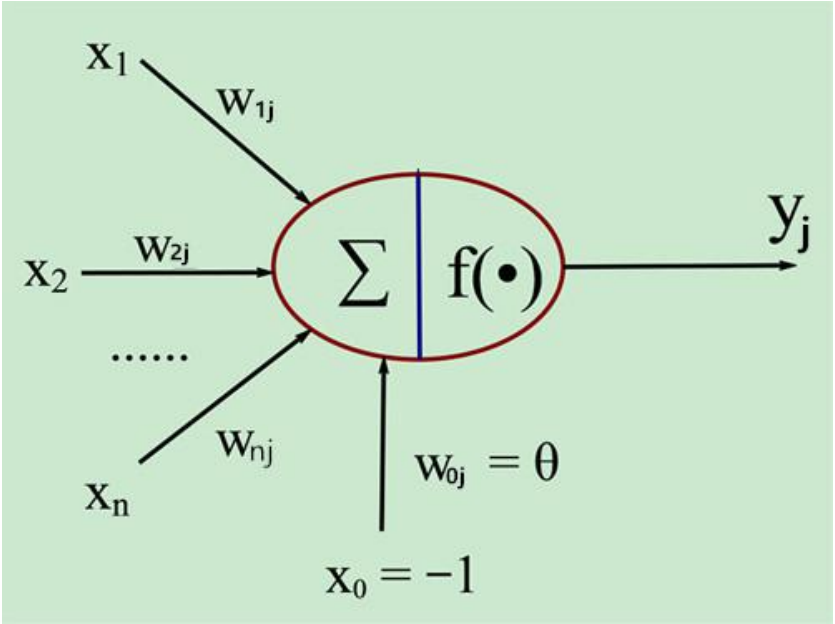
人机图像感知中的视觉歧义

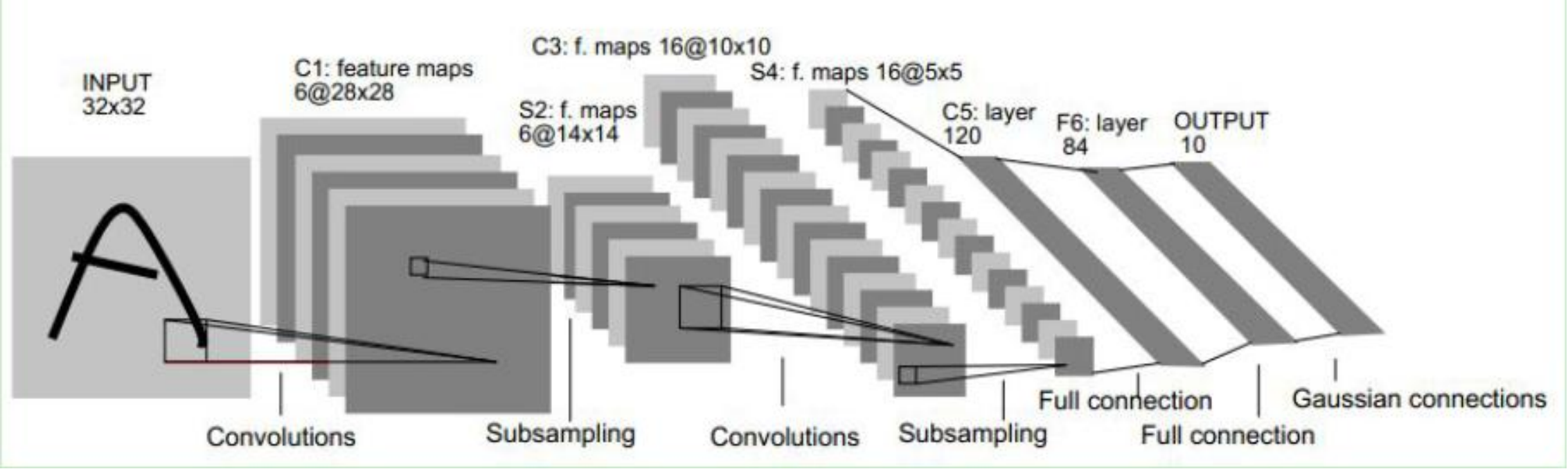
王志越 11810125
唐云龙 11911607
徐思婷 11911635
张成杰 11911918
江欣乐 11810806

- 只要有足够的简单神经元，在这些神经元互相连接并同步运行的情况下，可以模拟任何计算函数（M-P模型）。该模型借鉴了已知的神经细胞生物过程原理，是第一个神经元数学模型，是人类历史上第一次对大脑工作原理描述的尝试。
- Donald Olding Hebb在《The Organization of Behavior》中对神经元之间连接强度的变化进行了分析，首次提出来一种调整权值的方法，称为Hebb学习规则。Hebb学习规则主要假定机体的行为可以由神经元的行为来解释。Hebb受启发于巴普罗夫的条件反射实验，认为如果两个神经元在同一时刻被激发，则它们之间的联系应该被强化。在这种学习中，由对神经元的重复刺激，使得神经元之间的突触强度增加。
- 美国神经学家Frank Rosenblatt提出可以模拟人类感知能力的机器，并称之为“感知机”。1957年，在Cornell航空实验室中，他成功在IBM704机上完成了感知机的仿真，并于1960年，实现了能够识别一些英文字母的基于感知机的神经计算机
- 1985年，Geoffrey Hinton使用多个隐藏层来代替感知机中原先的单个特征层，并使用BP算法来计算网络参数。进而有后续的深度神经网络的应用。

