideal para inferencia remota, minimizando la pérdida de rendimiento en imágenes redimensionadas.	—

Exemplo(s) de aplicação

Contexto	Tarefa de Visão Computacional	Impacto Esperado
Reconhecimento Facial em Sistemas de Segurança	Classificação de Identidades	Menos falsos positivos/negativos, maior confiabilidade em ambientes críticos
Diagnóstico Médico por Imagens	Classificação de Anomalias	Maior precisão na detecção de condições médicas, como tumores ou fraturas

Exemplo(s) de aplicação

Veículos Autônomos	Classificação de Objetos e Cenas	Decisões mais seguras, com menor erro em detecção de pedestres e sinais
Avaliação de Qualidade em Plataformas de Streaming	Avaliação de Qualidade de Imagem (IQA)	Avaliação mais precisa da qualidade de vídeos, melhorando a experiência do usuário
Monitoramento Ambiental por Drones	Classificação de Padrões Ambientais	Identificação mais precisa de áreas degradadas, apoiando ações de conservação

Comparação com outros algoritmos

Método	Desempenho em Tarefas	Qualidade Perceptual	Complexidade Computacional	Flexibilidade
CNN Resizer	Alto (ex.: erro top-1 de 24,0% vs. 26,7% com Inception-v2)	Baixa (não priorizada)	Moderada (11,87 mil parâmetros, 5,07B FLOPS)	Alta (resoluções arbitrárias)
Bilinear	Moderado (erro top-1 de 26,7%)	Moderada	Baixa (não treinável, 3,88B FLOPS)	Moderada (fixo para downscaling)
Bicúbica	Moderado (correlação de 0,642 em IQA)	Alta	Baixa (não treinável)	Moderada (fixo para downscaling)
Superresolução (SRResNet, EDSR)	Baixo para downscaling	Muito alta	Alta (milhões de parâmetros)	Baixa (otimizado para upscaling)
Pré-processamento com Perdas Perceptivas	Moderado a alto	Alta	Alta (treinamento separado)	Moderada (depende da tarefa)

Validação e Refutação

O método de redimensionamento aprendido é validado pelos resultados do artigo, que mostram **melhorias consistentes** no desempenho de tarefas de visão computacional (**Tabela 1**), como **redução de erro top-1** em classificação e aumento da correlação de Pearson em IQA. Sua flexibilidade para resoluções arbitrárias e sua arquitetura leve (**Tabela 2**) o tornam superior a métodos tradicionais e de super resolução em cenários com restrições computacionais e foco na precisão da tarefa. No entanto, a dependência de treinamento conjunto e a falta de priorização da qualidade perceptual podem limitar sua aplicação em contextos onde a estética visual é crucial, como interfaces de usuário. Assim, o método é validado para aplicações técnicas, mas pode ser refutado em cenários que exigem alta qualidade visual sem treinamento adicional.

Referências

•H. Talebi e P. Milanfar, Aprendendo a Redimensionar Imagens para Tarefas de Visão Computacional. arXiv preprint arXiv:2103.09950, 2021.

•K. He, X. Zhang, S. Ren, e J. Sun, Aprendizado residual profundo para reconhecimento de imagens. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

•C. Ledig et al., Super-resolução de imagem única fotorrealística usando uma rede adversária generativa. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4681–4690, 2017.

•I. Goodfellow, Y. Bengio, e A. Courville, Aprendizado Profundo. MIT Press, 2016.

•https://github.com/keras-team/keras-io/blob/master/examples/vision/learnable_resizer.py

QUIZ

QUIZ SEMINARIO

GITHUB

Obrigado!