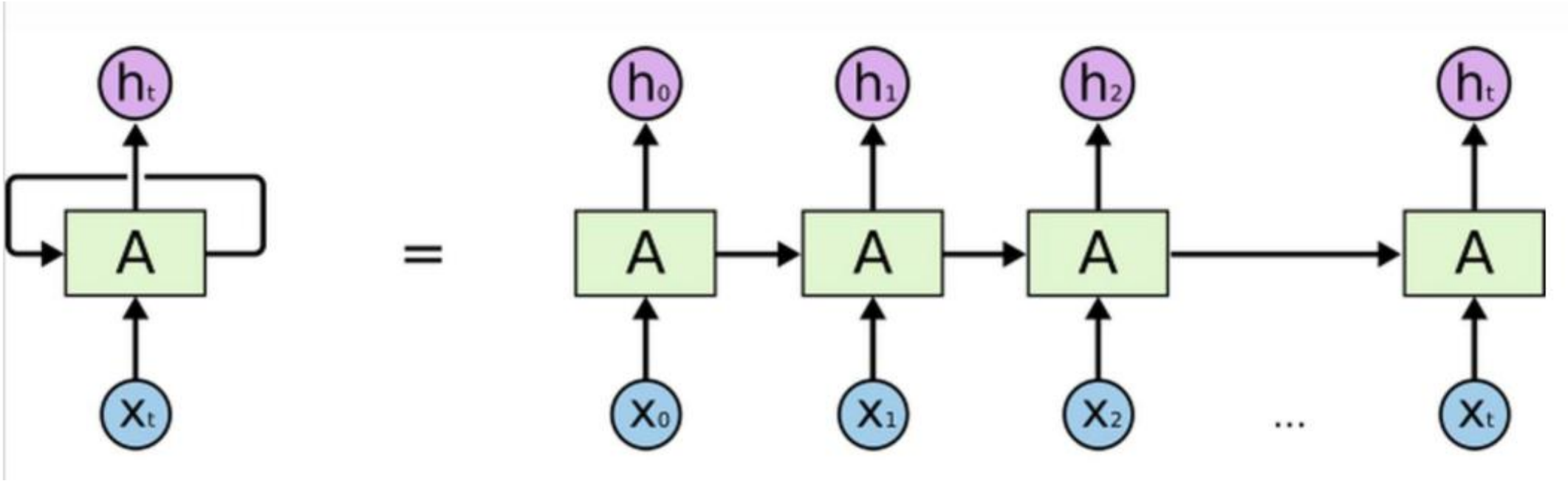


$$W_{ij}(t+1) := W_{ij}(t) + \alpha x_i x_j$$

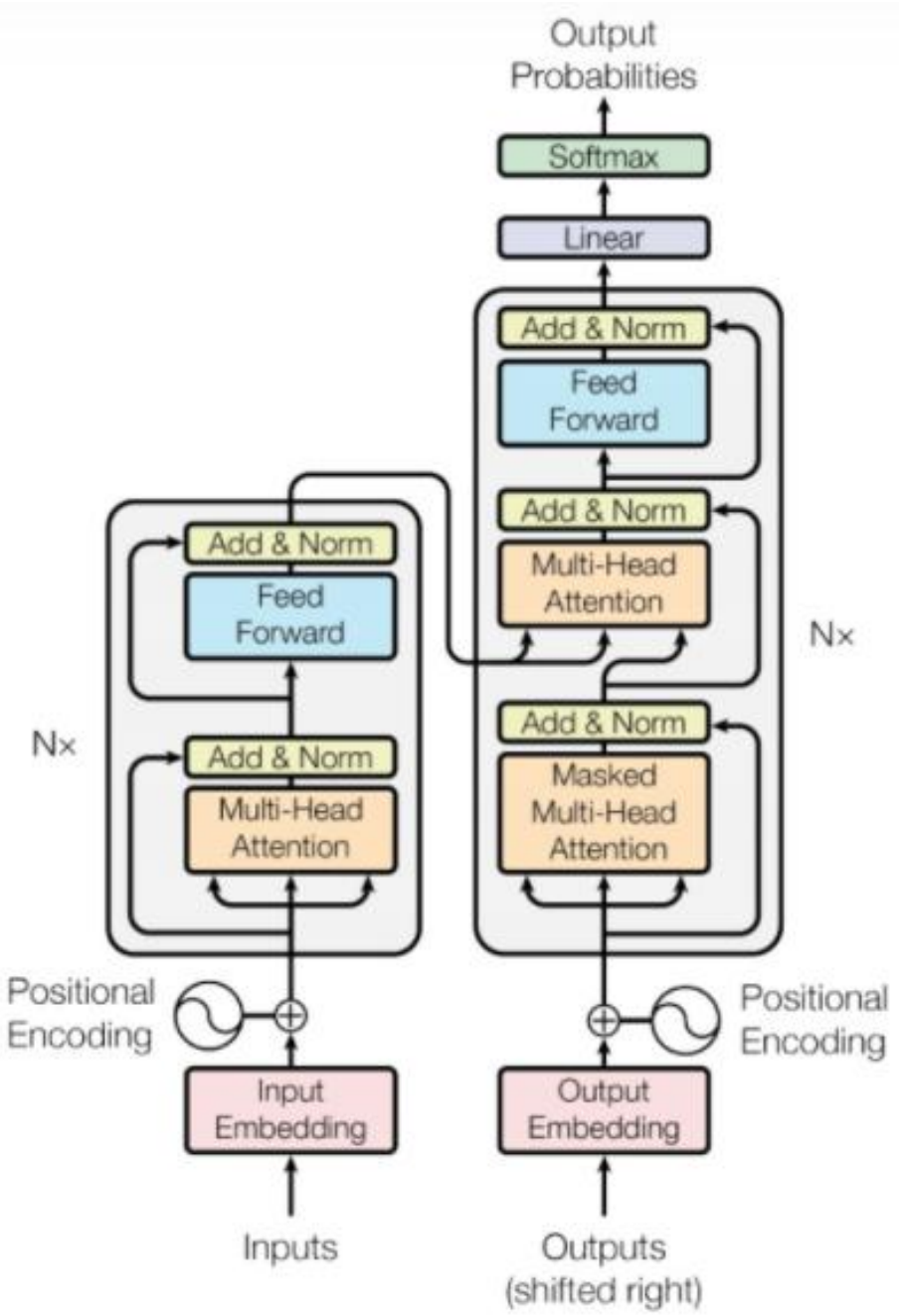
- CNN中，卷积操作对应的是人眼在看外界物体时的扫视过程。池化操作对应视网膜中的神经活动，在视网膜中，局部一个或多个感光细胞信号汇聚到一个双极细胞；一个或多个双极细胞的输出信号汇聚到一个神经节细胞。神经信号的脉冲信号与连续信号分别对应了两种不同的激活函数。



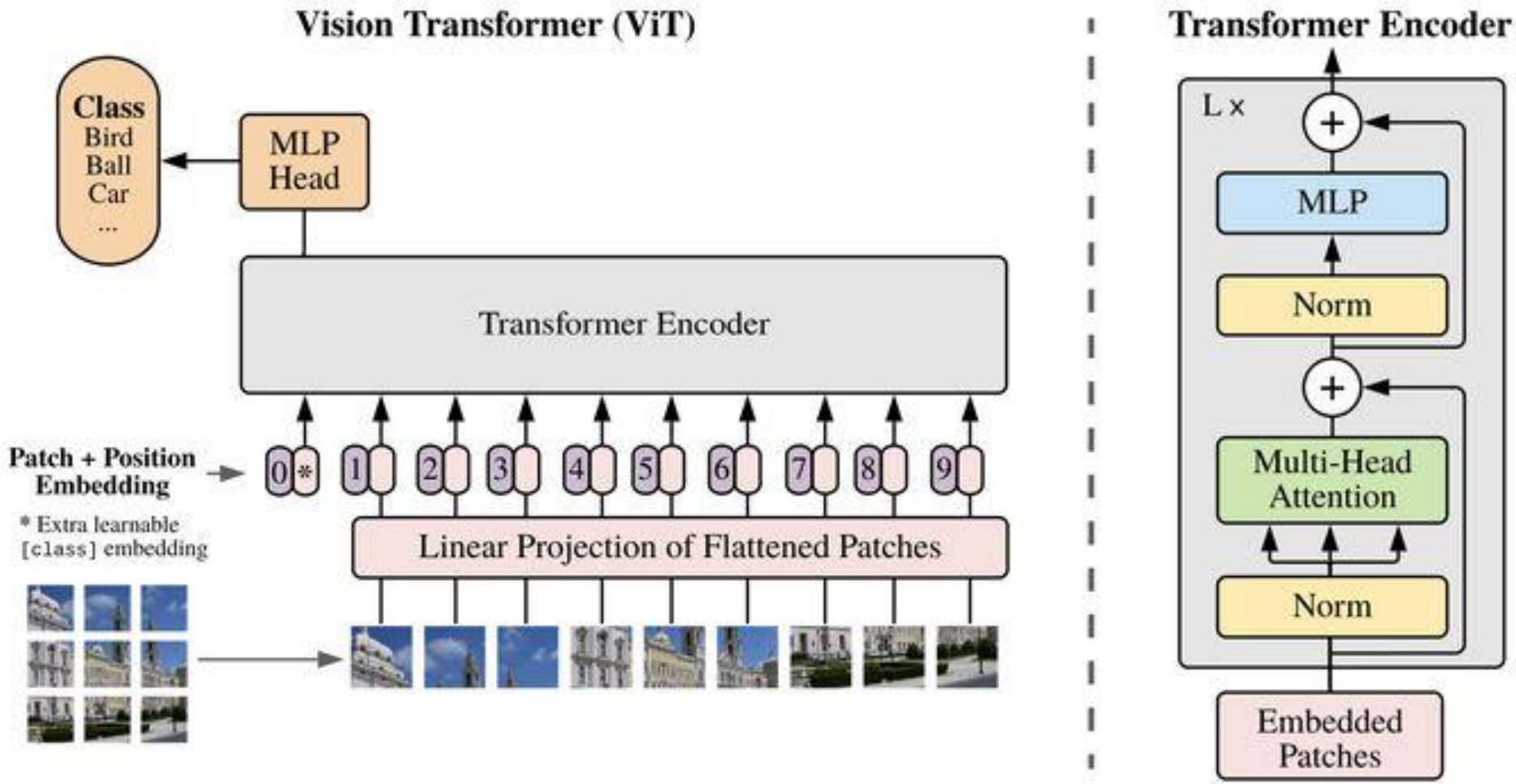
- RNN中，RNN的发展源于hopfield神经网络，hopfield神经网络是全连接的，神经元的输入源自于其他神经元的输出，进而出现基于时序的传统RNN，在传统RNN的基础上，添加对记忆与遗忘模拟，产生了LSTM模型。



- Transformer，通过特征向量与位置向量相加来替代RNN中的时序信息，并通过查询，关键词，价值矩阵获得自注意力矩阵，从而实现自注意力机制。



ViT


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches
Transformer Encoder
L x
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

ViLT

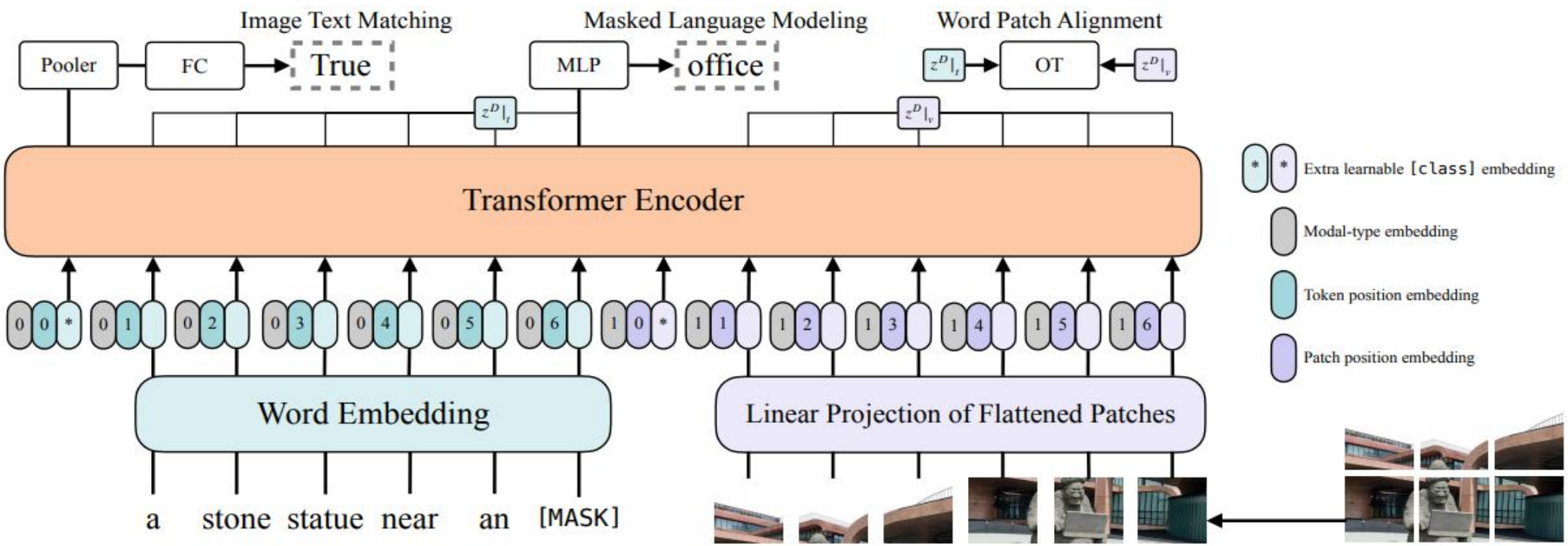
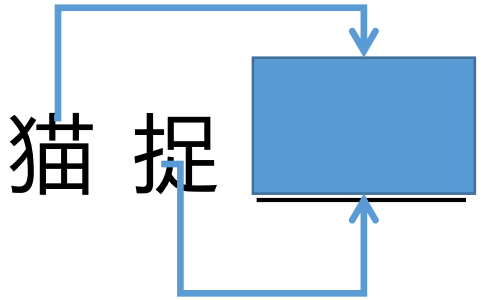


Image Text Matching
Masked Language Modeling
Word Patch Alignment
Pooler
FC
True
MLP
office
OT
Transformer Encoder
Word Embedding
Linear Projection of Flattened Patches
a stone statue near an [MASK]
Extra learnable [class] embedding
Modal-type embedding
Token position embedding
Patch position embedding

ViLT

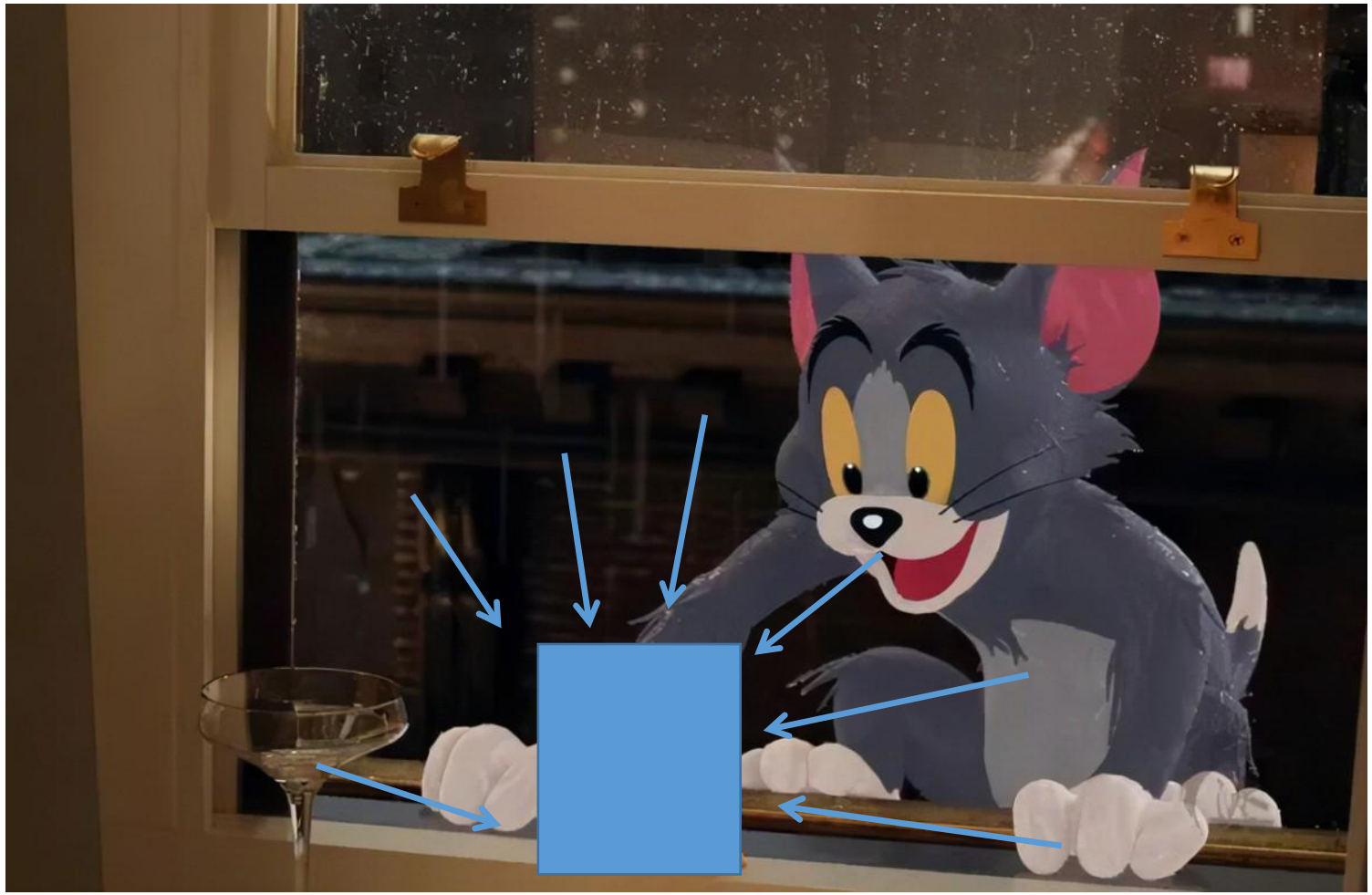


ViLT

猫 捉 老 鼠

ViLT

猫 捉 老 鼠

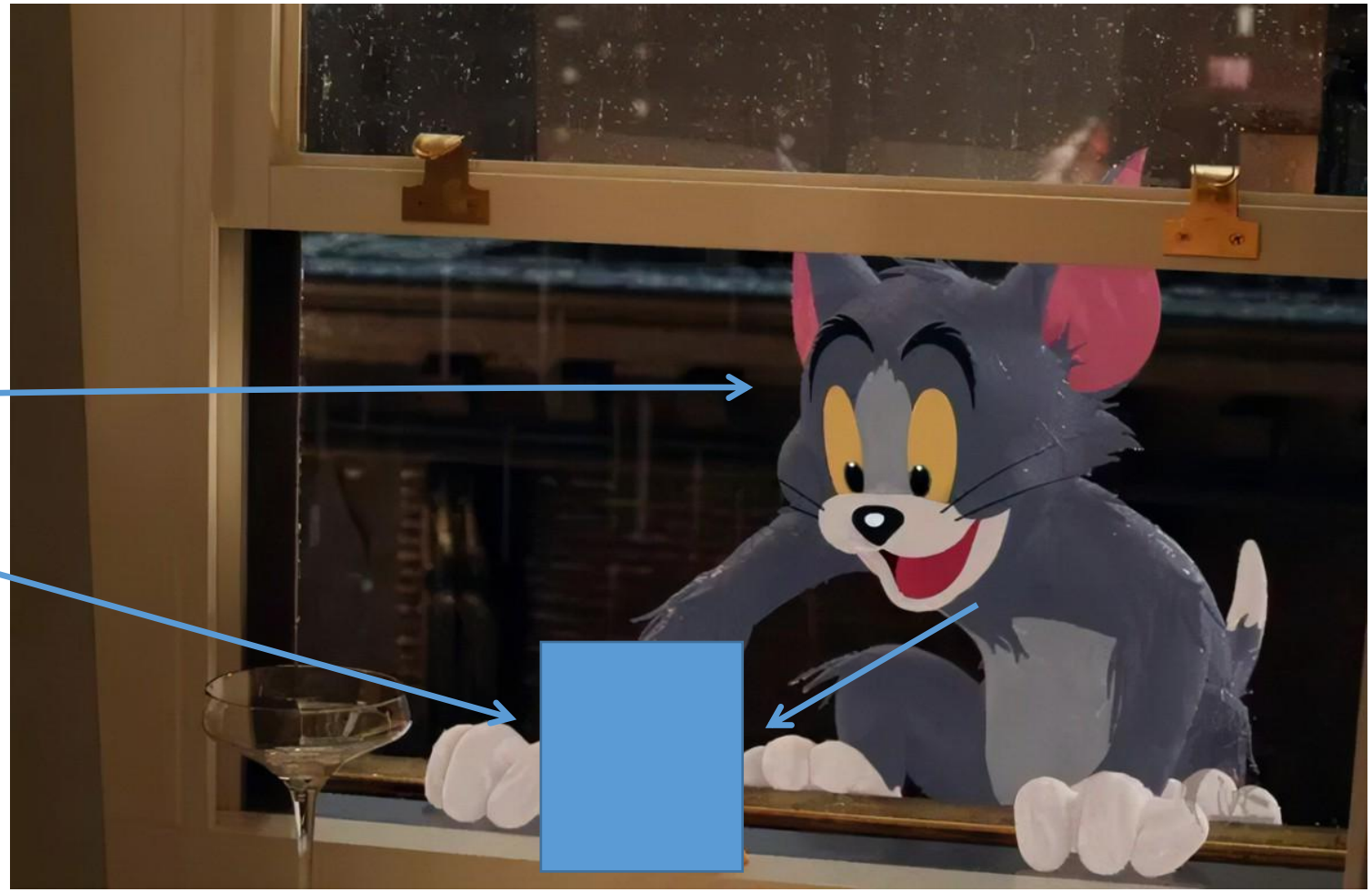


ViLT

猫 捉 老 鼠

ViLT

猫 捉 老 鼠



ViLT

猫 捉 老 鼠

ViLT

猫 捉 [redacted]

ViLT

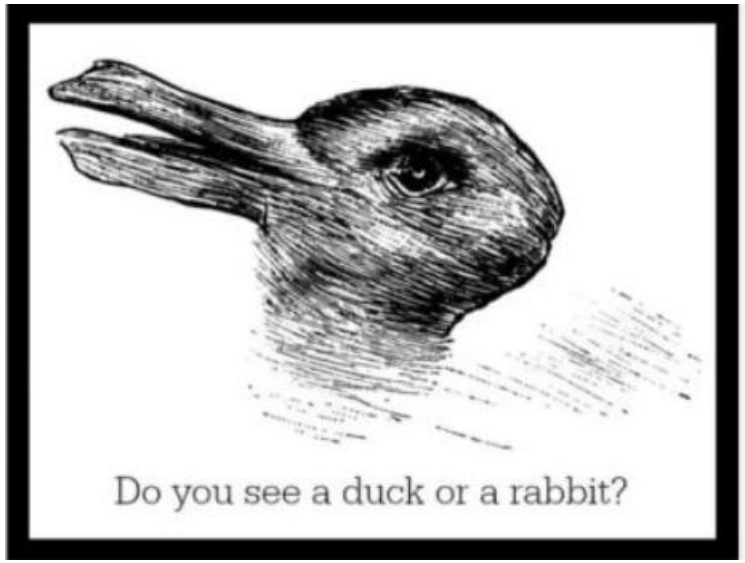
猫 捉 老 鼠

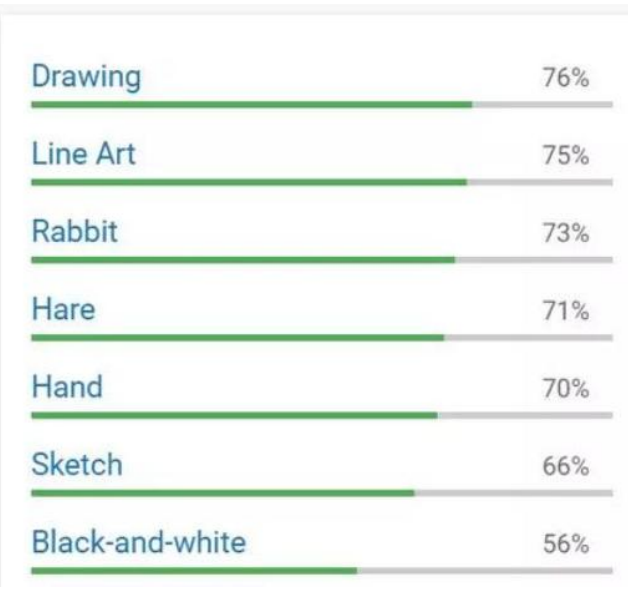
VQA

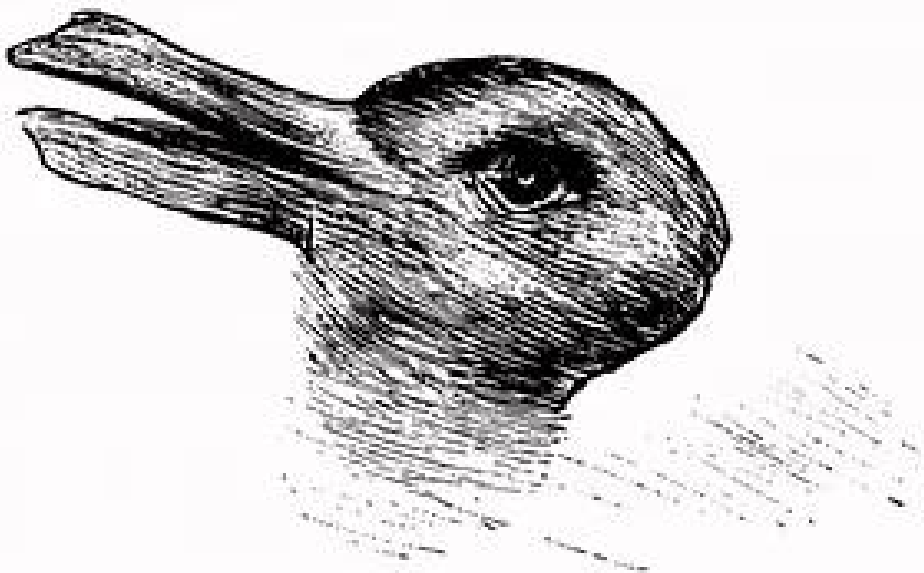
视觉问答（Visual Question Answering,VQA）是一项结合计算机视 觉和自然语言处理的学习任务。 计算机视觉主要是对给定图像进行处 理，包括图像识别，图像分类等任务。 自然语言处理主要是对自然语言， 文本形式的内容进行处理以及理解，包括机器翻译，信息检索，生成文 本摘要等任务。 **视觉问答是需要对给定图像和问题进行处理，经过一定 的视觉问答技术处理过后生成自然语言答案，是对二者的结合**

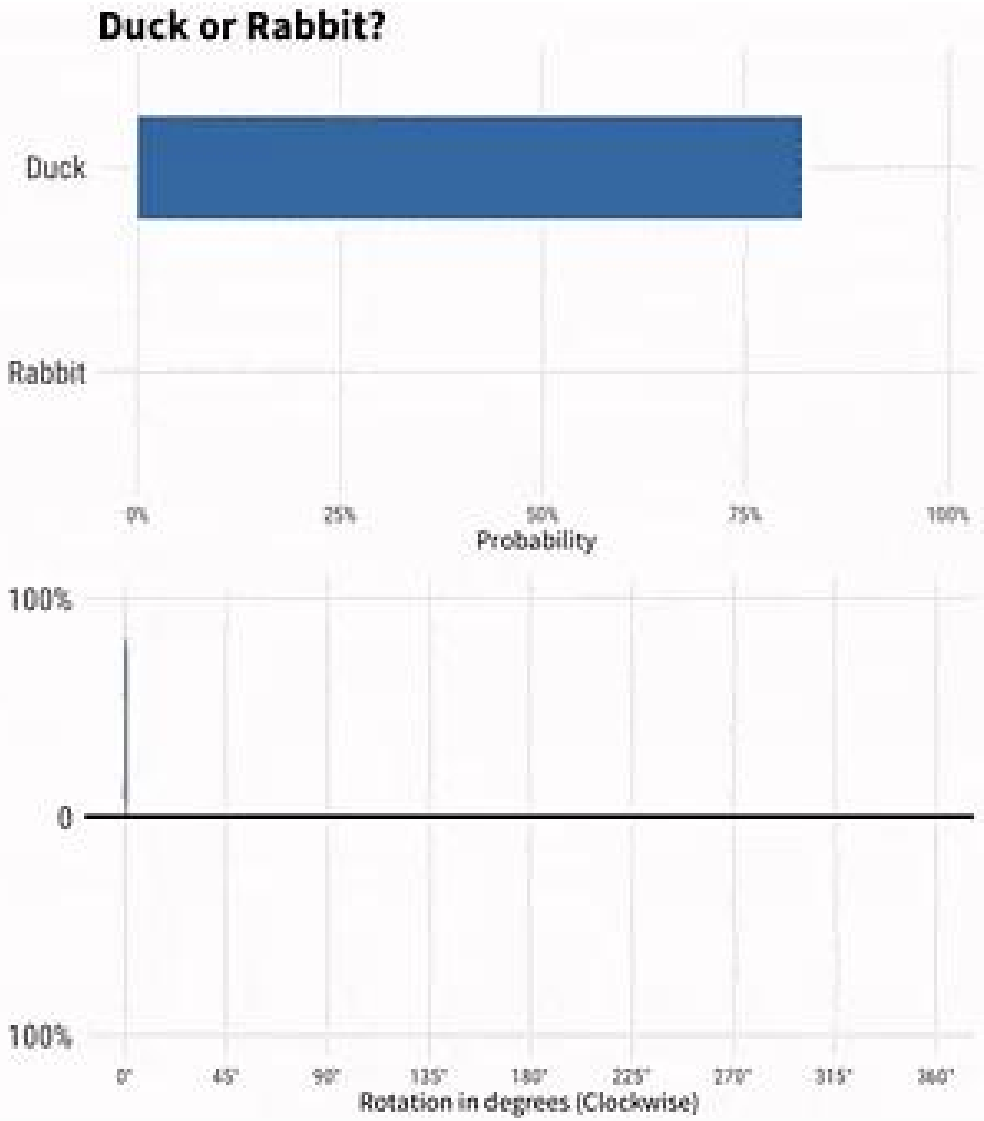
AI对图像的感知歧义

- 谷歌AI对图片的认知也存在歧义。不同的角度有不同的结果。









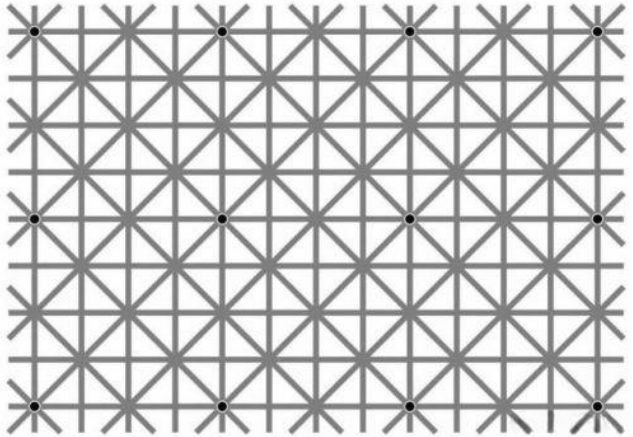
Predictions provided by the Google Cloud Vision API
Animation by Max Woolf (@minimaxir)

问卷设计与结果统计

总共11道题

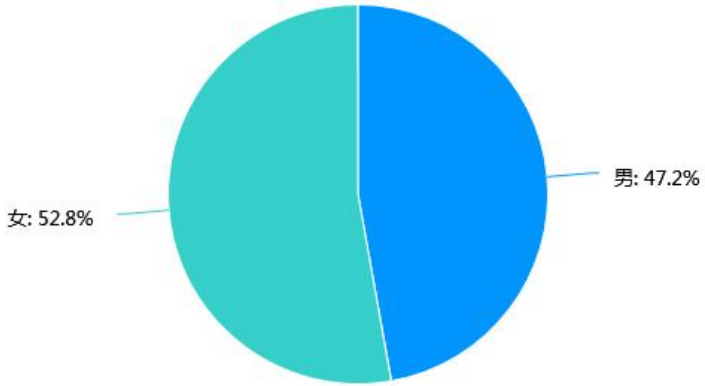
- 基本信息：性别和年龄（2道）
- 无歧义的视觉识别
  颜色辨认、物体识别和细节认知（3道）
- 歧义的视觉识别
  几何错觉和物体识别错觉等（6道）

图中左边绿色轿车车牌号是多少？（无歧义）

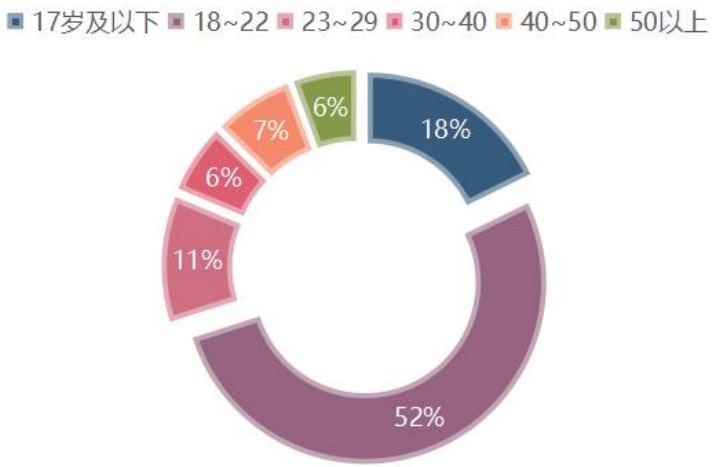


是否能看到黑色小点？（歧义）

性别统计结果



年龄统计结果



男女均衡，18-22岁占比较大，参与者以年轻人为主

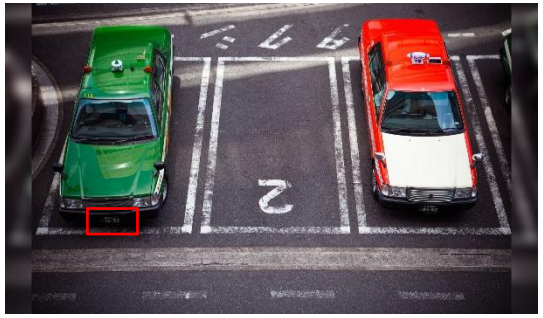
Q3 图片中的围墙是石头做的吗？

Q4 图片中的花朵主要是粉色的吗？

90%以上的人选择是

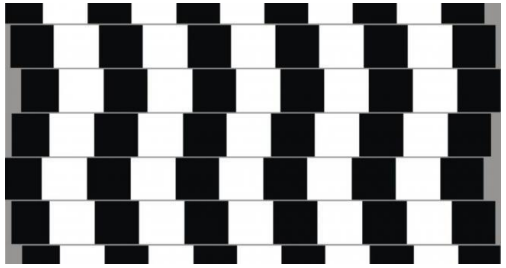
机器回答：**Yes**

Q5 绿色车的车牌号存在以下数字？



36和91: 5.61%　yes
35和91: 2.8%　yes
36和92: 23.36%　no
35和92: 3.27%　no
识别不出来：64.95%

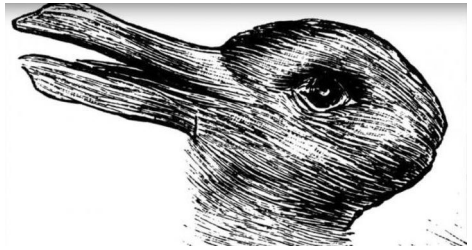
Q6 图中的线平行吗？



是：65.42%
不是：34.58%

机器回答：**Yes**

Q7图片中你第一眼看到的是什么？



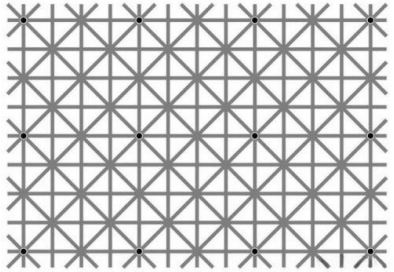
鸟：86.92%
兔子：13.08%

机器回答：**bird**

Q8 是鸟？或者兔子？　能够识别：90%以上

机器也能够准确识别出兔子

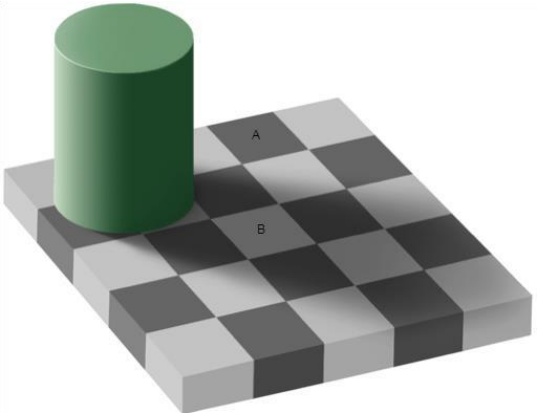
Q9 能否看到黑色的点？



95%以上的人能够看到

机器回答：yes

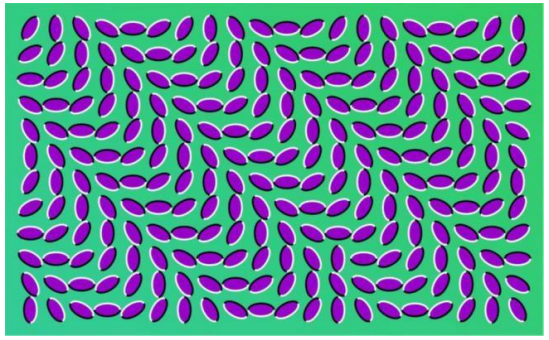
Q10 A和B哪个区域更黑？



A区：60%
B区：40%

机器回答：
颜色不一样
且B区更黑

Q11 图中叶子在动？



是：69.75%
不是：39.25%

机器回答：不是

未来的发展方向

- 从网络架构上看，vilt主要采用的是transformer，其论文中也对比采用CNN对图片信息处理后，转化为特征向量。但CNN可能能够应用到其他形式的处理来实现CNN与transformer的结合，进而有可能产生积极影响。
- 从与生物相关的方向上看，生物注意力机制作用原理尚且没有完整详尽的解释，因此可能会与transformer的注意力机制存在差异性，将来可能会产生基于生物注意力机制的算法来替代transformer。
- 从实验结果上看，vilt模型无法感觉到人类的错觉现象，与人的认知图片，处理图片存在一定的差异，并且，人的错觉可能因为眼睛的移动导致注意力的改变以及对注意区域的边缘区域感知而产生错觉。因此VILT可能能够在注意力转移的方向与对注意力区域的边缘区域的感知做出进一步的功能提升。